UNIVERSIDADE FEDERAL DE SÃO CARLOS
Departamento de Estatística

Extensões do resíduo quantílico

Ana Carolina do Couto Andrade

Tese de Doutorado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

Ana Carolina do Couto Andrade

Extensões do resíduo quantílico

Tese 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 Doutora em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Gustavo Henrique de Araujo Pereira

USP – São Carlos
Novembro de 2022

Ana Carolina do Couto Andrade

Extensions of the quantile residual

Doctoral 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 Doctorate Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Gustavo Henrique de Araujo Pereira

USP – São Carlos
November 2022

UNIVERSIDADE FEDERAL DE SÃO CARLOS

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

Folha de Aprovação

Defesa de Tese de Doutorado da candidata Ana Carolina do Couto Andrade, realizada em 20/12/2022.

Comissão Julgadora:

Prof. Dr. Gustavo Henrique de Araujo Pereira (UFSCar)

Profa. Dra. Cibele Maria Russo Novelli (USP)

Prof. Dr. Cristian Marcelo Villegas Lobos (ESALQ/USP)

Prof. Dr. Manoel Ferreira dos Santos Neto (UFCG)

Prof. Dr. Tiago Maia Magalhães (UFJF)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

AGRADECIMENTOS

Agradeço ao meu filho Carlos, minha maior fonte de motivação.

À Deus, pelas oportunidades e condições que tive até chegar aqui.

Aos meus pais, Thales e Alessandra, que me ensinaram desde cedo o valor do estudo. Mesmo frente às dificuldades, sempre me proporcionaram as melhores condições de estudo e conforto. Por todo suporte e incentivo para que eu alcançasse meus sonhos. Meus sucessos são e sempre serão divididos com vocês.

Ao meu irmão Iago e sua esposa Láisa, por se fazerem presentes nos cuidados com meus pais, já que a distância geográfica não me permite estar perto o quanto eu gostaria.

Ao meu esposo Alessandro, por todo apoio e compreensão. Por ouvir com tanta atenção quando explico algo sobre minha pesquisa, mesmo não entendendo a maioria das palavras.

Ao meu orientador, Gustavo, pela confiança em me aceitar como orientanda e pelo saber que transmitiu. Por ser presente em todas as etapas do doutorado, sempre disposto a me ouvir, ajudar, incentivar e corrigir. Sobretudo, por todo respeito e compreensão, que sem dúvidas tornaram essa caminhada mais leve. Por viabilizar nossas reuniões semanais durante a gestação e após o nascimento do meu filho, adequando seus horários às sonecas do Carlos.

À todos os funcionários e professores do ICMC-USP e DEs-UFSCar, por serem sempre cordiais, atenciosos e dedicados, em especial ao Julio Cezar de Barros e à Monique da Conceição, e aos professores Alexsandro Giacomo Grimbert Gallo e Daiane Aparecida Zuanetti.

Ao professor Rinaldo Artes, pela enorme contribuição no Capítulo 2.

À UNIVESP, pelo apoio financeiro e oportunidade de vivência prática da docência, podendo também estar no papel de orientadora de alunos da graduação.

Agradeço aos membros da banca examinadora, Cibele Russo Maria Novelli, Cristian Marcelo Villegas Lobos, Manoel dos Santos Neto e Tiago Maia Magalhães, pelo interesse, disponibilidade e pelas valiosas críticas e sugestões ao trabalho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

RESUMO

ANDRADE, A. C. C. **Extensões do resíduo quantílico**. 2022. 129 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

Os modelos de regressão possuem profunda importância em análises que têm por objetivo investigar a relação entre uma variável dependente e um conjunto de variáveis preditoras. A análise de diagnóstico é uma etapa fundamental na verificação do ajuste de um modelo de regressão, cujos objetivos são identificar possíveis pontos discrepantes e/ou influentes e verificar possíveis afastamentos das suposições feitas para a modelagem. Nesse caso, é desejável obter resíduos cuja distribuição seja bem aproximada pela distribuição Normal padrão, visto que suas propriedades e comportamento são conhecidos. O resíduo quantílico é uma importante classe de resíduos com essa característica, em que sua distribuição é assintoticamente Normal padrão quando os parâmetros do modelo são consistentemente estimados. Outro problema comum no âmbito da análise de regressão abrange a seleção de modelos, cujo intuito consiste em selecionar o melhor modelo teórico dentre um conjunto de modelos candidatos. O objetivo desse trabalho foi desenvolver extensões do resíduo quantílico, em aspectos de análise diagnóstica e seleção de modelos. Para verificação da adequação do modelo, introduz-se um resíduo assintoticamente Normal padrão distribuído, que pode ser usado para qualquer modelo de regressão circular-linear paramétrico. Para detecção de possíveis pontos *outliers*, é proposta uma extensão em modelos de regressão beta inflacionados em dois e três pontos, cuja distribuição caudal é semelhante a distribuição Normal padrão. Por fim, são introduzidos três critérios de seleção de modelos por meio de testes de bondade do ajuste com o uso do resíduo quantílico, em um contexto específico de seleção da distribuição da variável resposta em modelos aditivos generalizados para localização, escala e forma (GAMLSS).

Palavras-chave: análise de diagnóstico, regressão beta inflacionada, regressão circular, resíduo quantílico, seleção de modelos.

ABSTRACT

ANDRADE, A. C. C. **Extensions of the quantile residual**. 2022. 129 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

Regression models have profound importance in analyses that aim to investigate the relationship between a dependent variable and a set of predictor variables. The diagnostic analysis is a fundamental step in validating a regression model, whose objectives are to identify possible discrepant and/or influential points and to verify possible deviations from the assumptions made for modeling. In this case, it is desirable to obtain residuals whose distribution is close to the standard Normal distribution, since their properties and behavior are known. The quantile residual is an important class of residuals with this characteristic, where its distribution is asymptotically standard Normal when the model parameters are consistently estimated. Another common problem in regression analysis is model selection, which consists in selecting the best theoretical model from a set of candidate models. The objective of this work is to develop extensions of the quantile residuals, in aspects of diagnostic analysis and model selection. To check the model fit, an asymptotically distributed standard Normal residual is introduced, which can be used for any parametric circular-linear regression model. For the detection of possible outliers, an extension is proposed on two and three-point inflated beta regression models, whose tail distribution is similar to the standard Normal distribution. Finally, three model selection criteria are introduced by testing goodness of fit using the quantile residuals in a specific context of response variable distribution selection in generalized additive models for location, scale and shape (GAMLSS).

Keywords: circular regression, diagnostic analysis, inflated beta regression, model selection, quantile residuals.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Os modelos de regressão possuem profunda importância em análises que têm por objetivo investigar a relação entre uma ou mais variáveis, denominadas dependentes, e outras, ditas explicativas. Existem várias classes de modelos, que variam de acordo com as especificidades de cada problema e conjunto de variáveis. Dessa forma, uma etapa importante na regressão é a análise de diagnóstico. Seus principais objetivos são identificar pontos discrepantes e/ou influentes e verificar possíveis afastamentos das suposições feitas para a modelagem. Para tal, faz-se uso dos resíduos gerados pelo modelo ajustado.

Os resíduos são funções de uma medida de afastamento de uma observação para seu próprio valor ajustado. Isso posto, um grande desafio é encontrar resíduos com propriedades e comportamento conhecidos, o que facilita a interpretação dos gráficos e as posteriores identificações de pontos influentes e/ou *outliers* no modelo. Para verificar a adequação do modelo, é desejável, por exemplo, usar resíduos que são bem aproximados pela distribuição Normal padrão.

Diversos tipos de resíduos foram propostos por alguns autores, como por exemplo os resíduos quantílicos (DUNN; SMYTH, 1996). Esta classe de resíduos fundamenta-se no teorema da inversa da função distribuição acumulada, e tem distribuição assintoticamente Normal padrão quando os parâmetros do modelo são consistentemente estimados. Feng, Sadeghpour e Li (2017) estudaram o comportamento deste resíduo no contexto de modelos lineares generalizados (MLG), realizando a comparação do mesmo com os resíduos *deviance* e de Pearson. Sob os critérios considerados pelos autores, observou-se que em situações de má especificação do modelo e em amostras finitas o resíduo quantílico apresenta melhor desempenho que os demais. Pereira (2019) mostrou que para os modelos de regressão beta, a distribuição do resíduo quantílico é bem aproximada pela Normal padrão, mesmo em amostras pequenas. Lemonte e Moreno-Arenas (2019) observaram o mesmo para o contexto dos modelos generalizados de Johnson S_B. Scudilio e Pereira (2020) propuseram um ajuste aos resíduos quantílicos em MLG, constatando que, na

análise de diagnóstico, este resíduo se sobressai ao resíduo de Pearson padronizado e ao resíduo *deviance* padronizado.

Em alguns contextos, o resíduo quantílico não é adequado para a condução da análise de diagnóstico. Dessa forma, este trabalho propõe extensões para esta classe de resíduos sob os contextos da regressão circular e regressão inflacionada em dois ou três pontos.

Primeiramente, considerou-se o contexto de dados circulares, em que as variáveis estão definidas em um espaço limitado e fechado e, portanto, devem ser analisados conforme técnicas apropriadas, geralmente distintas daquelas para dados em um espaço euclidiano usual (GOULD, 1969). Aproveitando a escassez de estudos no âmbito da análise de diagnóstico em modelo de regressão circular (LIU *et al.*, 2016), propõe-se neste trabalho uma extensão do resíduo quantílico para modelos de regressão circular, visto que as diferenças topológicas entre a reta real e o círculo não permitem a utilização da definição de Dunn e Smyth (1996), sem que haja nenhuma correção ou adaptação. Para tal, considerou-se a classe de modelos de regressão proposta por Fisher e Lee (1992), cuja variável resposta possui distribuição von Mises. Além disso, foram propostas duas extensões para o modelo de Fisher e Lee, contemplando as distribuições *sine-skewed* von Mises e *wrapped* Cauchy.

O segundo contexto abordado diz respeito às variáveis do tipo taxa ou proporção, cujo suporte da variável resposta possui dois ou três pontos com massa de probabilidade positiva. Neste caso, a distribuição beta não se faz pertinente. Ospina e Ferrari (2010) apresentaram uma nova distribuição, denominada distribuição beta inflacionada, capaz de modelar variáveis com probabilidade positiva de assumir valores em pelo menos um dos limites do intervalo $(0,1)$. Para além deste cenário, surgem ainda variáveis cujas observações não podem assumir valores em um determinado intervalo $(0,c)$. Em tal caso, tem-se como proposta para o ajuste deste tipo de variável, a distribuição beta inflacionada truncada, que consiste na mistura de uma distribuição beta com suporte no intervalo $(c,1)$ e uma distribuição trinomial que assume os valores 0, 1 e c (PEREIRA; BOTTER; SANDOVAL, 2012).

Diante destas extensões da distribuição beta, surgem os modelos de regressão beta inflacionados (OSPINA; FERRARI, 2012) e beta inflacionados truncados (PEREIRA; BOTTER; SANDOVAL, 2013). Na análise de diagnóstico destes modelos, é comum que se utilize o resíduo quantílico aleatorizado (DUNN; SMYTH, 1996), o que é útil para verificar falhas na especificação do modelo. No entanto, pode falhar para identificar *outliers*, conforme observado por Pereira *et al.* (2020) no contexto dos modelos de regressão beta inflacionados em zero. O resíduo quantílico aleatorizado apresenta limitações na detecção destes pontos cujos valores se aproximam da componente discreta. O mesmo ocorre para os modelos beta inflacionados em dois e três pontos. Pereira *et al.* (2020) propuseram um resíduo com boas propriedades para detecção *outliers*, sobretudo quando não são identificados pelo resíduo quantílico aleatorizado. Neste trabalho, foi introduzida uma generalização do resíduo desses autores, abrangendo os modelos de regressão beta inflacionados em dois e três pontos, mas que também pode ser facilmente estendida para

mais pontos de inflação.

Além da análise de diagnóstico, outro problema comum no âmbito da análise de regressão consiste na seleção de modelos. Na prática, existem inúmeros modelos teóricos que podem ser ajustados, variando a distribuição da variável resposta, a quantidade de variáveis preditoras e a função de ligação, por exemplo. Em geral, não existe um modelo real que represente a relação entre a variável resposta e as preditoras. Portanto, é necessário utilizar de critérios de seleção, visando eleger o melhor modelo teórico dentre um conjunto de candidatos.

De um modo geral, os métodos para comparação de modelos podem ser divididos entre abordagens baseadas na informação e processos tradicionais baseados em testes de hipóteses (LEWIS; BUTLER; GILBERT, 2011). Muitos trabalhos versam sobre o uso de critérios de seleção de modelos baseados em estatísticas de testes, porém se restringem ao problema de seleção de variáveis. Por outro lado, testes de qualidade do ajuste são propostos para avaliar as premissas sobre a distribuição da variável dependente. Devido à propriedade do resíduo quantílico ser assintoticamente Normal padrão distribuído, quando os parâmetros do modelo são consistentemente estimados, serão introduzidos nesse trabalho três critérios de seleção de modelos por meio de testes de bondade do ajuste com o uso do resíduo quantílico, avaliando o desempenho relativo destes métodos em um contexto específico de seleção da distribuição da variável resposta em modelos aditivos generalizados para localização, escala e forma (GAMLSS).

1.1 Organização da tese

A tese está organizada da seguinte forma. O Capítulo 2 introduz um resíduo assintoticamente Normal padrão distribuído, que pode ser usado para qualquer modelo de regressão circular-linear paramétrico. Para avaliar o comportamento da distribuição de probabilidade do resíduo proposto em pequenas amostras, foram realizados estudos de simulação Monte Carlo. Para estudar o comportamento desse resíduo, dois modelos de regressão são introduzidos, e duas aplicações são usadas para mostrar que o resíduo proposto pode detectar erros de especificação nos modelos ajustados. O Capítulo 3 introduz a extensão do resíduo quantílico para detecção de *outliers* em modelos de regressão beta inflacionados em dois e três pontos. São realizados estudos de simulação Monte Carlo a fim de comparar a distribuição desses resíduos nas caudas com a distribuição Normal padrão. O Capítulo 4 propõe três critérios de seleção de modelos com o uso do resíduo quantílico, para seleção da distribuição da variável resposta em modelos GAMLSS. Estudos de simulação Monte Carlo foram realizados com o intuito de comparar o desempenho desses critérios a outros dois baseados na informação. Com o objetivo de enfatizar a importância da análise de diagnóstico do modelo selecionado, foram consideradas duas aplicações. Por fim, o Capítulo 5 apresenta as conclusões dessa tese e sugestões para trabalhos futuros.

CAPÍTULO

2

RESÍDUO QUANTÍLICO CIRCULAR

A regressão circular-linear é frequentemente utilizada para modelar a relação entre uma variável dependente circular e um conjunto de variáveis preditoras lineares. Para verificar a adequação do modelo, é desejável usar resíduos que são aproximadamente Normal padrão distribuídos. Contudo, a maioria dos resíduos usados em modelos de regressão circular não atendem esse requisito e são utilizados especialmente para identificação de *outliers*. Outros resíduos são limitados aos modelos de regressão von Mises. Neste capítulo é introduzido um resíduo assintoticamente Normal padrão distribuído, que pode ser usado para qualquer modelo de regressão circular-linear paramétrico. Estudos de simulação de Monte Carlo sugerem que a distribuição deste resíduo é bem aproximada pela distribuição Normal padrão mesmo em pequenas amostras. Para estudar o comportamento desse resíduo, dois modelos de regressão são introduzidos, e duas aplicações são usadas para mostrar que o resíduo proposto pode detectar erros de especificação nos modelos ajustados.

2.1 Introdução a dados circulares

Diversas áreas do conhecimento possuem interesse em fenômenos de natureza direcional ou periódica, cujas observações - denominadas *dados direcionais* - são, na prática, comumente tratadas em espaços bi ou tridimensionais. Alguns exemplos frequentes são:

1. meteorologia: orientação do vento (LANG *et al.*, 2020), horário que ocorrem tempestades;
2. oceanografia: direção de correntes marinhas;
3. geologia: orientação da magnetização remanescente de uma rocha;
4. biologia: orientação das aves durante a migração, movimento dos animais em resposta a estímulos (RODRÍGUEZ; NÚÑEZ-ANTONIO; ESCARELA, 2020);

5. medicina: horário de internação na UTI de um determinado hospital, ângulo das regiões dos olhos humanos (ABUZAID, 2020).

Medidas angulares bidimensionais podem ser geometricamente representadas por ângulos ou pontos sobre a circunferência de um círculo unitário, com relação a uma direção zero (origem) e um sentido de orientação preestabelecidos. À vista disso, essas observações são também chamadas de *dados circulares*. Analogamente, direções em um espaço de três dimensões podem ser caracterizadas por meio de uma esfera unitária, sendo então denominados como *dados esféricos*. Para além dessas dimensões, dados na forma angular podem ser configurados por pontos sobre a superfície de uma hiperesfera. É válido ressaltar que não há um consenso sobre a escolha da direção zero; frequentemente, para o caso circular, utiliza-se o sentido anti-horário com origem no 0^o ou $-p$.

Note que, tanto as superfícies esféricas quanto as circunferências constituem espaços limitados e fechados, cujo conceito de origem não é bem definido. Em decorrência disso, dados direcionais devem ser analisados conforme técnicas apropriadas, que, em geral, diferem da análise estatística para dados na reta (GOULD, 1969). O uso de medidas lineares no contexto direcional provoca, dentre outras desvantagens, respostas sensíveis a escolha da direção zero (MARDIA, 1972).

Considere, por exemplo, uma amostra de tamanho 2 mensurada com relação ao eixo das abcissas, cujos ângulos observados são 15^o e 345^o. É intuitivo pensar que a direção média esteja em torno do 0^o, e que o desvio padrão não seja muito alto, uma vez que as observações são próximas entre si (como mostra a Figura 2.1(a)). Entretanto, nesse caso, a média aritmética amostral equivale a 180^o, e o desvio padrão amostral - calculado utilizando a definição não viesada para dados em R^n - a $233,34^o$, aproximadamente. Contudo, se o eixo das ordenadas representar a direção zero, os dados se tornam 285^o e 255^o, respectivamente (como visto na Figura 2.1(b)). Nesse novo sistema de coordenadas, a média aritmética e o desvio padrão amostrais resultam em 270^o e aproximadamente $21,21^o$, nessa ordem. Portanto, além de serem medidas voláteis, observa-se por meio da Figura 2.1 que, a depender da origem escolhida, a média aritmética não corresponde a uma medida de "centro", assim como o desvio padrão amostral (linear) não descreve o grau de dispersão dos dados.

2.1.1 Notação e representações gráficas

Uma amostra de n dados circulares pode ser representada de diferentes formas, tais como:

1. $q_1, \ldots, q_n$, com $q_j \in G$, $\forall j \in \{1, \ldots, n\}$, em que G representa um arco de comprimento $2p$. Nesse contexto, q_j refere-se ao ângulo entre o eixo x e o *j-ésimo* ponto observado, considerando o sentido anti-horário de rotação.

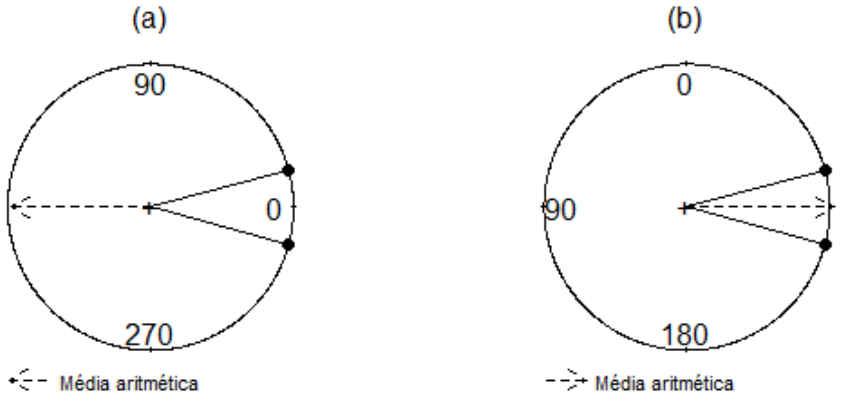

Figura 2.1 – Posição correspondente a média aritmética, considerando como direção zero: o eixo das abcissas (a) e o eixo das ordenadas (b).

2. $(1, q_1), \ldots, (1, q_n)$, utilizando coordenadas polares.

3. $\mathbf{x}_1, \ldots, \mathbf{x}_n$, em que $\mathbf{x}_j$ corresponde ao vetor unitário $\mathbf{x}_j = (\cos q_j, \text{ sen } q_j)^T, \forall j \in \{1, \ldots, n\}$.

4. $Z_1, \ldots, Z_n$, sendo Z_j na forma de números complexos, obtido por meio da fórmula de Euler $Z_j = e^{iq_j} = \cos q_j + i \text{ sen } q_j, \forall j \in \{1, \ldots, n\}$.

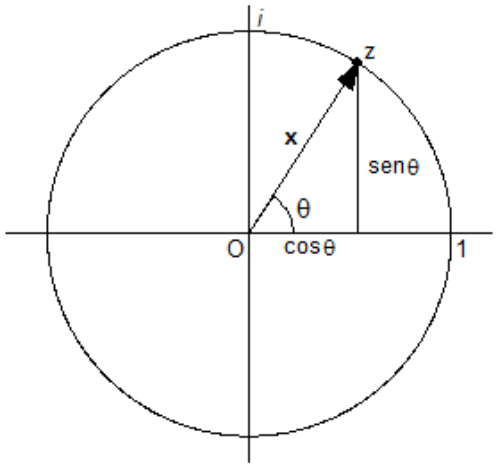

Figura 2.2 – Formas de representar um dado circular.

A Figura 2.2 exibe algumas das representações mencionadas anteriormente, considerando uma amostra de tamanho $n = 1$. Note que o vetor unitário $\mathbf{x} = (\cos q, \text{ sen } q)^T$ corresponde ao número $Z = \cos q + i \text{ sen } q$ no plano complexo, bem como, ao ângulo q mensurado com relação a origem e sentido anti-horário de rotação.

A maneira mais simples e usual de representar graficamente uma amostra circular se dá por meio do gráfico de dispersão no círculo, em que cada observação é vista como um ponto

na circunferência do círculo unitário. A Figura 2.3(a) ilustra esse método para os dados da Tabela 2.1, obtidos em Luschi *et al.* (2001). Os registros consistem nas direções resultantes de 10 tartarugas-verdes (*Chelonia mydas*), ao se moverem em direção a sua ilha de Ascensão. Nesse experimento, estabeleceu-se como direção zero o Norte (equivalente ao eixo das ordenadas) e sentido horário de rotação.

Tabela 2.1 – Direções resultantes da nidificação de 10 tartarugas-verdes.

20^o	40^o	285^o	296^o	296^o	299^o	303^o	308^o	314^o	326^o

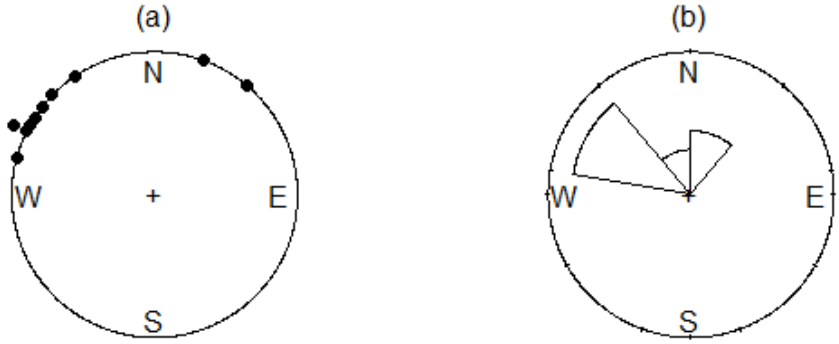

Figura 2.3 – Diagrama de dispersão circular (a) e *rose diagram* (b) para os dados da Tabela 2.1.

Alternativamente, pode-se utilizar um gráfico - denominado *rose diagram* (MARDIA, 1972) - similar ao histograma para dados na reta real, como mostra a Figura 2.3(b). Ao invés de barras, o *rose diagram* compõe-se de setores cuja área é proporcional à frequência observada na respectiva classe.

2.1.2 Estatísticas descritivas no círculo

Seja $q_1, \ldots, q_n$ os ângulos correspondentes aos respectivos vetores unitários $\mathbf{x}_1, \ldots, \mathbf{x}_n$. A direção média (ou ângulo médio), $\bar{q}$, é definido como a direção do vetor resultante $\bar{\mathbf{x}}$, cuja coordenada cartesiana é dada por

$$(\bar{C}, \bar{S}) = \left(\frac{\sum_{j=1}^{n} \cos q_j}{n}, \frac{\sum_{j=1}^{n} \operatorname{sen} q_j}{n} \right).$$

Dessa forma, $\bar{q}$ pode ser entendido como o ângulo entre o eixo das abscissas e o vetor $(\bar{C}, \bar{S})$, obtido por meio de

$$\bar{q} = Arg\left\{ \frac{1}{n} \sum_{j=1}^{n} \cos q_j + i \frac{1}{n} \sum_{j=1}^{n} \operatorname{sen} q_j \right\},$$

ou como solução do sistema de equações

$$\begin{cases} \bar{C} = \bar{R}\cos\bar{q}, \\ \bar{S} = \bar{R}\,\text{sen}\,\bar{q}, \end{cases}$$

em que $\bar{R}$ representa o comprimento do vetor resultante $\bar{\mathbf{x}}$. Portanto

$$\bar{R} = \sqrt{\bar{C}^2 + \bar{S}^2}.$$

A quantidade $\bar{R}$, denominada comprimento resultante da média, satisfaz $0 \leq \bar{R} \leq 1$, uma vez que os vetores $\mathbf{x}_1, \ldots, \mathbf{x}_n$ são unitários, e pode ser utilizada como uma medida de concentração dos dados. Nesse caso, quanto mais dispersas as direções estão em relação ao ângulo médio, mais próximo $\bar{R}$ estará de 0. Sendo assim, quando os dados estão distribuídos uniformemente no círculo, a concentração é mínima. A partir de $\bar{R}$, pode-se estabelecer uma medida de dispersão simples e útil para comparações com dados na reta,

$$V = 1 - \bar{R},$$

cognominada variância circular amostral. Contudo, diferentemente do âmbito linear, define-se o desvio padrão circular amostral como

$$u = \sqrt{-2\log(1-V)}.$$

Para elucidar tais medidas, considere os dados da Tabela 2.1, em que $n = 10$. Calculando $\cos q_j$ e $\text{sen}\, q_j$, $j = 1, \ldots, 10$, obtém-se

Tabela 2.2 – Coordenadas cartesianas para os dados da Tabela 2.1.

q_j	$\cos q_j$	$\text{sen}\, q_j$
20^o	0,9397	0,3420
40^o	0,7660	0,6428
285^o	0,2588	-0,9659
296^o	0,4384	-0,8988
296^o	0,4384	-0,8988
299^o	0,4848	-0,8746
303^o	0,5446	-0,8387
308^o	0,6157	-0,7880
314^o	0,6947	-0,7193
326^o	0,8290	-0,5592
Média	0,6010	-0,5559

Portanto,

$$(\bar{C}, \bar{S}) \simeq (0,60,\ -0,56), \quad \bar{q} \approx 317^o \quad \text{e} \quad \bar{R} \approx 0,8187,$$

consequentemente

$$V \approx 0,1813 \quad \text{e} \quad u = 0,6326.$$

Observe que $\bar{R} \approx 0,8187$ está mais próximo de 1, indicando que as direções estão, de certa forma, concentradas em relação a média. A Figura 2.4(a) ilustra a direção média $\bar{q} \approx 317^o$, indicada pela seta, bem como seu comprimento resultante.

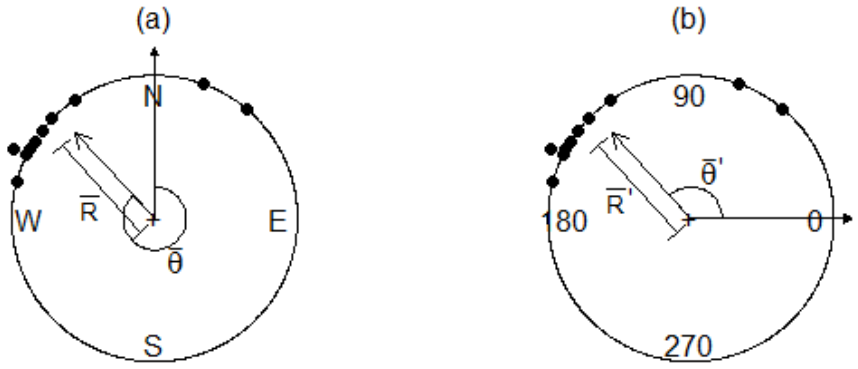

Figura 2.4 – Direção média e comprimento resultante da média, referentes aos dados da Tabela 2.1 (a) e Tabela 2.3 (b).

Para fins de comparação, considere os dados da Tabela 2.1 transformados para o sistema com origem no eixo das abcissas e sentido anti-horário de rotação, exibidos na Tabela 2.3. Analogamente, obtém-se

$$(\bar{C}', \bar{S}') \simeq (0,60,\ -0,56), \quad \bar{q}' \approx 133^o \quad \text{e} \quad \bar{R}' \approx 0,8187, \quad V' \approx 0,1813 \quad \text{e} \quad u' = 0,6326.$$

Tabela 2.3 – Direções resultantes da nidificação de 10 tartarugas-verdes, mensuradas com relação a origem no eixo X e sentido de rotação anti-horário.

70^o	50^o	165^o	154^o	154^o	151^o	147^o	142^o	136^o	124^o

Note que os valores de $\bar{R}$, V e u não se alteram conforme a direção zero é modificada. Observe também que, $\bar{q} = 317^o$ - considerando origem no Norte e sentido horário de rotação - e $\bar{q}' = 133^o$ - com origem no eixo X e sentido de rotação anti-horário - equivalem ao mesmo ponto sobre a circunferência (como ilustra a Figura 2.4). Isto é, diferentemente das medidas lineares, tais estatísticas não são sensíveis à escolha da direção zero.

2.1.2.1 Direção mediana amostral

Considere uma amostra circular $q_1, \ldots, q_n$, equivalente aos pontos $P_1, \ldots, P_n$ na circunferência do círculo unitário. De acordo com Mardia (1972), a direção mediana $\tilde{q}$ corresponde ao ponto P que satisfaz:

1. O diâmetro $\overline{PQ}$ divide o círculo em dois semi-círculos com igual número de observações;

2. A maioria dos dados observados está mais próxima de P do que da anti-mediana Q.

Seja, por exemplo, uma amostra de tamanho $n = 9$, cujos ângulos observados são 43^o, 45^o, 52^o, 61^o, 75^o, 88^o, 88^o, 279^o, 357^o, medidos com relação ao eixo das abcissas e sentido anti-horário de rotação (MARDIA, 1972). Por meio da Figura 2.5(a), note que o ponto P, correspondente ao ângulo 52^o, satisfaz as condições anteriores e, portanto, $\tilde{q} = 52^o$. É válido ressaltar que $\tilde{q}$ não equivale a mediana calculada segundo a definição para dados lineares - que, nesse caso, resultaria em $\tilde{q}^* = 75^o$ -, em virtude da dificuldade em ordenar as observações, uma vez que o conceito de origem não é bem definido.

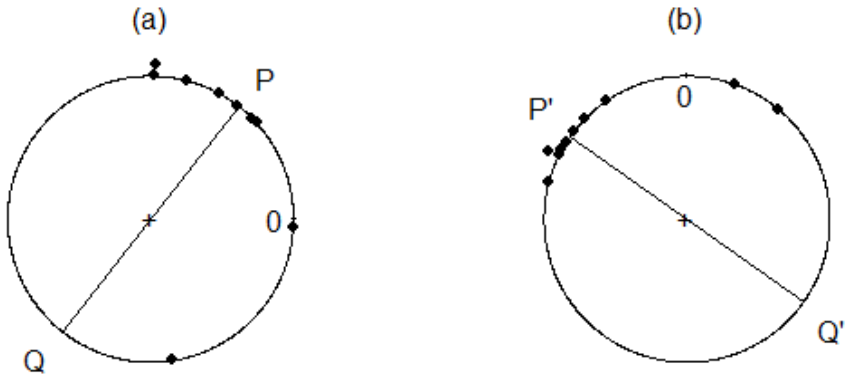

Figura 2.5 – Direção mediana para uma amostra de tamanho ímpar, retirada de Mardia (1972) (a), e para dos dados da Tabela 2.1, em que n é par (b).

Tal como ocorre com a mediana para dados na reta, quando n é ímpar, a mediana circular amostral coincide com uma das direções observadas. No entanto, se n é par, considera-se o ponto médio de duas observações adjacentes apropriadas (OTIENO, 2002). Desse modo, para os dados da Tabela 2.1, tem-se que a mediana está compreendida entre os pontos 303^o e 308^o, portanto $\tilde{q}' = 305,5^o$, como mostrado na Figura 2.5(b).

Novamente, utilizando a definição adequada para dados no espaço euclidiano usual, tem-se que a mediana para a amostra da Tabela 2.1 consiste no ângulo $297,5^o$, enquanto que para as observações transformadas (Tabela 2.3), equivale a $144,5^o$. Isso posto, assim como as medidas anteriores, a mediana amostral (linear) revela-se inapropriada para dados circulares.

2.1.3 Função distribuição

As distribuições circulares são caracterizadas por concentrar a probabilidade total sobre a circunferência de um círculo unitário. Dessa forma, o suporte de uma variável aleatória circular q é descrito por um arco predefinido de comprimento $2p$. Usualmente utiliza-se $[0, 2p)$ ou $[-p, p)$. Além disso, faz-se necessário estabelecer uma direção de origem e sentido de orientação.

É válido ressaltar que, por convenção, utiliza-se letras gregas minúsculas para representar tanto variáveis aleatórias circulares ainda não observadas, quanto seus valores já observados. Os parâmetros da população de interesse também são simbolizados, em geral, por meio de letras gregas.

Isso posto, seja q um ângulo aleatório com suporte em $[w_0, w_0+2p)$, $w_0 \in \mathrm{R}$, e sentido de orientação anti-horário. A função distribuição acumulada circular F é definida satisfazendo (MARDIA; JUPP, 2009):

$$F(q_0) = \mathrm{P}(w_0 \leq q \leq q_0),\ w_0 \leq q_0 < w_0 + 2p \tag{2.1}$$

e

$$F(q_0 + 2p) - F(q_0) = 1,\ -¥ < q_0 < ¥. \tag{2.2}$$

Observe que a condição (2.2) garante probabilidade igual a 1 para qualquer arco de comprimento $2p$ sobre o círculo unitário. Para $w_0 + 2pk \leq q_0 < w_0 + 2p(k+1)$, com $k = \pm 1, \pm 2, \ldots$, tem-se

1. $w_0 \leq q_0 - 2pk < w_0 + 2p \overset{(2.1)}{\Rightarrow} F(q_0 - 2pk) = \mathrm{P}(w_0 \leq q \leq q_0 - 2pk)$;
2. $F(q_0 - 2pk) = F(q_0) - k$, pois

$$\begin{aligned} F(q_0 - 2pk) &\overset{(2.2)}{=} F(q_0 - 2pk + 2p) - 1 \\ &\overset{(2.2)}{=} F(q_0 - 2pk + 2p + 2p) - 1 - 1 \\ &\vdots \\ &\overset{(2.2)}{=} F(q_0 - 2pk + 2pk) - k \\ &= F(q_0) - k. \end{aligned}$$

Portanto,

$$F(q_0) = k + \mathrm{P}(w_0 \leq q \leq q_0 - 2pk). \tag{2.3}$$

Dessa forma, diferentemente das funções distribuição definidas no espaço euclidiano usual,

$$\lim_{q_0 \to ¥} F(q_0) = ¥ \quad \text{e} \quad \lim_{q_0 \to -¥} F(q_0) = -¥.$$

Por definição, $F(w_0) = 0$ e $F(w_0 + 2p) = 1$. Para $a \leq b \leq a + 2p$,

$$\mathrm{P}(a \leq q \leq b) = F(b) - F(a) = \int_a^b dF(q_0), \tag{2.4}$$

considerando a integral de Lebesgue-Stieltjes (MARDIA; JUPP, 2009). Note que, apesar de F depender da direção de origem w_0, $F(b) - F(a)$ independe dessa escolha. À vista disso, a

função densidade de probabilidade (*f.d.p.*) f é definida baseando-se em $F(b)-F(a)$. Se F é absolutamente contínua, f existe e é tal que

$$\int_a^b f(q)\,dq = F(b)-F(a),\ -¥ < a \leq b < ¥, \tag{2.5}$$

satisfazendo as seguintes propriedades (JAMMALAMADAKA; SENGUPTA, 2001):

(i) $f(q) \geq 0$;

(ii) $\int_G f(q)\,dq = 1$;

(iii) $f(q) = f(q + k\cdot 2p),\ \forall\, k \in Z$.

2.1.4 Função Característica

Seja q uma variável aleatória circular com distribuição $F(q)$. A função característica de q é definida por

$$j_p = E[e^{ipq}] = \int_0^{2p} e^{ipq}\,dF(q),\quad p = 0, \pm1, \pm2, \ldots. \tag{2.6}$$

Observe que $j_p : Z \longmapsto C$, diferentemente do que ocorre no âmbito das variáveis aleatórias com suporte nos reais, visto que a variável aleatória q é periódica e, portanto,

$$E\left[e^{ipq}\right] = \int_0^{2p} e^{ipq}\,dF(q) \overset{(2.2)}{=} \int_0^{2p} e^{ip(q+2p)}\,dF(q) = E\left[e^{ip(q+2p)}\right]$$

$$\Rightarrow j_p = e^{ip2p} j_p,$$

que ocorre quando $j_p = 0$ ou $e^{ip2p} = 1$, isto é, apenas em valores inteiros de p. Além disso, os coeficientes de Fourier de F, j_p, satisfazem as seguintes propriedades (MARDIA; JUPP, 2009):

(i) $j_0 = 1$;

(ii) $\bar{j}_p = j_{-p}$, em que $\bar{j}_p$ representa o complexo conjugado de j_p;

(iii) $|j_p| \leq 1$.

2.1.5 Momentos Trigonométricos

O p-ésimo momento trigonométrico de q corresponde ao valor de sua função característica aplicada em p (JAMMALAMADAKA; SENGUPTA, 2001), que pode ser descrito, também, por

$$j_p = a_p + ib_p,$$

em que

$$a_p = E[\cos(pq)] = \int_{w_0}^{w_0+2p} \cos(pq)\,dF(q) \tag{2.7}$$

e

$$b_p = \mathrm{E}[\,\mathrm{sen}(pq)] = \int_{w_0}^{w_0+2p} \mathrm{sen}(pq)\, dF(q). \tag{2.8}$$

Em particular, para $p = 1$, tem-se $\underset{\sim}{j}_1 = (r, m)$, em coordenadas polares. Esse primeiro momento trigonométrico proporciona mais informação do que o primeiro momento ordinário de uma variável aleatória com suporte nos reais, uma vez que $m = \arctan^*\left(\frac{b_1}{a_1}\right)$ representa a média populacional de q e $r = \sqrt{a_1^2 + b_1^2}$ corresponde a uma medida de concentração teórica, denominada comprimento resultante da média. A função $\arctan^*$ é definida como

$$\arctan^*\left(\frac{b}{a}\right) = \begin{cases} \arctan\left(\frac{b}{a}\right) + w_0, & \text{se } a > 0 \text{ e } b \geq 0, \\ \frac{p}{2} + w_0, & \text{se } a = 0 \text{ e } b > 0, \\ \arctan\left(\frac{b}{a}\right) + w_0 + p, & \text{se } a < 0, \\ \frac{3p}{2} + w_0, & \text{se } a = 0 \text{ e } b < 0, \\ \arctan\left(\frac{b}{a}\right) + w_0 + 2p, & \text{se } a > 0 \text{ e } b < 0, \\ \text{indefinido}, & \text{se } a = 0 \text{ e } b = 0, \end{cases} \tag{2.9}$$

com o intuito de "corrigir" a função *arco tangente*, uma vez que esta fornece valores no intervalo $\left(-\frac{p}{2}, \frac{p}{2}\right)$. A definição (2.9) leva em consideração os sinais de a e b para gerar o inverso único e correto em $[w_0, w_0 + 2p)$. Note que $\arctan^*\left(\frac{b}{a}\right)$ é definida mesmo quando a igual a zero, desde que b seja diferente de zero (JAMMALAMADAKA; SENGUPTA, 2001).

2.1.6 Mediana Populacional

Seja q uma variável aleatória circular com densidade $f(q)$. De maneira análoga a definição de mediana amostral (Subseção 2.1.2.1), a mediana direcional populacional corresponde ao x_0 que soluciona

$$\int_{x_0}^{x_0+p} f(q)\, dq = \int_{x_0+p}^{x_0+2p} f(q)\, dq = \frac{1}{2}, \tag{2.10}$$

satisfazendo $f(x_0) > f(x_0 + p)$ (MARDIA, 1972).

2.1.7 Distribuições circulares

Muitas distribuições circulares são obtidas a partir de transformações de modelos probabilísticos da reta real ou de vetores aleatórios bidimensionais, bem como, por meio de analogias no círculo de importantes caracterizações univariadas. Os mecanismos utilizados fundamentam-se, por exemplo, no método da máxima entropia, em distribuições deslocadas (convertendo um vetor aleatório linear bivariado) ou em projeções estereográficas, como descreve Jammalamadaka e Sengupta (2001).

Em particular, destaca-se a distribuição von Mises, derivada de maneira similar a distribuição Normal para dados na reta, por meio da caracterização de propriedades. Esse modelo

possui uma vasta utilização em problemas aplicados (JAMMALAMADAKA; SENGUPTA, 2001), e desempenha uma grande importância na inferência estatística circular (MARDIA, 1972), semelhantemente a distribuição Normal no contexto linear. Em decorrência disso, vários autores propuseram generalizações e extensões dessa distribuição, como por exemplo: von Mises generalizada (GATTO; JAMMALAMADAKA, 2007), *sine-skewed* von Mises (UMBACH; JAMMALAMADAKA, 2009) e von Mises assimétrica generalizada (KIM; SENGUPTA, 2012).

Outros modelos probabilísticos são obtidos por meio de mapeamentos do tipo *muitos-para-um* (JAMMALAMADAKA; SENGUPTA, 2001), originando as chamadas distribuições arqueadas (*wrapped distributions*). O método consiste em *enrolar* a função densidade de probabilidade de uma variável com suporte na reta, em torno do círculo unitário, acumulando os valores das funções densidades de probabilidade das regiões sobrepostas. Dessa forma, seja X uma variável aleatória linear com densidade $f(x)$, define-se

$$q = X \,(\text{mod } 2p), \tag{2.11}$$

em que $X \,(\text{mod } 2p)$ representa o resto da divisão de X por $2p$. Portanto, a densidade circular de q é calculada por

$$g(q) = \sum_{m=-\infty}^{\infty} f(q + 2pm), \quad 0 \leq q < 2p. \tag{2.12}$$

Em algumas situações, a depender da distribuição de X, utiliza-se uma representação alternativa da densidade g, obtida por meio das seguintes propriedades (MARDIA, 1972):

a) Se $f_X(t)$ corresponde a função característica de X, então o momento trigonométrico de ordem p, j_p, para a distribuição circular arqueada equivale a

$$j_p = f_X(p).$$

b) Se $f_X(t)$ é integrável, então

$$g(q) = \sum_{m=-\infty}^{\infty} f(q + 2pm) = \frac{1}{2p}\left[1 + 2\sum_{p=1}^{\infty}\Big(a_p \cos(pq) + b_p \,\text{sen}(pq)\Big)\right], \tag{2.13}$$

em que $f_X(p) = a_p + ib_p$.

São exemplos de distribuições obtidas por meio do método supramencionado: *wrapped* Normal, *wrapped* Cauchy, *wrapped* t de Student e *wrapped* Exponencial (MARDIA, 1972; PEWSEY; LEWIS; JONES, 2007). É válido ressaltar que essa metodologia não se restringe ao uso de variáveis contínuas com suporte na reta real. De forma análoga a descrita anteriormente, pode-se, por exemplo, obter uma distribuição arqueada a partir da distribuição Poisson (MARDIA, 1972).

Nas seções seguintes, serão detalhadas as distribuições utilizadas para avaliar o resíduo proposto neste trabalho. São elas: von Mises, visto que é considerada a distribuição mais

importante no contexto de dados circulares (MARDIA, 1975); *wrapped* Cauchy, em razão de seu uso abrangente, uma vez que surge como uma alternativa à distribuição von Mises para dados simétricos no círculo (KENT; TYLER, 1988); e *sine-skewed* von Mises (UMBACH; JAMMALAMADAKA, 2009), devido à sua capacidade em descrever dados circulares não simétricos, diferentemente das distribuições anteriores.

2.1.7.1 Distribuição von Mises

Mises (1918) introduziu a distribuição que recebe seu nome em um problema físico, ao estudar desvios de pesos atômicos em relação a valores inteiros. O autor derivou a distribuição por meio do princípio da máxima verossimilhança, supondo unicamente que a média é o valor mais provável, baseando-se, portanto, na abordagem utilizada por Gauss para derivar a distribuição Normal. Por esse motivo e, também, devido às similaridades com a distribuição Normal para dados na reta real, Gumbel, Greenwood e Durand (1953) nomearam-a como distribuição Normal Circular.

Em virtude dessa analogia, a distribuição von Mises vem sendo exaustivamente estudada e aplicada por diversos autores, em diferentes áreas do conhecimento. Stephens (1982), por exemplo, utilizou esse modelo probabilístico para analisar proporções contínuas de tempo, considerando o tempo gasto por 130 alunos em oito atividades durante um dia. Gabarda e Cristóbal (2012) propuseram uma metodologia de avaliação da qualidade de imagens, por meio da distribuição Normal Circular. Harris e Johnson (2007) fizeram uso dessa distribuição para caracterizar a orientação de flocos em painéis de fibras orientadas (*flakeboard*). Outras referências de aplicações, sobretudo no âmbito da zoologia e geologia, podem ser encontradas em Batschelet (1965).

Já na esfera teórica, destacam-se alguns trabalhos importantes, como os de Gumbel, Greenwood e Durand (1953) e Greenwood, Durand *et al.* (1955). Esses autores forneceram tabelas que facilitam os cálculos da estimativa de máxima verossimilhança de um dos parâmetros da von Mises, bem como, tabelas de significância para testes de hipótese. Vários testes foram sugeridos por Watson e Williams (1956), tanto para comparar vetores de duas ou mais amostras, quanto seus parâmetros concentração. Best e Fisher (1979), por sua vez, forneceram um algoritmo capaz de gerar valores dessa distribuição, com base no método de aceitação-rejeitação, utilizando a densidade de uma *wrapped* Cauchy como função envelope.

Isso posto, uma variável aleatória circular q tem distribuição von Mises, denotada por $q \sim vM(m,k)$, se sua função densidade de probabilidade for descrita como

$$f(q;m,k) = \frac{1}{2p\,I_0(k)} \exp\{k\cos(q-m)\}, \tag{2.14}$$

em que $q \in [-p,p)$, $m \in [-p,p)$, $k > 0$, e $I_0(k)$ representa a função de Bessel modificada de

1^o tipo e ordem zero, definida como

$$I_0(k) = \sum_{r=0}^{\infty} \frac{1}{r!^2} \left(\frac{1}{2}k\right)^{2r}. \tag{2.15}$$

O parâmetro m equivale a média circular da variável q, enquanto k mensura a concentração dos dados, sendo assim um parâmetro de precisão. Quando $k \to 0$, a distribuição von Mises converge para uma uniforme circular, e quando $k \to \infty$, para uma distribuição degenerada no ponto $y = m$. Além disso, como ilustra a Figura 2.6, a distribuição é unimodal e simétrica em relação a direção m. Outras propriedades importantes dessa distribuição são mencionadas e demonstradas por Jammalamadaka e Sengupta (2001).

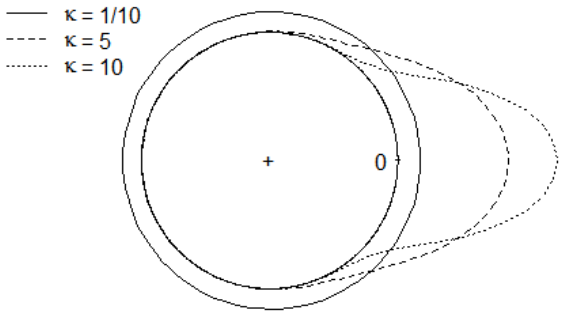

Figura 2.6 – Densidade da distribuição von Mises para $m = 0$ e $k = 0, 5$ e 10.

2.1.7.2 Distribuição sine-skewed von Mises

Umbach e Jammalamadaka (2009) propuseram uma ampla classe de modelos assimétricos, obtidos por meio de uma adaptação da abordagem de Azzalini (1985) para o contexto circular. Os autores apresentaram um caso particular dessa construção, derivado de perturbações com a função seno, o qual resultou em uma família de distribuições denominada *sine-skewed*. Abe e Pewsey (2011) apresentaram resultados gerais para qualquer distribuição pertencente à essa família, relativos a função distribuição, momentos trigonométricos, medidas circulares e sobre a estimativa de máxima verossimilhança dos parâmetros. Recentemente, Miyata, Shiohama e Abe (2019) provaram a identificabilidade destes modelos.

O método consiste em construir distribuições assimétrias a partir de uma densidade f_m simétrica conhecida e uma distribuição G, tal que G' exista e seja simétrica em relação a 0.

Teorema 2.1.1. Suponha f e g densidades circulares simétricas em relação a 0, definidas sobre o intervalo $[-p, p)$, e $G(q_0) = \int_{-p}^{q_0} g(q)dq$. Se d é uma função ímpar e periódica, tal que $|d(q)| \leq p$ e $d(q) = d(q + 2pk),\ \forall k \in Z$, então

$$f_m(q) = 2f(q-m)G\Big(d(q-m)\Big), \tag{2.16}$$

é uma densidade circular (UMBACH; JAMMALAMADAKA, 2009).

Em particular, os autores apresentaram uma versão assimétrica da distribuição von Mises, proveniente das escolhas $G(q) = \frac{p+q}{2p}$ (distribuição uniforme circular), e $d(q) = l\,p\text{sen}\,q$, cuja função densidade de probabilidade é dada por

$$f(q;m,k,l) = \frac{\exp\{k\cos(q-m)\}}{2pI_0(k)}(1+l\,\text{sen}(q-m)), \tag{2.17}$$

em que $q \in [-p,p)$, $m \in [-p,p)$, $k > 0$ e $l \in [-1,1]$. Abe e Pewsey (2011) nomearam esse modelo de distribuição *sine-skewed* von Mises, denotada aqui por $q \sim SSvM(m,k,l)$. Essa distribuição de probabilidade é um caso particular da distribuição k *sine-skewed* generalizada (RAD; BEKKER; ARASHI, 2020). Por outro lado, a distribuição von Mises é um caso particular da distribuição *sine-skewed* von Mises quando $l = 0$, e a assimetria da distribuição aumenta à medida que $|l|$ cresce (Figura 2.7).

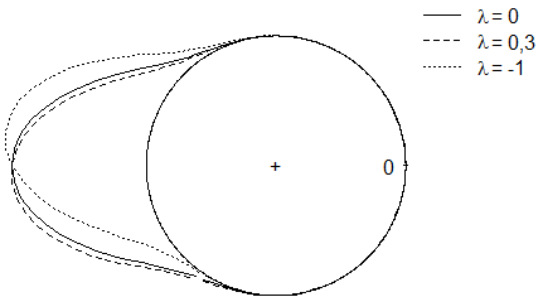

Figura 2.7 – Densidade da distribuição *sine-skewed* von Mises para $m = p$, $k = 7$ e $l = 0, \frac{3}{10}$ e -1.

Abe e Pewsey (2011) forneceram resultados mais específicos para essa distribuição, obtendo, por exemplo, os estimadores dos parâmetros pelo método dos momentos. Uma aplicação desse modelo pode ser encontrada em Umbach e Jammalamadaka (2009), referente a direção escolhida por formigas após receberem um estímulo. Os autores mostraram que a distribuição *sine-skewed* von Mises oferece um melhor ajuste aos dados, quando comparada a distribuição von Mises.

2.1.7.3 Distribuição wrapped Cauchy

A distribuição *wrapped* Cauchy, introduzida por Levy (1939), surge como uma importante alternativa ao modelo von Mises para representação de dados circulares simétricos. Schulgasser (1985), por exemplo, mostrou que essa distribuição é mais adequada para o estudo sobre a orientação das fibras de uma folha de papel fabricada à maquina, do que o modelo von Mises.

A distribuição *wrapped* Cauchy é obtida por meio de uma variável aleatória X com distribuição Cauchy, cuja densidade é dada por

$$f(x) = \frac{g}{pg^2 + p(x-m)^2}, \; -¥ < x < ¥, \tag{2.18}$$

em que $m \in \mathrm{R}$ e $g > 0$ são parâmetros de locação e forma, respectivamente.

Utilizando a equação (2.12), tem-se que a função densidade de probabilidade de uma variável q *wrapped* Cauchy distribuída, denotada como $q \sim WC(m,g)$, é dada por

$$g(q) = \sum_{m=-\infty}^{\infty} f(q+2pm) = \sum_{m=-\infty}^{\infty} \frac{g}{pg^2 + p(q+2pm-m)^2}, \tag{2.19}$$

que se reduz, com base em Jammalamadaka e Sengupta (2001) e na equação (2.13), a

$$g(q) = \frac{1}{2p}\left[1 + 2\sum_{p=1}^{\infty} e^{-gp}\cos\Big(p(q-m)\Big)\right] \tag{2.20}$$

$$= \frac{1}{2p}\frac{1-e^{-2g}}{1+e^{-2g}-2e^{-g}\cos(q-m)}, \quad -p \leq q < p, \tag{2.21}$$

em que m representa a direção média de q e g é um parâmetro de dispersão, com $Var[q] = 1-e^{-g}$. Neste caso, o comprimento resultante da média é dado por $r = e^{-g}$. Note que a distribuição *wrapped* Cauchy converge para uma uniforme no círculo, a medida que $r \to 0$. Assim como a von Mises, o modelo em questão é unimodal e simétrico em relação a m. A Figura 2.8 exibe a densidade da distribuição *wrapped* Cauchy, considerando diferentes valores para o parâmetro g.

Outra forma bastante útil para expressar a *f.d.p.* de uma variável *wrapped* Cauchy distribuída, se dá por meio das funções seno e cosseno hiperbólico, em que obtém-se (CAO *et al.*, 2014)

$$g(q) = \frac{1}{2p}\frac{\sinh(g)}{\cosh(g)-\cos(q-m)}, \tag{2.22}$$

como demonstrado no Apêndice A.1.

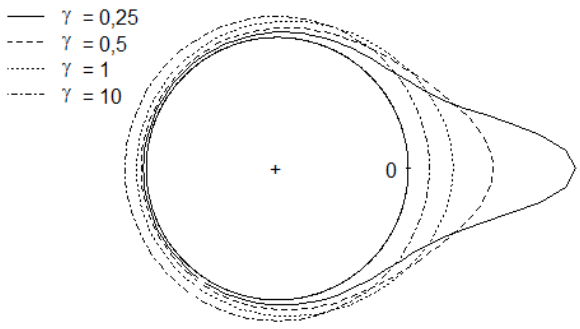

Figura 2.8 – Densidade da distribuição *wrapped* Cauchy para $m = 0$ e $g = \frac{1}{4}, \frac{1}{2}, 1$ e 10.

Como visto anteriormente, a distribuição *wrapped* Cauchy pode ser representada por meio de uma forma fechada de densidade. Em geral, o mesmo não ocorre para modelos *wrapped*, ocasionando, por exemplo, dificuldades na estimação via método da máxima verossimilhança (MMV). Kent e Tyler (1988) mostraram que para uma amostra de tamanho $n \geq 3$, o MMV para a distribuição *wrapped* Cauchy existe e é único.

2.2 Regressão Circular

Métodos de regressão circular possuem grande importância no estudo de dados bi ou multivariados, quando pelo menos uma das componentes tem natureza circular e há o interesse em entender a relação entre duas ou mais variáveis. Os modelos de regressão circular podem ser classificados levando em consideração o tipo de associação entre as variáveis, que podem ser: circular-circular, circular-linear e linear-circular. O primeiro caso refere-se aos cenários em que tanto a variável resposta quanto as preditoras possuem natureza circular. Rivest (1997), por exemplo, investigou a associação entre o movimento do solo durante um terremoto e a direção da descida mais íngreme. Por sua vez, a regressão circular-linear é utilizada quando há o interesse em predizer valores de uma variável circular a partir de valores de variáveis lineares. Gould (1969) analisou dados provenientes de um vetorcardiograma, por meio de um modelo cuja variável preditora é linear. Por último, tem-se a regressão linear-circular, que pode ser utilizada para investigar a associação entre uma variável linear e preditoras circulares. No contexto multivariado, Lund (1999) relacionou o tempo de desova de um peixe em particular, com duas características referentes à maré do ambiente, sendo uma delas também circular - tempo de maré baixa - e a outra linear - amplitude da maré baixa.

No que tange os modelos de regressão com resposta circular, convém-se utilizar a família de distribuição von Mises (FISHER; LEE, 1992), devido às similaridades com a distribuição Normal e suas principais propriedades de inferência. Nesse sentido, Gould (1969) apresentou um modelo de regressão múltipla com resposta circular von Mises distribuída e preditoras lineares, cujos parâmetros são ajustados por meio do método de máxima verossimilhança. Entretanto, Johnson e Wehrly (1978) verificaram a presença de máximos locais na função de probabilidade desses modelos, o que provoca problemas de não identificabilidade. Diante disso, os autores propuseram uma alternativa à abordagem de Gould, baseando-se em uma função distribuição marginal completamente especificada para a variável explicativa, que, nesse caso, é única. Por fim, Fisher e Lee (1992) generalizaram a metodologia de Johnson e Wehrly, por meio de uma função de ligação g que mapeia a reta real para o círculo. O modelo permite relacionar a direção média e a dispersão da variável resposta - ainda von Mises distribuída - a um conjunto de variáveis explicativas lineares.

Em relação aos modelos de Gould, Laycock (1975) afirma que o método de máxima verossimilhança é equivalente ao método de mínimos quadrados (MMQ). Contudo, Lund (1999) ressalta que a regressão via MMQ não é adequada no contexto circular, uma vez que a diferença de quadrados não é uma medida conveniente para o círculo. Dessa forma, o autor propõe uma metodologia análoga ao MMQ, adaptada para o contexto circular, cujas estimativas produzidas também são iguais àquelas obtidas pelo MMV, se a variável dependente for von Mises distribuída.

Para além da distribuição von Mises, Sarma e Jammalamadaka (1993) introduziram uma classe de modelos de regressão para dados bivariados, em que ambas as componentes possuem natureza circular. O método consiste em uma aproximação por polinômios trigonométricos,

e pode ser utilizado tanto de forma não paramétrica, quanto paramétrica, ao se especificar a distribuição da variável dependente condicionada à variável independente. Em particular, os autores apresentaram a forma exata do modelo para os casos paramétricos com a distribuição von Mises, *wrapped* Cauchy e *wrapped* Normal. Jammalamadaka e Lund (2006) utilizaram a metodologia proposta para investigar a relação entre a direção do vento e o horário em que a observação foi mensurada, considerando dados sobre a qualidade do ar do Texas.

Contudo, muito dos fenômenos periódicos ou direcionais não são necessariamente simétricos, portanto, surge também a necessidade por modelos que envolvam a assimetria quando presente. Note que os modelos de Sarma e Jammalamadaka se encaixam nessa conjuntura. As classes de modelos propostas por SenGupta, Kim e Arnold (2013), SenGupta e Kim (2015) e Kim e SenGupta (2016), também possibilitam o uso de distribuições assimétricas ou, ainda, assimétricas bimodais. Kim e Rifat (2019), por exemplo, utilizaram os modelos de SenGupta e Kim (2015), assumindo erros assimétricos por meio da distribuição von Mises generalizada (KIM; SENGUPTA, 2012).

Entetanto, como afirmado por Fisher e Lee (1992) e, mais recentemente, por Hall e Shen (2015), há uma certa negligência em estudos teóricos de modelos de regressão com resposta circular. Diante dessa limitação, sobretudo no âmbito de distribuições que diferem da von Mises, propõe-se, neste trabalho, duas extensões à metodologia de Fisher e Lee (1992), na qual abrangem as distribuições *sine-skewed* von Mises e *wrapped* Cauchy.

É válido destacar que a regressão circular não se restringe aos cenários em que a variável resposta é angular. No entanto, quando a variável dependente possui suporte na reta, os métodos de inferência estatística usuais (isto é, para dados lineares) são também válidos. Jammalamadaka e Lund (2006), por exemplo, utilizaram um modelo de regressão linear múltipla para prever o nível de ozônio, segundo um conjunto de preditoras lineares e circulares. As covariáveis circulares em questão - direção do vento e mês de medição - foram incorporadas ao modelo por meio das funções *seno* e *cosseno*. Devido a esse paralelo com os métodos lineares, este trabalho abordará apenas os modelos cuja variável dependente possui natureza circular, restringindo-se, ainda, aos casos em que as preditoras são lineares.

2.2.1 Modelo von Mises de regressão

Fisher e Lee (1992) propuseram três generalizações aos modelos de Johnson e Wehrly, assumindo que a variável resposta é von Mises distribuída. A primeira delas, denominada por *modelo de médias*, consiste em modelar a média direcional da variável dependente, em termos do vetor de covariáveis lineares. Gill e Hangartner (2010) consideram este modelo como o mais importante dentre os três apresentados por Fisher e Lee, e o utilizaram na análise de dados sobre terrorismo doméstico nos Estados Unidos. Diversas aplicações deste modelo podem ser encontradas nos campos da biologia e ecologia, como por exemplo, nos trabalhos de Borgioli *et al.* (1999), Fitzgerald e Taylor (2008) e Artes (2008). O segundo modelo apresentado corresponde

a uma extensão do *modelo de dispersão* de Johnson e Wehrly, que associa o parâmetro de concentração da von Mises à uma única covariável com suporte na reta. Hensz (2015) utilizou esse modelo para investigar os fatores ambientais que influenciam a rota de migração das andorinhas do ártico. Por fim, o terceiro modelo provém da combinação direta entre os dois primeiros e, portanto, foi denotado por *modelo misto*. Nesse caso, os dois parâmetros da distribuição von Mises são associados, conjuntamente, às variáveis explicativas. Para exemplificar seu uso, Fisher e Lee utilizaram os dados provenientes de um estudo com 31 pervincas azuis, modelando a direção do movimento em função da distância percorrida, após terem sido transplantadas para baixo da altura em que vivem normalmente.

Este trabalho se restringirá ao uso do modelo de médias. Portanto, considere $g = (q_1, q_2, ..., q_n)^T$ variáveis aleatórias circulares independentes, tais que $q_i \sim vM(m_i, k)$, $\forall i = 1, ..., n$. O modelo em questão relaciona cada média direcional m_i à um conjunto de p variáveis explicativas lineares, por meio de uma função de ligação g, de modo que

$$m_i = g(h_i), \tag{2.23}$$

em que $h_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}$. A função de ligação $g : \mathrm{R} \to [-p, p)$ é monótona e duplamente diferenciável, o que evita os problemas de não identificabilidade dos estimadores de máxima verossimilhança (EMV) nos modelos de Gould.

Uma outra abordagem possível para o modelo de médias von Mises, consiste em considerar

$$m_i = m + g(b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}), \tag{2.24}$$

cuja função g deve satisfazer $g(0) = 0$, a fim de garantir que o parâmetro m tenha interpretação de origem. É válido ressaltar que, na prática, raramente há o interesse em interpretar o parâmetro de origem e, portanto, não há perdas em se utilizar a primeira abordagem. Por meio de estudos preliminares, optou-se pela utilização do modelo (2.23), como feito por Medeiros, Ferrari e Lemonte (2017) e Maitra e Braun (2012), uma vez que traz menos problemas de convergência dos estimadores.

Fisher e Lee (1992) discutem algumas funções de ligação possíveis, dentre elas

$$g(t) = 2\arctan(t), \tag{2.25}$$

bastante usual, tanto em trabalhos teóricos (SOUZA; PAULA, 1999), quanto práticos (BORGIOLI *et al.*, 1999; GILL; HANGARTNER, 2010). Neste trabalho, propõe-se também a utilização de

$$g(t) = 2p\frac{e^t}{1+e^t} - p, \tag{2.26}$$

em que $g : \mathrm{R} \to [-p, p)$ satisfaz as condições de monotonicidade e diferenciabilidade, bem como, $g(0) = 0$.

Para o ajuste do modelo, Fisher e Lee sugerem o uso do MMV, utilizando o algoritmo de mínimos quadrados reponderados (IRLS) de Green (1984), uma vez que o método não oferece solução explícita e necessita de um procedimento iterativo. O logaritmo da função de verossimilhança é dado por

$$\ell(h_i, k|g) = -n\ln(2p\ I_0(k)) + k\sum_{i=1}^{n}\cos(q_i - g(h_i)). \tag{2.27}$$

No estudo em questão, as estimações serão obtidas pelo MMV, por meio do método Nelder-Mead (NELDER; MEAD, 1965), utilizado no âmbito de regressão circular por Kim (2009) e Hidayah (2018), por exemplo. Será utilizada a função *optim*, presente no pacote *stats* em ambiente R (R Core Team, 2021).

2.2.2 Modelo sine-skewed von Mises de regressão

Apesar de raramente dados circulares serem simétricos (KIM; RIFAT, 2019), não há muitos avanços na literatura acerca de modelos de regressão circular assimétricos. SenGupta, Kim e Arnold (2013) e Kim e Rifat (2019) propuseram modelos de regressão que abrangem erros cuja distribuição é assimétrica, ou assimétrica bimodal. No entanto, ambos os modelos versam sobre regressão inversa e se restringem ao caso circular-circular. Já no contexto da regressão linear-circular, SenGupta e Ugwuowo (2006) apresentaram um modelo de regressão multivariado assimétrico. Mais sobre os recentes avanços na estatística circular, sobretudo no âmbito de modelos de regressão, podem ser encontrados em Kim e SenGupta (2018) e Pewsey e García-Portugués (2020).

Em virtude da escassez de estudos sobre regressão circular assimétrica, principalmente no âmbito circular-linear, este trabalho propõe uma extensão ao modelo de médias von Mises, desenvolvido por Fisher e Lee (1992) e descrito na seção anterior. Então, sejam $q_1, \ldots, q_n$ variáveis aleatórias circulares independentes, tais que $q_i \sim SSvM(m_i, k, l)$. O modelo de regressão é dado por

$$m_i = g(h_i), \tag{2.28}$$

em que h_i e a função de ligação g são definidos de maneira análoga ao modelo de médias von Mises. Dessa forma, o logaritmo da função de verossimilhança do modelo pode ser expresso por

$$\ell(h_i, k, l|g) = -n\ln(2p\ I_0(k)) + k\sum_{i=1}^{n}\cos(q_i - g(h_i)) + \sum_{i=1}^{n}\ln(1 + l\,\mathrm{sen}(q_i - g(h_i))). \tag{2.29}$$

2.2.3 Modelo wrapped Cauchy de regressão

No que tange os modelos de regressão com resposta *wrapped* Cauchy, pode-se citar os trabalhos de Kato, Shimizu e Shieh (2008), Abuzaid e Allahham (2015) e Jha e Biswas (2017). No primeiro, os autores desenvolveram um modelo de regressão baseado na transformação de Mobius, supondo que o erro circular segue distribuição *wrapped* Cauchy. Jha e Biswas

(2017) generalizaram este modelo de Kato, Shimizu e Shieh para o caso múltiplo, considerando, também, outras distribuições para os erros circulares. Por fim, Abuzaid e Allahham propuseram um modelo de regressão simples, também considerando erros *wrapped* Cauchy distribuídos. Entretanto, todos os trabalhos mencionados se restringem aos casos onde tanto a variável resposta quanto a explicativa possuem natureza circular.

Devido à limitação de estudos no âmbito da regressão circular-linear com resposta *wapped* Cauchy distribuída, propõe-se estender o modelo de médias de Fisher e Lee (1992) também para esse caso. Então, sejam $q_1, \ldots, q_n$ variáveis aleatórias circulares independentes, tais que $q_i \sim WC(m_i, g)$. O modelo de regressão é dado por

$$m_i = g(h_i), \tag{2.30}$$

em que h_i e a função de ligação g são definidos de maneira análoga ao modelo de médias von Mises. Dessa forma, o logaritmo da função de verossimilhança do modelo, considerando a parametrização definida em (2.22), pode ser expresso por

$$\ell(h_i, g|g) = -n\ln(2p) + n\ln(\sinh(g)) - \overset{n}{\underset{i=1}{\mathring{a}}} \ln(\cosh(g) - \cos(q_i - g(h_i))). \tag{2.31}$$

2.3 Resíduos para dados circulares

A análise de diagnóstico consiste em uma etapa importante na regressão, cujos principais objetivos são: identificar pontos discrepantes e/ou influentes e avaliar possíveis afastamentos das suposições feitas para o ajuste do modelo. De início surge a análise de resíduos, que é útil tanto na identificação de pontos discrepantes, quanto na verificação das suposições do modelo. Os resíduos configuram-se como uma medida capaz de mensurar a discrepância entre o valor observado e o valor ajustado pelo modelo (COX; SNELL, 1968). É conveniente buscar por resíduos que sejam bem aproximados pela distribuição Normal padrão, visto que suas propriedades e comportamento são conhecidos, o que facilita a interpretação dos gráficos e as posteriores identificações de pontos discrepantes no modelo.

Devido às diferenças topológicas entre a reta real e o círculo, muitos dos resíduos existentes na literatura para dados lineares não se aplicam no contexto circular, sobretudo àqueles obtidos a partir da diferença $Y - m$ (sendo m a esperança da variável resposta linear Y). Contudo, há uma escassez de estudos relacionados a análise de diagnóstico em regressões circulares (LIU *et al.*, 2016). Grande parte dos trabalhos existentes versa apenas sobre diagnósticos de influência (LIU *et al.*, 2016; RAMBLI *et al.*, 2010) ou detecção de *outliers* (ABUZAID; HUSSIN; MOHAMED, 2013; RAMLEE *et al.*, 2020), negligenciando, de certa forma, a bondade de ajuste para a distribuição imposta à variável dependente.

Mardia (1972), por exemplo, propôs um resíduo baseado na distância circular (RAO, 1969), definido como

$$e_i^* = 1 - \cos\left(q_i - \hat{q}_i\right), \tag{2.32}$$

para a *i-ésima* observação. Note que $e_i^* \in [0,2]$ e, portanto, não pode ser utilizado na verificação da distribuição imposta aos erros (IBRAHIM, 2013), para os casos em que essa suposição existe. O mesmo ocorre com os resíduos baseados da distância circular definida por Jammalamadaka e Sengupta (2001), em que

$$e_i^{**} = p - \left| p - |q_i - \hat{q}_i| \right|, \tag{2.33}$$

com $e_i^{**} \in [0,p]$. Além disso, em ambos os casos, não parece razoável que os resíduos sejam bem aproximados pela distribuição Normal padrão, justamente pelo intervalo limitado em que se encontram.

Por outro lado, D'elia (2001) apresentou uma definição de resíduo com suporte no intervalo $[-p,p]$, o que possibilita a investigação da distribuição imposta no modelo. Maruotti *et al.* (2016), por exemplo, utilizaram esse resíduo para checar as características marginais do modelo ajustado, por meio de gráficos quantil-quantil. Abuzaid, Hussin e Mohamed (2008) também propuseram uma definição de resíduo circular, baseada na distância circular, capaz de avaliar a adequanção do modelo, bem como, detectar possíveis *outliers*. Entretanto, os autores restringiram-se ao modelo de regressão circular-circular desenvolvido por Hussin, Fieller e Stillman (2004). Por fim, Souza e Paula (2002) desenvolveram resíduos que, além de satisfazerem essa necessidade (verificar a aderência do modelo) , dispõem de uma importante propriedade: distribuição de probabilidade equivalente à distribuição Normal padrão. Estes resíduos consistem em uma padronização do resíduo *deviance*, mas limitam-se aos modelos de regressão von Mises.

Diante do exposto, este trabalho tem por finalidade propor um novo resíduo para regressões de resposta circular e preditoras lineares, cujo desempenho será comparado com alguns dos resíduos existentes na literatura (descritos nas seções seguintes). Bem como, por meio de estudos de simulação, será verificado a aproximação do resíduo proposto com a distribuição Normal padrão.

2.3.1 Resíduo deviance

Comumente, nos modelos lineares generalizados (MLG) (NELDER; WEDDERBURN, 1972), utiliza-se a função desvio como medida de avaliação da qualidade do ajuste, cuja definiçao é dada por $D(\underset{\sim}{q},\hat{\underset{\sim}{m}}) = \sum_{i=1}^{n} d_i^2$, em que

$$d_i = sgn(q_i - \hat{m}_i)\sqrt{2}\Big(\ell_i(q_i|\tilde{m}_i,k) - \ell_i(q_i|\hat{m}_i,k)\Big)^{1/2}, \tag{2.34}$$

sendo *sgn* a função que retorna o sinal de um número real, $\ell_i(q_i|.)$ a contribuição de q_i para o logaritmo da função de verossimilhança e $\tilde{m}_i$ e $\hat{m}_i$ os estimadores de máxima verossimilhança de m_i com base no modelo saturado e no modelo ajustado, respectivamente.

No contexto de dados circulares, Paula (1996) e Maitra (2012) empregaram o resíduo componente do desvio em modelos de dispersão com resposta von Mises. Mutwiri (2015) e Mutwiri (2016), por sua vez, fizeram uso desse resíduo nos modelos de média von Mises,

definidos em (2.24). Nesse último caso, como demonstrado no Apêndice A.2, o resíduo *deviance* é dado por

$$d_i = d(q_i|\hat{m}_i,\hat{k}) = sgn(q_i-\hat{m}_i)\sqrt{2\hat{k}}\Big(1-\cos(q_i-\hat{m}_i)\Big)^{1/2}, \tag{2.35}$$

que equivale a

$$d_i = sgn(q_i-\hat{m}_i)\,2\sqrt{\hat{k}}\,\text{sen}\left(\frac{q_i-\hat{m}_i}{2}\right), \tag{2.36}$$

com base em algumas relações trigonométricas (SOUZA; PAULA, 2002).

Para o modelo *sine-skewed* von Mises, proposto neste trabalho, obtém-se (ver Apêndice A.3)

$$d_i = d(q_i|\hat{m}_i,\hat{k},\hat{l}\,) = sgn(q_i-\hat{m}_i)\sqrt{2}\left\{\hat{k}\Big(1-\cos(q_i-\hat{m}_i)\Big) - \ln\Big[1+\hat{l}\,\text{sen}\,(q_i-\hat{m}_i)\Big]\right\}^{1/2}. \tag{2.37}$$

Entretanto, assim como ocorre na regressão beta (ESPINHEIRA; FERRARI; CRIBARI-NETO, 2008), para a qual a variável resposta também possui como suporte um intervalo limitado, $\ell_i(q_i|\tilde{m}_i,\hat{k},\hat{l}\,) - \ell_i(q_i|\hat{m}_i,\hat{k},\hat{l}\,)$ pode ser negativo, tornando impossível o cálculo do resíduo *deviance*. Para exemplificar, considere o caso em que $q_i = 2,43$, $\hat{m}_i = 2,42$, $\hat{k} = 4,57$ e $\hat{l} = 0,16$. Dessa forma, tem-se que

$$\begin{aligned}\ell_i(q_i|\tilde{m}_i,\hat{k},\hat{l}\,) - \ell_i(q_i|\hat{m}_i,\hat{k},\hat{l}\,) &= \hat{k}\Big(1-\cos(q_i-\hat{m}_i)\Big) - \ln\Big[1+\hat{l}\,\text{sen}\,(q_i-\hat{m}_i)\Big]\\ &= 4,57\Big(1-\cos(2,43-2,42)\Big) - \ln\Big[1+0,16\,\text{sen}\,(2,43-2,42)\Big]\\ &= -0,0014,\end{aligned}$$

inviabilizando, portanto, o cálculo de d_i nos modelos de regressão com resposta SSvM.

Por fim, para o modelo com resposta *wrapped Cauchy*, tem-se que (ver Apêndice A.4)

$$d_i = d(q_i|\hat{m}_i,\hat{g}) = sgn(q_i-\hat{m}_i)\sqrt{2}\left\{\ln\left(\frac{\cosh(\hat{g})-\cos(q_i-\hat{m}_i)}{\cosh(\hat{g})-1}\right)\right\}^{1/2}. \tag{2.38}$$

2.3.2 Resíduo d_i^*: correção tipo matriz hat

Souza e Paula (2002) apresentaram duas extensões do resíduo *deviance* para modelo de regressão von Mises, sendo a primeira delas uma correção da componente da função desvio, definida como

$$d_i^* = sgn(q_i-\hat{m}_i)\,2\sqrt{\hat{k}}\,\frac{\text{sen}\left(\frac{q_i-\hat{m}_i}{2}\right)}{(1-h_{ii}^*)^{1/2}}, \tag{2.39}$$

em que h_{ii}^* é o *i-ésimo* elemento da diagonal principal da matriz $\mathbf{H}^* = \dot{\mathbf{G}}\mathbf{X}(\mathbf{X}^\mathbf{T}\dot{\mathbf{G}}^\mathbf{2}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\dot{\mathbf{G}}$, e $\dot{\mathbf{G}} = \text{diag}\{g'(h_i)\}$, avaliada no vetor de estimativas de máxima verossimilhança $(\hat{m},\hat{h},\hat{k})^T$. Por meio de estudos de simulação, os autores verificaram que o resíduo d_i^* possui distribuição aproximadamente Normal padrão.

2.3.3 Resíduo r_i: aproximação via série de Taylor

A segunda extensão do resíduo deviance para modelo de regressão von Mises, desenvolvida por Souza e Paula (2002), fundamenta-se em transformar o componente do desvio, expressando-o como função de uma variável aleatória com distribuição conhecida, e realizar uma expansão em série de Taylor até termos de segunda ordem. Com isso, o resíduo é dado por

$$r_i = sgn(q_i - \hat{m}_i)(2/p)^{1/2} \frac{\text{sen}\left(\frac{q_i - \hat{m}_i}{2}\right)}{I_0(\hat{k})e^{-\hat{k}}}, \tag{2.40}$$

em que I_0 representa a função de Bessel modificada de 1º tipo, definida na equação (2.15). Como verificado pelos autores, r_i também possui distribuição aproximadamente Normal padrão.

2.3.4 Resíduo quantílico circular

Seja $F(y; m, f)$ a função distribuição acumulada de uma variável aleatória linear Y, contínua, com média m e parâmetro de precisão ou dispersão f. Em um modelo de regressão assumindo que m varia em função de variáveis preditoras e f é constante, o resíduo quantílico (DUNN; SMYTH, 1996) é definido como

$$r_{q_i} = \text{F}^{-1}\{F(y_i; \hat{m}_i, \hat{f})\}, \tag{2.41}$$

em que $\text{F}(.)$ denota a função de distribuição acumulada da Normal padrão. Dunn e Smyth (1996), ainda, propuseram uma extensão ao resíduo quantílico, denominada resíduo quantílico aleatorizado, adequada aos casos em que Y possui distribuição discreta. Além disso, se a distribuição de Y contém mais de dois parâmetros, o resíduo quantílico pode ser definido de maneira similar.

Pode-se demonstrar que tanto o resíduo quantílico, quanto o resíduo quantílico aleatorizado, possuem distribuição assintótica Normal padrão, quando os parâmetros do modelo são consistentemente estimados. Contudo, como já mencionado no Capítulo 1, Pereira (2019), Scudilio e Pereira (2020) e Lemonte e Moreno-Arenas (2019) mostraram que a distribuição do resíduo quantílico é aproximadamente Normal padrão, mesmo em pequenas amostras, considerando, respectivamente, os modelos de regressão beta, MLG e os modelos de regressão generalizados de Johnson S_B. Já para o resíduo quantílico aleatorizado, Feng, Li e Sadeghpour (2020) mostraram que sua distribuição também é bem aproximada pela Normal padrão em pequenas amostras, nos modelos de regressão Poisson e binomial negativo.

Em um primeiro momento, pode-se pensar em utilizar o resíduo quantílico no âmbito dos modelos de regressão circular, sem qualquer tipo de correção ou adaptação, isto é, por meio da definição (2.41). Contudo, é válido lembrar que a função distribuição acumulada de uma variável aleatória circular depende da posição de origem, como definido em (2.1). Para elucidar os problemas dessa abordagem, considere o seguinte exemplo: suponha que a posição de origem

seja 0, $q_i = 11p/12$, $\hat{m}_i = 0$ e $\hat{g} = e^{-1/2}$, sendo os dois últimos estimados com base no modelo de médias *wrapped* Cauchy, proposto na Seção 2.2.3. Por (2.41), obtém-se

$$\begin{aligned} r_{q_i} &= \mathrm{F}^{-1}\{F(q_i;\hat{m}_i,\hat{g})\} = \mathrm{F}^{-1}\left\{F\left(\frac{11p}{12};\hat{m}_i = 0,\hat{g} = e^{-1/2}\right)\right\} \\ &= \mathrm{F}^{-1}\left\{\mathrm{P}\left(0 \leq q \leq \frac{11p}{12}\,\middle|\, q \sim WC(0,e^{-1/2})\right)\right\} = \mathrm{F}^{-1}(0,4877) \\ &= -0,0309, \end{aligned}$$

ou seja, apesar da observação q_i estar praticamente oposta à sua respectiva estimativa $\hat{m}_i$, o resíduo quantílico apresenta valor próximo de zero. A Figura 2.9(a) ilustra essa situação, destacando o intervalo de integração $[0, q_i]$.

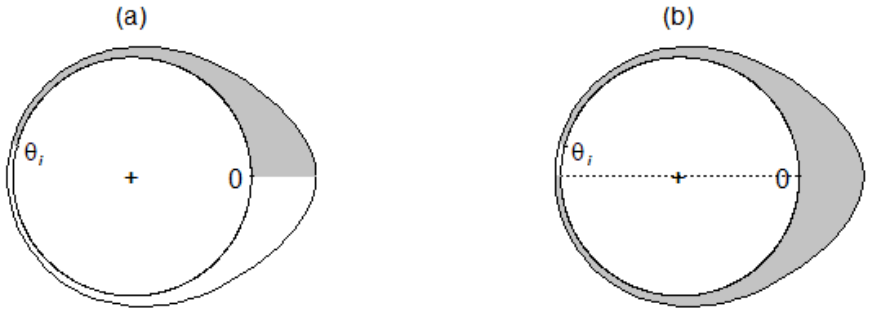

Figura 2.9 – Regiões de integração para os exemplos em que $q_i = 11p/12$, $\hat{m}_i = 0$, $\hat{g} = e^{-1/2}$, com origem no 0, considerando r_{q_i} (a) e $r^*_{q_i}$ (b).

Entretanto, por (2.4), pode-se observar que mudar a posição de origem implica, simplesmente, em adicionar uma constante a F (MARDIA; JUPP, 2009). Dito isto, a função distribuição acumulada, no contexto de dados circulares, pode ser definida sob qualquer intervalo de comprimento $2p$. No pacote *circular*, implementado no *R* por Lund, Agostinelli e Agostinelli (2017), por exemplo, os autores expressam-na a partir de $m - p$, por padrão. Essa escolha é conveniente para distribuições simétricas, uma vez que possibilita representar a função distribuição acumulada na reta real, ao "desembrulhá-la" no intervalo $[m-p, m+p]$.

Nesse sentido, para o emprego do resíduo quantílico no âmbito da regressão circular, propõe-se utilizar a função distribuição acumulada definida da seguinte forma

$$F^*(q_0) = \mathrm{P}(\mathrm{Md}_i - p \leq q_i \leq q_0), \tag{2.42}$$

em que Md_i representa a mediana circular da variável aleatória $q_i \sim F(\underset{\sim}{t}_i)$, sendo $\underset{\sim}{t}_i$ o vetor de parâmetros que define a distribuição F. Portanto, a extensão do resíduo quantílico, apresentada neste trabalho, é dada por

$$r^*_{q_i} = \mathrm{F}^{-1}\{\hat{F}^*(q_i;\hat{\underset{\sim}{t}}_i)\}, \tag{2.43}$$

sendo $\hat{F}^*$ o estimador de F^*, com base em $\widehat{\mathrm{Md}_i}$.

Voltando ao exemplo anterior, tem-se $\widehat{\text{Md}}_i = 0$, uma vez que a distribuição *wrapped* Cauchy é simétrica com $\hat{m}_i = 0$. Utilizando a definição (2.43),

$$\begin{aligned} r^*_{q_i} &= \text{F}^{-1}\{F^*(q_i;\hat{m}_i,\hat{g})\} = \text{F}^{-1}\left\{F\left(\frac{11p}{12};\hat{m}_i = 0, \hat{g} = e^{-1/2}\right)\right\} \\ &= \text{F}^{-1}\left\{\text{P}\left(0 - p \leq q \leq \frac{11p}{12} \middle| q \sim WC(0, e^{-1/2})\right)\right\} = \text{F}^{-1}(0,9877) \\ &= 2,2468, \end{aligned}$$

o que corrobora com o fato de $\hat{m}_i$ e q_i serem extremamente distantes no círculo. Assim como no cenário anterior, a Figura 2.9(b) evidencia a região de integração. A comparação entre os gráficos (a) e (b) da Figura 2.9 traz indícios, mesmo que intuitivos, sobre a plausibilidade de $r^*_{q_i}$, em confronto à r_{q_i}.

É razoável que se questione a escolha de $\text{Md}_i - p$ como limite inferior do intervalo, uma vez que, para algumas distribuições, coincide com o ponto $m_i - p$. De fato, para distribuições simétricas como a *wrapped* Cauchy, não há distinção entre as duas alternativas. Contudo, o mesmo não ocorre no contexto das regressões com resposta assimétrica. O uso da mediana circular fundamenta-se justamente na sua definição, isto é, no fato de garantir que $F^*(\text{Md}_i) = \frac{1}{2}$, o que implica, por exemplo, $r^*_{q_i} \approx 0$ para os casos em que $\widehat{\text{Md}}_i \approx q_i$. Analisando a Figura 2.10, observa-se que o resíduo $r^*_{q_i}$ assume valores negativos e decrescentes em módulo, quando $q_i \in [\widehat{\text{Md}}_i - p, \widehat{\text{Md}}_i]$, e valores positivos crescentes, se $q_i \in [\widehat{\text{Md}}_i, \widehat{\text{Md}}_i + p]$. Ou seja, $r^*_{q_i}$ se aproxima de zero, a medida que $\widehat{\text{Md}}_i \to q_i$, indo de encontro ao que se espera de um resíduo. Da mesma forma, $r^*_{q_i}$ se afasta de zero, a medida que a distância entre $\widehat{\text{Md}}_i$ e q_i cresce.

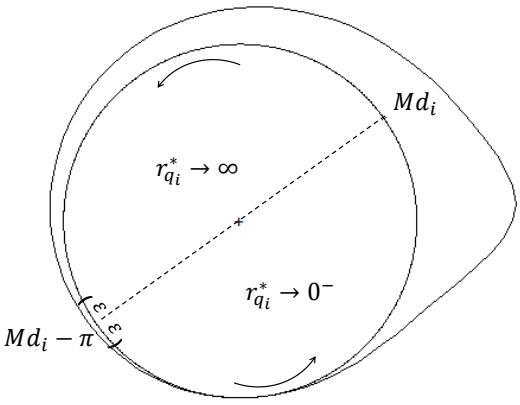

Figura 2.10 – Comportamento do resíduo $r^*_{q_i}$.

No âmbito da regressão circular, é essencial que os resíduos sejam invariantes rotacionais, ou seja, que não alterem seu valor se todas as observações sofrerem uma rotação de mesmo ângulo. Considere uma observação q_i e sua respectiva mediana estimada $\widehat{\text{Md}}_i$, o valor calculado de $r^*_{q_i}$ deverá ser igual ao resíduo da observação rotacionada $q_i + c$, com mediana ajustada $\widehat{\text{Md}}_i + c$. Assumindo que os parâmetros da variável resposta não relacionados à direção mediana são mantidos constantes, pode-se observar na Figura (2.11) que o resíduo quantílico circular

satisfaz essa propriedade. Observe que a curva de distribuição de probabilidade estimada da variável resposta gira de acordo com a observação, pois a mediana estimada foi aumentada em c. Desta forma, as regiões de integração correspondentes ao cálculo do resíduo quantílico são equivalentes, como pode ser verificado pela congruência dos ângulos.

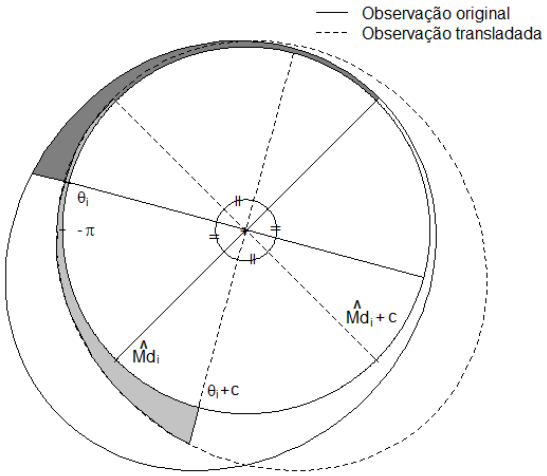

Figura 2.11 – Comportamento do resíduo $r^*_{q_i}$ ao girar uma observação q_i.

Para verificar a adequação do modelo, é interessante utilizar resíduos cuja distribuição seja bem aproximada pela distribuição Normal padrão. Quando os resíduos não possuem essa propriedade, às vezes eles também não são identicamente disribuídos (ver, por exemplo, Tabela 1 de Anholeto, Sandoval e Botter (2014)). Quando isso ocorre, modelos bem ajustados podem ser descartados (ver aplicação 3 de Pereira (2019)), mesmo quando utilizados gráficos de probabilidade Normal com envelope simulado (ATKINSON, 1981) na análise diagnóstica. O teorema a seguir afirma que o resíduo quantílico circular é assintoticamente Normal padrão distribuído.

Teorema 2.3.1. Seja $q_1, ..., q_n$ variáveis aleatórias circulares independentes, em que cada q_i possui um vetor de parâmetros $\underset{\sim}{t}_i$. Se $\underset{\sim}{t}_i$ e Md_i são consistentemente estimados, sob o modelo verdadeiro,

$$r^*_{q_i} \xrightarrow{D} N(0,1),$$

em que $\xrightarrow{D}$ denota convergência em distribuição.

O teorema é provado no Apêndice A.5. Isso sugere que a distribuição do resíduo quantílico circular é bem aproximada pela distribuição Normal padrão quando o tamanho da amostra é grande. Para avaliar o comportamento da distribuição de probabilidade do resíduo quantílico

circular em pequenas amostras, principalmente em comparação com os demais resíduos descritos, foram realizados estudos de simulação apresentados na seção seguinte.

É válido ressaltar que no ponto $\widehat{\text{Md}}_i - p = q_i$, $r^*_{q_i} = -¥$, uma vez que $F^*(\widehat{\text{Md}}_i - p; \underset{\sim}{\hat{t}}_i) = 0$. Similarmente, para o caso em que $(\widehat{\text{Md}}_i - p) + 2p = q_i$, $r^*_{q_i} = ¥$, uma vez que $F^*((\widehat{\text{Md}}_i - p) + 2p; \underset{\sim}{\hat{t}}_i) = 1$. Contudo, considerando que a variável dependente possui natureza contínua,

$$\text{P}(\widehat{\text{Md}}_i - p = q_i) = \text{P}((\widehat{\text{Md}}_i - p) + 2p = q_i) = 0. \tag{2.44}$$

Entretanto, em virtude das limitações de precisão e exatidão das máquinas e, consequentemente, dos métodos computacionais, poderá haver casos em que $r^*_{q_i}$ diverge. Como solução, sugere-se a escolha de um $e > 0$, tão pequeno quanto se queira, onde

$$r^*_{q_i} = \begin{cases} \text{F}^{-1}\{F^*(\widehat{\text{Md}}_i - p - e; \underset{\sim}{\hat{t}}_i)\}, & \text{se } q_i - (\widehat{\text{Md}}_i - p) \in [-e, 0] \\ \text{F}^{-1}\{F^*(\widehat{\text{Md}}_i - p + e; \underset{\sim}{\hat{t}}_i)\}, & \text{se } q_i - (\widehat{\text{Md}}_i - p) \in (0, e]. \end{cases} \tag{2.45}$$

Para avaliar o comportamento da distribuição de probabilidade do resíduo proposto em pequenas amostras, sobretudo em comparação aos demais resíduos descritos, realizaram-se estudos de simulação apresentados na seção seguinte. Considerou-se, neste caso, $e = 10^{-10}$. Para as distribuições simétricas, a direção mediana coincide com a média circular. Portanto, nos modelos de regressão com resposta von Mises ou *wrapped* Cauchy, utilizou-se o estimador de máxima verossimilhança de m_i como estimador de Md_i, que corresponde a direção mediana para observação i. Para o modelo de regressão com reposta *sine-skewed* von Mises, Md_i não tem uma forma fechada. Neste caso, a mediana foi estimada numericamente pelo método do trapézio (PURVES, 1992).

Mediante as estimativas de máxima verossimilhança de $b_0, b_1, ..., b_p$, foi obtida a estimativa de máxima verossimilhança de m_i para cada observação. Com base em em $\hat{m}_i$ e $\hat{b}_0, \hat{b}_1, ..., \hat{b}_p$, e nas estimativas de máxima verossimilhança dos demais parâmetros da distribuição resposta, procurou-se numericamente um ponto no intervalo $[-p, p)$ que satisfaz a Equação (2.10), considerando um erro de três casas decimais. Nota-se que o estimador da mediana obtido dessa forma é um estimador consistente de Md_i, uma vez que corresponde a uma função dos estimadores de máxima verossimilhança do modelo.

2.4 Estudos de simulação

Nessa seção, os estudos de simulação foram divididos de acordo com a distribuição da variável resposta, considerando, portanto, os três modelos de regressão circular descritos na Seção 2.2 e duas covariáveis, nas quais

$$m_i = g(b_0 + b_1 x_{i1} + b_2 x_{i2}). \tag{2.46}$$

Paralelamente, foram realizados estudos de simulação para avaliar a convergência dos estimadores de máxima verossimilhança. Os resultados destes estudos foram omitidos por questões

de brevidade. Os modelos considerados nos estudos de simulação das Seções 2.4.1 a 2.4.3 não apresentam problemas de convergência dos estimadores.

Com o intuito de estudar as características da distribuição do resíduo quantílico circular, foram realizadas 5000 réplicas de Monte Carlo. Para cada um dos modelos, foram avaliadas diferentes condições derivadas de um cenário denominado como base, por meio de variações dos valores pré definidos para os parâmetros, da função ligação e/ou da distribuição atribuída às covariáveis. Para todos os casos, considerou-se os tamanhos amostrais $n = 20$ e 50. Em cada um dos cenários considerados, as variáveis preditoras foram mantidas fixas em todas as réplicas de Monte Carlo.

Tendo em vista o interesse em resíduos cuja distribuição se aproxima de uma Normal padrão, foram calculadas algumas medidas descritivas para cada um dos resíduos explicitados, como média, variância, assimetria e curtose. Bem como, foram calculados os valores para a estatística de Anderson-Darling (ANDERSON; DARLING, 1954), utilizada para testar se cada resíduo possui distribuição Normal padrão. O valor da estatística de Anderson-Darling é usada como uma medida de proximidade entre cada distribuição residual e a distribuição Normal padrão.

2.4.1 Estudos de simulação para o modelo com resposta von Mises

Sabe-se que, para os modelos de regressão com resposta von Mises, em particular, as extensões propostas por Souza e Paula (2002) - isto é, os resíduos d_i^* e r_i - possuem distribuição aproximadamente Normal padrão. À vista disto, para esta classe de modelos, avaliou-se o comportamento do resíduo proposto, comparando-o com o resíduo *deviance* (d_i) (Equação 2.35) e os dois mencionados anteriormente. Para tal, foram considerados cinco cenários, conforme mostrado na Tabela 2.4. O cenário I é definido como cenário base. Para o cenário II, a função de ligação foi alterada. No cenário III, há uma diminuição no valor de do parâmetro de precisão dos dados. O cenário IV tem covariáveis geradas a partir das distribuições Normal e Gama e não da distribuição uniforme que foi usada nos demais cenários. Por fim, no cenário V, alterou-se os valores dos parâmetros do modelo para aumentar a variância circular de m_i ao longo das n observações.

Tabela 2.4 – Descrição dos cenários para o modelo de regressão von Mises.

Cenário	Função de Ligação	x_{i1}	x_{i2}	b_0	b_1	b_2	k
I-base	$g(t) = 2\arctan(t)$	$U(0,1)$	$U(0,1)$	1,75	-2,2	-1,2	4
II	$g(t) = 2p\frac{e^t}{1+e^t} - p$	$U(0,1)$	$U(0,1)$	1,75	-2,2	-1,2	4
III	$g(t) = 2\arctan(t)$	$U(0,1)$	$U(0,1)$	1,75	-2,2	-1,2	2
IV	$g(t) = 2\arctan(t)$	$N\left(\frac{1}{2}, \frac{1}{12}\right)$	$Gama(3,6)$	1,75	-2,2	-1,2	4
V	$g(t) = 2\arctan(t)$	$U(0,1)$	$U(0,1)$	2,50	-3,0	-2,0	4

A Tabela 2.5 apresenta a média e variância amostrais, os coeficientes de assimetria e curtose, e a estatística do teste de Anderson-Darling referente aos resíduos $r_{q_i}^*$, r_i e d_i, considerando

as 5000 réplicas via Monte Carlo para o cenário I e $n = 20$. Os resultados para o resíduo d_i^* foram omitidos, uma vez que sua distribuição é pior aproximada pela Normal padrão, quando comparada à distribuição de d_i, em todos os cenários. O resíduo d_i^* possui média, assimetria e curtose semelhantes ao resíduo *deviance*, contudo, sua variância se afasta de 1. O mesmo ocorre com o resíduo ponderado padronizado 2 na regressão beta, cuja padronização também consiste em uma divisão por $(1-h_{ii})^{1/2}$ (ANHOLETO; SANDOVAL; BOTTER, 2014; PEREIRA, 2019).

Para a maioria das observações, os resíduos r_i e d_i apresentam média e coeficiente de assimetria próximos de zero e coeficiente de curtose ligeiramente inferior a 3. O mesmo ocorre para o resíduo d_i^* e, portanto, os resultados obtidos para os resíduos r_i e d_i^* corroboram com àqueles encontrados por Souza (1999) e Souza e Paula (2002). Como também observado por esses autores, tem-se que o resíduo d_i^* possui variância maior que r_i, apresentando valores que se afastam de 1. É justamente neste ponto onde a distribuição do resíduo *deviance* melhor se aproxima da Normal padrão, em comparação à padronização via $(1-h_{ii})^{1/2}$. Em geral, a variância também é próxima de um para r_i, mas é superior a 1,1 para d_i em metade das 20 observações. No mesmo cenário, $r_{q_i}^*$ tem média, variância e coeficiente de curtose semelhantes a r_i. No entanto, $r_{q_i}^*$ tem coeficiente de assimetria mais próximo de zero do que r_i quando o valor de m_i está longe do valor de m_i da maioria das outras observações. Como consequência, em média, o resíduo $r_{q_i}^*$ apresenta valores inferiores para a estatística de Anderson-Darling, quando comparado aos demais resíduos.

Analogamente, as Tabelas A.1-A.4, encontradas no Apêndice A.6, exibem os resultados

Tabela 2.5 – Resultados da simulação para $r_{q_i}^*$, r_i e d_i, considerando o cenário I com $n = 20$ - modelo de regressão von Mises.

i	m_i	Média			Variância			Assimetria			Curtose			Estatística A-D		
		$r_{q_i}^*$	r_i	d_i	$r_{q_i}^*$	r_i	d_i	$r_{q_i}^*$	r_i	d_i	$r_{q_i}^*$	r_i	d_i	$r_{q_i}^*$	r_i	d_i
1	1,88	0,02	−0,06	−0,06	1,08	1,09	1,16	0,05	−0,17	−0,18	2,78	2,76	2,77	5,08	11,19	18,34
2	−0,83	−0,01	−0,01	−0,01	0,86	0,87	0,93	0,03	0,03	0,03	2,89	2,93	2,94	6,30	5,89	1,83
3	−1,07	0,04	0,04	0,04	1,00	1,01	1,08	0,03	0,03	0,04	2,82	2,85	2,86	3,17	3,24	6,05
4	1,00	0,00	0,00	0,00	1,00	1,01	1,08	−0,04	−0,04	−0,05	2,78	2,82	2,82	0,74	0,79	3,81
5	−1,38	−0,01	0,00	0,00	1,05	1,06	1,13	0,00	0,03	0,03	2,81	2,81	2,81	2,47	2,64	7,86
6	−0,75	0,00	0,00	0,00	0,99	1,00	1,06	0,04	0,04	0,04	2,85	2,89	2,90	0,47	0,45	2,24
7	−0,02	0,00	0,00	0,00	0,89	0,90	0,96	−0,02	−0,02	−0,02	2,87	2,92	2,93	3,40	3,08	0,58
8	1,00	−0,03	−0,03	−0,03	1,00	1,01	1,08	−0,05	−0,05	−0,05	2,82	2,85	2,87	2,36	2,40	5,41
9	1,30	0,02	0,01	0,01	1,05	1,06	1,13	−0,05	−0,10	−0,11	2,87	2,85	2,87	3,95	3,84	9,04
10	1,41	−0,01	−0,02	−0,02	1,03	1,04	1,11	−0,03	−0,09	−0,09	2,71	2,72	2,73	3,24	3,97	9,23
11	0,33	0,00	0,00	0,00	1,05	1,06	1,13	0,04	0,04	0,04	2,78	2,82	2,82	2,44	2,65	7,85
12	−1,79	−0,01	0,04	0,05	1,05	1,06	1,13	−0,06	0,17	0,18	2,90	2,90	2,91	2,23	5,60	10,63
13	−0,78	0,01	0,01	0,01	1,02	1,03	1,10	0,00	0,01	0,01	2,74	2,78	2,78	2,05	2,14	6,34
14	0,01	0,01	0,01	0,01	0,80	0,81	0,86	−0,02	−0,02	−0,02	2,87	2,91	2,92	13,89	13,31	6,66
15	−0,24	−0,02	−0,02	−0,02	0,90	0,91	0,97	0,03	0,02	0,02	2,88	2,88	2,89	4,20	3,92	2,04
16	−1,51	0,03	0,04	0,04	1,08	1,09	1,16	0,03	0,08	0,09	2,92	2,92	2,92	4,27	5,64	11,52
17	1,70	0,02	−0,02	−0,02	1,07	1,07	1,15	0,03	−0,13	−0,14	2,90	2,84	2,86	3,62	4,08	9,89
18	1,04	−0,01	−0,01	−0,02	0,99	1,00	1,07	−0,02	−0,02	−0,02	2,78	2,80	2,81	1,58	1,53	4,47
19	−1,23	0,00	0,00	0,00	1,03	1,04	1,11	0,09	0,11	0,11	2,83	2,85	2,86	1,91	2,10	6,39
20	1,08	0,01	0,01	0,01	1,05	1,06	1,13	−0,06	−0,06	−0,07	2,88	2,92	2,94	1,83	2,07	6,54
Média		0,00	0,00	0,00	1,00	1,01	1,08	0,00	−0,01	−0,01	2,83	2,85	2,86	3,46	4,03	6,84
Desv. Pad.		0,02	0,02	0,02	0,08	0,08	0,08	0,04	0,08	0,09	0,06	0,06	0,06	2,84	3,20	4,09

para os cenários II-V, respectivamente, considerando $n = 20$. No cenários I, II e IV, os três resíduos apresentam média próxima de 0 em todas as 20 observações da amostra. Já no cenários III e V, os resíduos r_i e d_i apresentam observações com valores substancialmente longe da referência, como por exemplo a primeira observação no cenário III, cuja média equivale a $-0,24$ e $-0,26$, respectivamente. A assimetria é próxima de 0 para todas as observações, em todos os cenários, apenas para o resíduo quantílico circular.

Tabela 2.6 – Média das medidas de distribuição dos resíduos $r^*_{q_i}$, r_i e d_i, considerando os cenários I-VI e os tamanhos amostrais $n = 20$ e $n = 50$ - modelo de regressão von Mises.

		$r^*_{q_i}$		r_i		d_i	
Cenários		$n = 20$	$n = 50$	$n = 20$	$n = 50$	$n = 20$	$n = 50$
	Média	0,00	0,00	0,00	0,00	0,00	0,00
	Variância	1,00	1,00	1,01	1,01	1,08	1,09
I	Assimetria	0,00	0,00	−0,01	0,01	−0,01	0,01
	Curtose	2,83	2,99	2,85	3,02	2,86	3,02
	Estatística A-D	3,46	1,28	4,03	2,16	6,84	4,47
	Média	0,00	0,00	0,00	0,00	0,00	0,00
	Variância	1,00	1,00	1,01	1,01	1,08	1,09
II	Assimetria	0,00	0,01	−0,01	0,00	−0,01	0,00
	Curtose	2,83	2,98	2,85	3,01	2,86	3,02
	Estatística A-D	2,37	1,24	2,73	2,08	5,52	4,40
	Média	0,00	0,00	−0,01	0,00	−0,01	0,00
	Variância	1,00	1,00	0,99	1,00	1,16	1,18
III	Assimetria	−0,01	0,01	−0,02	0,01	−0,02	0,01
	Curtose	2,98	3,02	2,78	2,76	2,77	2,75
	Estatística A-D	3,31	1,20	24,56	27,89	37,61	43,77
	Média	0,00	0,00	0,00	0,00	0,00	0,00
	Variância	1,00	1,00	1,01	1,01	1,08	1,09
IV	Assimetria	−0,01	−0,01	0,00	−0,01	0,00	−0,01
	Curtose	2,79	2,94	2,82	3,00	2,83	3,01
	Estatística A-D	3,45	1,25	3,40	1,28	6,35	3,61
	Média	0,00	0,00	0,01	0,01	0,01	0,01
	Variância	1,00	1,00	0,98	1,01	1,05	1,08
V	Assimetria	−0,01	0,00	−0,02	0,02	−0,02	0,02
	Curtose	2,84	2,99	2,83	2,99	2,82	2,99
	Estatística A-D	8,52	1,38	9,96	10,62	12,32	13,33

A Tabela 2.6 apresenta a média (entre as n observações) das medidas de distribuição, obtidas por cada um dos três resíduos, em todos os cenários e ambos tamanhos amostrais.Para $n = 20$ e para os três resíduos, os valores médios da média e assimetria são próximos de zero em todos os cenários, enquanto que os valores médios da curtose são ligeiramente inferiores a três para todos os resíduos. Para o mesmo tamanho de amostra, os valores médios da variância para $r^*_{q_i}$ e r_i são próximos de um, enquanto que a variância média de d_i é ligeiramente maior que um. A distribuição de $r^*_{q_i}$ é bem aproximada pela distribuição Normal padrão em todos os

cenários para $n = 20$, mas a aproximação é um pouco pior quando a variância de m_i ao longo das n observações é maior (cenário V). Por outro lado, a aproximação da distribuição dos demais resíduos à distribuição Normal padrão não é tão boa quando a variância da variável resposta é maior (cenário III). A estatística de Anderson-Darling revela que a distribuição de $r_{q_i}^*$ está mais próxima da distribuição Normal padrão, quando n aumenta de 20 para 50. Note que o mesmo não ocorre com os demais resíduos nos cenários III e V.

Tabela 2.7 – Comparação da estatística de Anderson-Darling, para $n = 20$ e $n = 50$, considerando os cenários I-VI - modelo de regressão von Mises.

Cenário	n	Resíduos	Média	DP	Mínimo	Q_1	Q_2	Q_3	Máximo
		$r_{q_i}^*$	3,46	2,84	0,47	2,01	2,82	4,01	13,89
	20	r_i	4,03	3,20	0,45	2,13	3,16	4,46	13,31
		d_i	6,84	4,09	0,58	4,31	6,46	9,09	18,34
I									
		$r_{q_i}^*$	1,28	1,02	0,23	0,52	0,99	1,66	4,57
	50	r_i	2,16	2,64	0,16	0,71	1,31	2,23	14,63
		d_i	4,47	3,31	0,28	2,37	4,05	5,46	17,42
		$r_{q_i}^*$	2,37	1,58	0,44	1,29	2,08	2,86	6,00
	20	r_i	2,73	1,63	0,49	1,63	2,60	3,31	5,81
		d_i	5,52	2,85	0,97	3,58	5,44	7,60	12,02
II									
		$r_{q_i}^*$	1,24	0,90	0,24	0,55	0,97	1,59	4,26
	50	r_i	2,08	3,14	0,18	0,66	1,25	2,07	20,49
		d_i	4,40	3,38	0,15	2,54	3,70	5,97	21,40
		$r_{q_i}^*$	3,31	4,08	0,58	1,26	2,26	3,07	18,45
	20	r_i	24,56	31,45	1,86	6,16	11,90	23,33	122,58
		d_i	37,61	40,04	1,32	15,93	23,42	42,44	156,96
III									
		$r_{q_i}^*$	1,20	0,98	0,26	0,53	0,87	1,32	4,11
	50	r_i	27,89	44,83	0,74	1,96	5,75	20,14	198,20
		d_i	43,77	51,21	3,02	15,68	19,77	38,25	233,00
		$r_{q_i}^*$	3,45	2,94	0,49	1,39	2,49	4,57	12,85
	20	r_i	3,40	2,98	0,51	1,36	2,00	4,94	12,27
		d_i	6,35	4,75	0,84	2,72	4,98	10,85	15,85
IV									
		$r_{q_i}^*$	1,25	0,87	0,15	0,56	1,00	1,62	4,18
	50	r_i	1,28	0,87	0,14	0,59	1,16	1,67	4,33
		d_i	3,61	2,11	0,32	1,74	3,54	5,38	8,45
		$r_{q_i}^*$	8,52	5,79	0,90	4,21	6,95	13,01	21,34
	20	r_i	9,96	11,14	1,67	3,28	5,05	8,72	37,78
		d_i	12,32	11,15	1,77	6,68	9,15	12,58	44,84
V									
		$r_{q_i}^*$	1,38	1,22	0,18	0,58	0,85	1,65	4,83
	50	r_i	10,62	23,29	0,17	0,62	1,50	4,46	108,79
		d_i	13,33	25,03	0,66	1,94	4,21	7,16	116,78

A Tabela 2.7 exibe, de maneira mais detalhada, os resultados para a estatística de Anderson-Darling, em todos os cenários considerados neste estudo de simulação. Com exceção do cenário IV, em que $r_{q_i}^*$ e r_i têm comportamento semelhante, a distribuição do resíduo proposto

revela-se mais próxima da distribuição Normal padrão, quando comparada a distribuição dos demais resíduos. Para ambos os tamanhos amostrais no cenário III, em que o parâmetro de concentração da variável resposta é menor, o valor máximo da estatística de Anderson-Darling referente ao resíduo quantílico circular é consideravelmente menor que a média para os demais resíduos. O mesmo ocorre no cenário V com $n = 50$, onde a variância de m_i também é aumentada. O resultado do cenário III é coerente com a conclusão de Souza e Paula (2002), que observaram que a distribuição dos resíduos r_i e d_i^* melhor se aproxima da distribuição Normal padrão para valores moderados ou grandes de k. Em alguns cenários, para $n = 20$, algumas observações apresentam valores mais elevados para estatística de Anderson-Darling, para todos os resíduos. Contudo, para o resíduo quantílico circular, esta situação minimiza-se substancialmente quando $n = 50$.

De uma forma geral, tem-se uma boa concordância entre a distribuição do resíduo proposto e a Normal padrão, mesmo em pequenas amostras, como ocorre nos modelos de regressão beta, MLG e nos modelos de regressão generalizados de Johnson S_B. Assim como observado no caso linear, e conforme o Teorema 2.3.1, a aproximação melhora quando n aumenta de 20 para 50.

2.4.2 Estudos de simulação para o modelo com resposta sine-skewed von Mises

Com o intuito de avaliar o comportamento do resíduo proposto em modelos de regressão cuja variável resposta possui distribuição assimétrica, considerou-se a classe de modelos com resposta SSvM. Para este caso, não foi encontrado nenhum outro resíduo com potencial distribuição Normal padrão, aproximadamente. É válido lembrar que, como mostrado na Seção 2.3.1, nessa classe de modelos o resíduo *deviance* (Equação 2.37) não pode ser calculado. Os cenários foram construídos de forma análoga ao caso von Mises, como pode ser observado na Tabela 2.8. Para esse modelo, tem-se um número maior de cenários, pois também foi considerado diferentes valores para o parâmetro de assimetria da distribuição da variável resposta.

Tabela 2.8 – Descrição dos cenários para o modelo de regressão *sine-skewed* von Mises.

Cenário	Função de ligação	x_{i1}	x_{i2}	b_0	b_1	b_2	k	l
I-base	$g(t) = 2\arctan(t)$	$U(0,1)$	$U(0,1)$	1,75	-2,2	-1,2	4	0,5
II	$g(t) = 2p\frac{e^t}{1+e^t} - p$	$U(0,1)$	$U(0,1)$	1,75	-2,2	-1,2	4	0,5
III	$g(t) = 2\arctan(t)$	$U(0,1)$	$U(0,1)$	1,75	-2,2	-1,2	2	0,5
IV	$g(t) = 2\arctan(t)$	$N(\frac{1}{2}, \frac{1}{12})$	$Gama(3,6)$	1,75	-2,2	-1,2	4	0,5
V	$g(t) = 2\arctan(t)$	$U(0,1)$	$U(0,1)$	1,75	-2,2	-1,2	4	0,9
VI	$g(t) = 2\arctan(t)$	$U(0,1)$	$U(0,1)$	1,75	-2,2	-1,2	4	-0,2
VII	$g(t) = 2\arctan(t)$	$U(0,1)$	$U(0,1)$	2,50	-3,0	-2,0	4	0,5

Para $n = 20$, no cenário base, semelhante à regressão von Mises, a média e o coeficiente de assimetria do resíduo quantílico circular não diferem substancialmente dos correspondentes

valores na distribuição Normal padrão para todas as observações (Tabela 2.9). Por outro lado, o coeficiente de curtose é ligeiramente inferior a 3 para todas as observações, e a variância não é tão próxima de um para algumas observações. A média e o desvio padrão da estatística de Anderson-Darling são maiores que os observados no cenário base do modelo de regressão von Mises. No entanto, o resíduo quantílico circular se aproxima satisfatoriamente da distribuição Normal padrão, uma vez que a estatística do teste não é tão alta para nenhuma das observações.

Tabela 2.9 – Resultados da simulação para $r_{q_i}^*$, considerando o cenário I com $n = 20$ - modelo de regressão *sine-skewed* von Mises.

i	m_i	Média	Variância	Assimetria	Curtose	Estatística A-D
1	1,88	0,04	1,00	−0,04	2,68	6,29
2	−0,83	−0,02	0,85	0,01	2,83	7,18
3	−1,07	0,04	1,00	0,00	2,77	3,86
4	1,00	−0,01	1,01	−0,04	2,76	0,83
5	−1,38	0,02	1,03	−0,05	2,76	3,85
6	−0,75	−0,02	0,98	0,01	2,80	1,40
7	−0,02	−0,04	0,88	−0,02	2,89	7,74
8	1,00	−0,03	1,01	−0,04	2,78	2,81
9	1,30	0,03	1,05	−0,06	2,85	6,48
10	1,41	0,00	1,03	−0,06	2,71	3,23
11	0,33	−0,03	1,04	0,02	2,79	3,82
12	−1,79	0,06	0,96	−0,04	2,90	11,72
13	−0,78	0,00	1,02	−0,01	2,73	1,77
14	0,01	−0,02	0,79	0,00	2,85	15,28
15	−0,24	−0,05	0,89	0,02	2,88	10,70
16	−1,51	0,06	1,04	−0,03	2,81	12,70
17	1,70	0,04	1,01	−0,03	2,76	6,25
18	1,04	−0,02	1,00	−0,03	2,76	1,96
19	−1,23	0,02	1,02	0,07	2,72	2,17
20	1,08	0,01	1,05	−0,09	2,86	2,41
Média		0,01	0,98	−0,02	2,79	5,62
Desv. Pad.		0,03	0,07	0,04	0,06	4,16

De maneira análoga, as Tabelas A.5-A.10 (Apêndice A.7) exibem os resultados para os cenários II-VII, respectivamente, com $n = 20$. Com exceção do cenário V, em que o coeficiente de assimetria é maior, todas as observações têm média próximo de zero. Nos cenários III, V e VII, algumas observações apresentam assimetria maior que 0,11. Entretanto, no cenário III o coeficiente de curtose se aproxima de 3. O aumento do tamanho amostral, ocasiona melhoras significativas no coeficiente de curtose (Tabela 2.10) para os demais cenários e aproxima a variância a 1. Como consequência, a estatística de Anderson-Darling diminui e, portanto, a distribuição do resíduo proposto se aproxima de uma Normal padrão quando n aumenta de 20 para 50.

Novamente, tem-se uma boa concordância entre a distribuição do resíduo proposto e a

Tabela 2.10 – Média das medidas de distribuição do resíduo $r_{q_i}^*$, considerando os cenários I-VII e os tamanhos amostrais $n = 20$ e $n = 50$ - modelo de regressão *sine-skewed* von Mises.

Cenários	Tamanho amostral	Média	Variância	Assimetria	Curtose	Estatística A-D
I	$n = 20$	0,01	0,98	−0,02	2,79	5,62
	$n = 50$	0,00	1,00	−0,04	2,95	1,61
II	$n = 20$	0,00	0,99	−0,01	2,77	3,45
	$n = 50$	0,00	1,00	−0,03	2,93	1,66
III	$n = 20$	0,00	0,98	−0,04	3,04	4,82
	$n = 50$	0,00	1,00	−0,04	3,04	1,74
IV	$n = 20$	0,00	0,98	−0,01	2,75	5,26
	$n = 50$	0,00	0,99	−0,02	2,91	1,65
V	$n = 20$	0,02	0,97	0,03	2,78	16,48
	$n = 50$	0,01	0,99	0,02	2,96	6,15
VI	$n = 20$	−0,01	0,99	−0,01	2,79	6,29
	$n = 50$	−0,01	1,00	0,00	2,94	3,22
VII	$n = 20$	0,01	0,98	−0,03	2,79	6,83
	$n = 50$	0,00	1,01	−0,05	2,97	1,56

Normal padrão, embora a aproximação não seja tão boa quanto na regressão von Mises. Isso pode ser notado, por exemplo, comparando-se os valores das estatísticas de Anderson-Darling dos dois modelos (Tabelas 2.7 e 2.11). A aproximação é pior especialmente quando o parâmetro de assimetria da variável resposta é alto (cenário V). Como esperado, a aproximação melhora com o aumento do tamanho amostral.

2.4.3 Estudos de simulação para o modelo com resposta wrapped Cauchy

Nesta última classe de modelos considerada, avaliou-se o comportamento do resíduo proposto, comparando-o com o resíduo *deviance* (d_i) (Equação 2.38). De maneira análoga aos casos anteriores, foram analisados diferentes cenários, como mostra a Tabela 2.12.

Para $n = 20$, em todos os cenários, a média e a assimetria do resíduo quantílico circular não difere substancialmente de zero para todas as observações (Tabelas 2.13 e A.11-A.15, no Apêndice A.8). A variância é ligeiramente superior a um para a maioria das observações, em todos os cenários. Para todos os cenários, a curtose, em média, é próxima de 3, embora se afaste para algumas observações. Como consequência, a média e o desvio padrão da estatística de Anderson-Darling são maiores do que os observados nos cenários do modelo de regressão von Mises, o que

Tabela 2.11 – Comparação da estatística de Anderson-Darling, para $n = 20$ e $n = 50$, considerando os cenários I-VII - modelo de regressão *sine-skewed* von Mises.

Cenário	n	Média	DP	Mínimo	Q_1	Q_2	Q_3	Máximo
I	20	5,62	4,16	0,83	2,35	3,85	7,32	15,28
	50	1,61	1,22	0,24	0,74	1,30	2,15	6,34
II	20	3,45	1,75	1,10	2,07	3,03	5,10	6,47
	50	1,66	1,62	0,23	0,81	1,08	1,75	9,56
III	20	4,82	4,61	0,34	1,92	2,54	8,83	16,38
	50	1,74	1,17	0,30	0,86	1,21	2,23	4,92
IV	20	5,26	6,14	0,49	1,72	3,77	5,02	26,37
	50	1,65	1,58	0,13	0,68	1,27	1,90	8,47
V	20	16,48	16,01	1,26	4,11	10,65	27,87	57,24
	50	6,15	7,73	0,38	1,41	3,11	6,06	38,31
VI	20	6,29	6,14	0,48	2,36	4,09	7,50	22,61
	50	3,22	3,03	0,20	1,07	2,12	4,10	14,60
VII	20	6,83	4,43	0,34	3,74	5,13	9,68	14,75
	50	1,56	1,26	0,33	0,71	1,00	2,10	6,24

Tabela 2.12 – Descrição dos cenários para o modelo de regressão *wrapped* Cauchy.

Cenário	Função de ligação	x_{i1}	x_{i2}	b_0	b_1	b_2	g
I-base	$g(t) = 2\arctan(t)$	$U(0,1)$	$U(0,1)$	1,75	-2,2	-1,2	0,15
II	$g(t) = 2p\frac{e^t}{1+e^t} - p$	$U(0,1)$	$U(0,1)$	1,75	-2,2	-1,2	0,15
III	$g(t) = 2\arctan(t)$	$U(0,1)$	$U(0,1)$	1,75	-2,2	-1,2	0,35
IV	$g(t) = 2\arctan(t)$	$U(0,1)$	$U(0,1)$	1,75	-2,2	-1,2	1,00
V	$g(t) = 2\arctan(t)$	$N\left(\frac{1}{2}, \frac{1}{12}\right)$	$Gama(3,6)$	1,75	-2,2	-1,2	0,15
VI	$g(t) = 2\arctan(t)$	$U(0,1)$	$U(0,1)$	2,50	-3,0	-2,0	0,15

indica que a distribuição do resíduo quantílico circular é mais próxima à distribuição Normal padrão no caso da regressão von Mises. No entanto, o resíduo quantílico circular aproxima-se satisfatoriamente da distribuição Normal padrão, principalmente em comparação com o resíduo *deviance*, cujos valores são significativamente maiores.

A Tabela 2.14 apresenta a média das medidas de distribuição, para os resíduos $r^*_{q_i}$ e d_i, em todos os cenários e em ambos os tamanhos de amostra. Para $n = 20$, a distribuição de $r^*_{q_i}$ é bem aproximada pela distribuição Normal padrão. No entanto, a aproximação é pior do que nos modelos anteriores, especialmente porque o resíduo quantílico circular tem variância ligeiramente superior a um na regressão *wrapped* Cauchy. Para $n = 50$, a aproximação da distribuição Normal padrão é muito boa e próximo ao observado para os demais modelos. Por outro lado, para ambos os tamanhos de amostra, a distribuição do resíduo deviance na regressão *wrapped* Cauchy não

Tabela 2.13 – Resultados da simulação para $r^*_{q_i}$ e d_i, considerando o cenário I com $n = 20$ - modelo de regressão *wrapped* Cauchy.

i	m_i	Média		Variância		Assimetria		Curtose		Estatística A-D	
		$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i
1	1,88	−0,02	−0,23	1,16	2,86	0,01	−0,04	2,71	2,09	14,91	865,54
2	−0,83	0,03	0,10	1,02	2,52	0,03	0,05	3,30	2,57	26,88	452,40
3	−1,07	−0,02	0,07	1,10	2,74	0,00	0,05	2,87	2,25	5,64	644,83
4	1,00	0,01	−0,08	1,05	2,60	0,04	−0,06	3,09	2,38	6,59	527,12
5	−1,38	−0,01	0,12	1,11	2,77	−0,02	0,04	2,84	2,21	6,87	691,22
6	−0,75	−0,01	0,06	1,07	2,64	−0,01	0,06	3,09	2,39	8,22	540,09
7	−0,02	0,01	0,03	1,03	2,53	−0,02	−0,01	3,27	2,51	14,84	453,92
8	1,00	0,00	−0,06	1,08	2,67	−0,11	−0,09	3,08	2,33	6,51	571,76
9	1,30	−0,01	−0,12	1,12	2,76	−0,03	−0,08	2,90	2,18	6,54	701,16
10	1,41	−0,02	−0,17	1,07	2,65	0,01	−0,07	2,97	2,24	4,12	639,56
11	0,33	0,03	0,01	1,12	2,79	−0,04	−0,04	2,84	2,21	8,66	685,52
12	−1,79	−0,02	0,16	1,08	2,68	0,00	0,09	2,90	2,19	4,20	673,22
13	−0,78	0,00	0,06	1,09	2,73	−0,01	0,03	2,81	2,23	4,53	645,97
14	0,01	−0,01	−0,03	1,01	2,44	0,04	0,01	3,58	2,78	45,87	368,33
15	−0,24	−0,03	−0,02	1,06	2,59	−0,01	0,03	3,22	2,50	16,63	480,89
16	−1,51	0,01	0,15	1,11	2,75	0,02	0,07	2,88	2,18	5,87	705,27
17	1,70	−0,01	−0,19	1,09	2,71	0,04	−0,06	2,85	2,17	5,51	716,56
18	1,04	0,01	−0,08	1,10	2,74	0,02	−0,03	2,86	2,23	5,16	655,78
19	−1,23	0,00	0,10	1,08	2,71	0,00	0,06	2,83	2,24	4,62	637,62
20	1,08	0,00	−0,11	1,10	2,70	0,07	−0,02	3,02	2,31	4,18	618,85
Média		0,00	−0,01	1,08	2,68	0,00	0,00	3,00	2,31	10,32	613,78
Desv. Pad.		0,01	0,12	0,04	0,10	0,04	0,06	0,21	0,17	10,14	115,59

pode ser aproximada pela distribuição Normal padrão, pois possui variância muito maior que um e coeficiente de curtose muito menor do que 3.

Embora o resíduo quantílico circular apresente algumas observações com valores elevados para estatística de Anderson-Darling, esta situação minimiza-se substancialmente em $n = 50$ (Tabela 2.15). É válido ressaltar que os valores máximos da estatística do teste, referente ao resíduo proposto, são consideravelmente menores que os mínimos observados no resíduo *deviance*. Portanto, novamente, tem-se uma boa concordância entre a distribuição do resíduo proposto e a Normal padrão, sobretudo em comparação ao resíduo d_i.

2.5 Aplicação

Na Seção 2.4, estudou-se a distribuição dos resíduos considerados neste trabalho, avaliando se são bem aproximados pela distribuição Normal padrão. No entanto, é de suma importância que os resíduos sejam capazes, também, de identificar erros de especificação do modelo. Para avaliar essa propriedade, foram consideradas duas aplicações, sendo a primeira via dados simulados e a segunda com um banco de dados reais.

Tabela 2.14 – Média das medidas de distribuição dos resíduos $r^*_{q_i}$ e d_i, considerando os cenários I-VI e os tamanhos amostrais $n = 20$ e $n = 50$ - modelo de regressão *wrapped* Cauchy.

		$r^*_{q_i}$		d_i	
Cenários		$n = 20$	$n = 50$	$n = 20$	$n = 50$
I	Média	0,00	0,00	−0,01	0,00
	Variância	1,08	1,02	2,68	2,53
	Assimetria	0,00	0,00	0,00	0,00
	Curtose	3,00	3,00	2,31	2,25
	Estatística A-D	10,32	1,74	613,78	556,76
II	Média	0,00	0,00	−0,01	0,00
	Variância	1,08	1,02	2,68	2,53
	Assimetria	0,00	0,00	0,00	0,00
	Curtose	2,99	3,00	2,31	2,25
	Estatística A-D	8,04	1,68	605,1	553,85
III	Média	0,00	0,00	−0,01	0,00
	Variância	1,09	1,02	2,32	2,16
	Assimetria	−0,01	0,00	0,00	0,00
	Curtose	2,98	2,99	2,13	2,02
	Estatística A-D	10,98	1,85	577,93	520,14
IV	Média	0,01	0,00	−0,06	0,00
	Variância	1,11	1,03	1,63	1,31
	Assimetria	−0,01	0,00	0,05	0,00
	Curtose	2,98	2,99	2,06	1,86
	Estatística A-D	15,32	2,49	330,85	350,13
V	Média	0,00	0,00	−0,01	−0,02
	Variância	1,08	1,02	2,68	2,54
	Assimetria	0,00	0,00	0,00	−0,02
	Curtose	2,99	3,01	2,32	2,27
	Estatística A-D	10,39	2,00	599,39	541,87
VI	Média	0,00	0,00	−0,01	0,01
	Variância	1,08	1,02	2,64	2,52
	Assimetria	0,00	0,00	0,01	0,00
	Curtose	3,00	3,00	2,32	2,24
	Estatística A-D	14,39	1,96	619,88	574,66

2.5.1 Aplicação com dados simulados

Para esta primeira aplicação, foram geradas 200 observações provenientes do modelo de regressão com resposta *sine-skewed* von Mises. As covariáveis foram geradas com base em uma $U(0,1)$, utilizando $b_0 = 3,5$, $b_1 = 1,8$, $b_2 = 1,2$, $k = 1$, e $l = -0,9$.

A Figura (2.12) (a) apresenta o gráfico de probabilidade normal com envelope simulado (ATKINSON, 1981) para o resíduo quantílico circular, quando o modelo com resposta *sine-skewed* von Mises é ajustado. Corretamente, o resíduo $r^*_{q_i}$ não indicou nenhuma falha de ajuste, uma vez que não há pontos fora dos limites do envelope simulado. Adiante, foram ajustados os modelos de regressão com resposta von Mises e com resposta *wrapped* Cauchy, em que claramente há um erro de especificação. Erroneamente, os resíduos r_i e d_i não sugerem falhas

Tabela 2.15 – Comparação da estatística de Anderson-Darling, para $n = 20$ e $n = 50$, considerando os cenários I-VI - modelo de regressão *wrapped* Cauchy.

Cenário	n	Resíduos	Média	DP	Mínimo	Q_1	Q_2	Q_3	Máximo
I	20	$r^*_{q_i}$	10,32	10,14	4,12	5,03	6,53	10,20	45,87
		d_i	613,78	115,59	368,33	536,85	642,19	686,95	865,54
	50	$r^*_{q_i}$	1,74	0,92	0,39	1,13	1,57	2,58	4,11
		d_i	556,76	55,46	443,97	520,14	556,20	592,55	724,17
II	20	$r^*_{q_i}$	8,04	6,16	3,93	5,01	5,85	7,34	28,32
		d_i	605,10	81,00	440,21	558,92	620,01	654,98	746,06
	50	$r^*_{q_i}$	1,68	0,86	0,51	0,97	1,55	2,13	3,96
		d_i	553,85	44,94	443,79	526,83	552,48	579,03	685,92
III	20	$r^*_{q_i}$	10,98	10,21	3,84	5,90	7,47	10,42	47,88
		d_i	577,93	158,42	286,30	480,55	595,58	653,46	973,55
	50	$r^*_{q_i}$	1,85	1,25	0,41	1,02	1,53	2,20	6,60
		d_i	520,14	109,24	375,45	443,52	503,96	549,17	903,76
IV	20	$r^*_{q_i}$	15,32	8,73	5,30	10,77	13,51	17,25	40,01
		d_i	330,85	137,81	96,62	237,49	341,77	409,22	647,40
	50	$r^*_{q_i}$	2,49	1,50	0,58	1,31	2,18	3,30	7,44
		d_i	350,13	220,24	122,16	190,16	246,70	441,28	937,00
V	20	$r^*_{q_i}$	10,39	9,85	4,44	5,79	7,67	9,96	49,66
		d_i	599,39	96,61	369,80	557,95	593,79	672,13	746,60
	50	$r^*_{q_i}$	2,00	1,57	0,30	0,94	1,52	2,56	8,37
		d_i	541,87	44,19	407,26	515,15	544,02	565,66	615,52
VI	20	$r^*_{q_i}$	14,39	16,77	4,23	6,13	8,34	13,49	77,72
		d_i	619,88	162,86	293,94	498,15	634,48	710,52	971,39
	50	$r^*_{q_i}$	1,96	1,32	0,32	1,13	1,51	2,41	6,62
		d_i	574,66	85,83	427,08	517,22	567,40	599,09	847,90

no ajuste dos modelos von Mises (Figura (2.12) (c)(d)). O mesmo acontece para o resíduo *deviance* nos modelos de regressão *wrapped* Cauchy (Figura (2.12) (f)). Por outro lado, o resíduo quantílico circular sugere corretamente problemas na especificação dos modelos, posto que é possível observar uma série de pontos consecutivos fora dos limites do envelope simulado (Figura (2.12) (b)(e))

2.5.2 Aplicação com dados reais

A segunda aplicação usa os dados de Scapini *et al.* (2002), correspondentes à orientação de dois crustáceos da espécie funil de areia *(sympatric sandhoppers)* de uma praia no Mediterrâneo, expostos a diferentes fatores ambientais. Os dados estão disponíveis no pacote

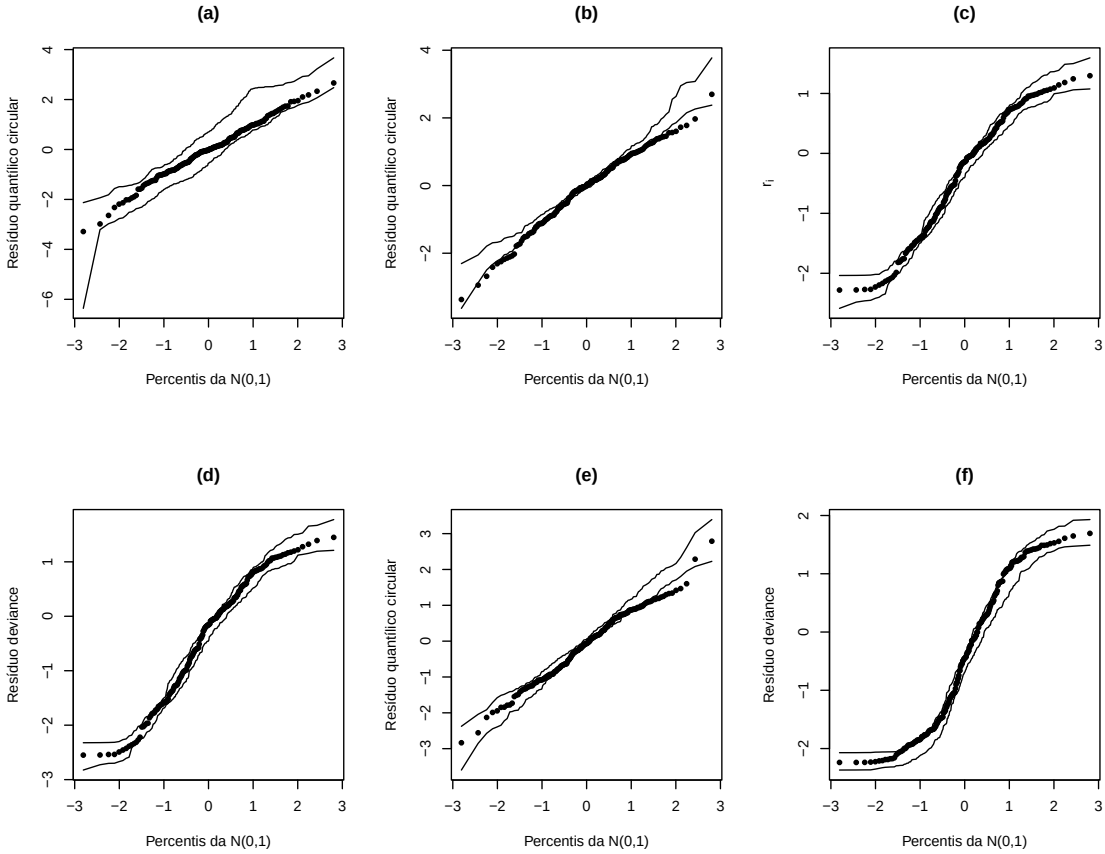

Figura 2.12 – Gráfico de probabilidade normal com envelope simulado considerando os dados simulados para o resíduo $r^*_{q_i}$ (a), quando ajustado corretamente o modelo *sine-skewed* von Mises; para $r^*_{q_i}$ (b), r_i (c), e d_i (d), quando ajustado incorretamente o modelo von Mises; para $r^*_{q_i}$ (e) e d_i (f), quando ajustado incorretamente o modelo *wrapped* Cauchy.

HDiR do software R (SAAVEDRA-NIEVES; CRUJEIRAS, 2021). Neste trabalho, utilizou-se o subconjunto referente as 229 observações de fêmeas da espécie *Talitrus saltator*, cuja visão de paisagem foi permitida (Figura 2.13 (a)).

Foram ajustados os modelos com resposta von Mises, *sine-skewed* von Mises e *wrapped* Cauchy com função de ligação definida pela Equação 2.25. A Tabela 2.16 apresenta as estimativas dos parâmetros e os valores-p do teste da razão de verossimilhança (TRV) para os três modelos, considerando a temperatura como covariável. Os resultados sugerem que a temperatura é significativa, como encontrado por Scapini *et al.* (2002) e Marchetti e Scapini (2003). De uma forma geral, $\widehat{b}_1 < 0$ implica que as direções médias estimadas sofrem uma translação em sentido horário, à medida que a temperatura aumenta. O mesmo pode ser observado pelos diagramas de dispersão circular da Figura 2.13. A Figura 2.13 (b) representa as direções dos crustáceos, cuja temperatura registrada foi menor ou igual à $23,5^oC$ (mediana das observações da covariável temperatura), enquanto que a Figura 2.13 (c) representa as direções dos crustáceos, cuja temperatura registrada foi maior que $23,5^oC$. Observa-se, portanto, que há uma translação em sentido horário na concentração dos pontos entre os dois gráficos, coerente com a estimativa

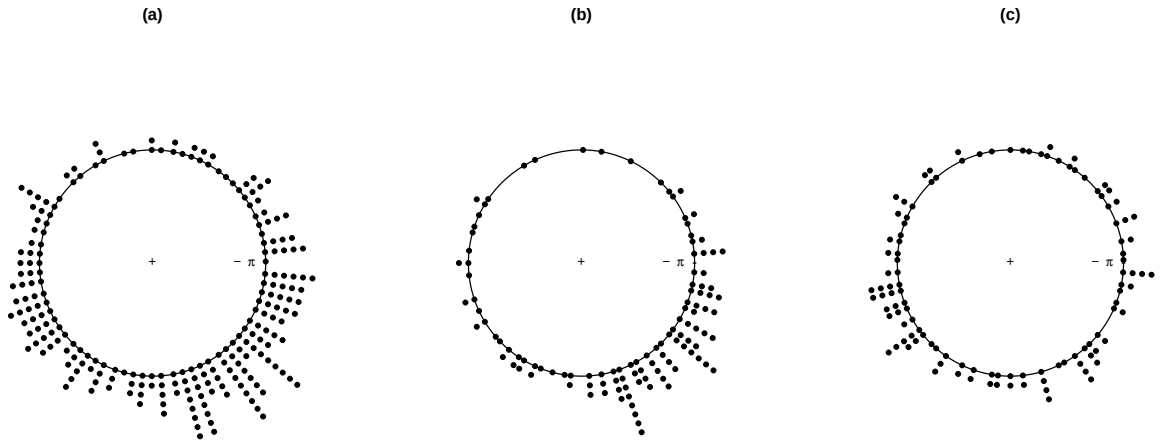

Figura 2.13 – Diagrama de dispersão circular para as 229 orientações dos crustáceos (a), para as observações dos crustáceos, cuja temperatura registrada é menor ou igual a $23,5^oC$ (b) e cuja temperatura registrada é maior que $23,5^oC$ (c).

obtida de $\widehat{b}_1$

Tabela 2.16 – Estimativas dos parâmetros e valor-p do (TRV) para dados de sandhopper.

	von Mises		sine-skewed von Mises		wrapped Cauchy	
	Estimativa	valor-p	Estimativa	valor-p	Estimativa	valor-p
b_0	3,275	$< 0,001$	1,314	0,003	3,487	$< 0,001$
b_1	$-0,083$	0,007	$-0,052$	0,003	$-0,084$	0,013

Para a verificação da adequação dos modelos, foram obtidos os três conjuntos de resíduos. O resíduo quantílico circular indica falhas no ajuste dos modelos com resposta simétrica, já que 13% dos pontos estão fora dos limites do envelope na Figura 2.14 (a), sendo um ponto bem evidente, e pela presença de 31% dos pontos para além dos limites na Figura 2.14 (e). O mesmo não ocorre com os demais resíduos (Figura 2.14 (b)(c)(f)), onde no máximo 5% dos pontos encontram-se fora dos envelopes e todos esses estão bem próximos aos seus limites. Para o modelo de regressão *sine-skewed* von Mises, o resíduo quantílico circular não sugere falhas no ajuste (Figura 2.14 (d)), uma vez que não há pontos visivelmente fora dos limites do envelope e mais de 90% deles de fato se encontram entre as curvas.

2.6 Conclusões

Neste capítulo, propôs-se uma extensão ao resíduo quantílico, adequando seu uso para os modelos de regressão circular. Por meio de estudos de simulação via Monte Carlo, comparou-se a distribuição do resíduo quantílico circular com outros três preexistentes. Com o intuito de investigar se os resíduos podem identificar erros de especificação no modelo, foram implementadas

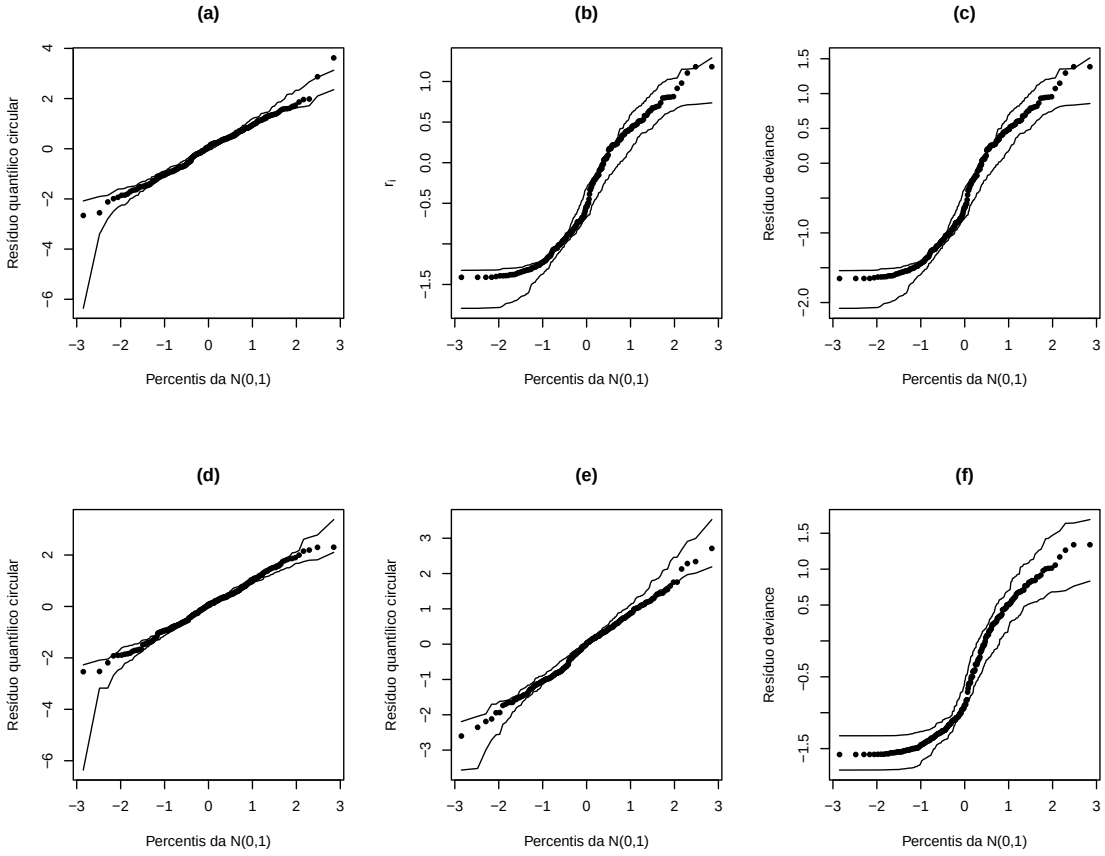

Figura 2.14 – Gráfico de probabilidade normal com envelope simulado para orientação sandhopper e $r^*_{q_i}$ (a), r_i (b), e d_i (c), quando o modelo de regressão von Mises é ajustado; $r^*_{q_i}$ (d), quando o modelo de regressão *sine-skewed* von Mises é ajustado; $r^*_{q_i}$ (e) e d_i (f), quando o modelo de regressão *wrapped* Cauchy é ajustado.

duas aplicações, sendo uma delas com dados simulados e outra com dados reais. Para realizar essas análises, foram introduzidos dois modelos de regressão.

É muito desejável que um resíduo tenha as duas propriedades a seguir. Primeiro, sua distribuição deve ser bem aproximada pela distribuição Normal padrão. Em segundo lugar, deve ser possível detectar a falta de ajuste do modelo. Os estudos de simulação sugerem que há uma boa concordância entre a distribuição do resíduo quantílico circular e a distribuição Normal padrão em todos os cenários, mesmo em pequenas amostras. A aproximação melhora quando o tamanho da amostra aumenta de 20 para 50, o que corrobora o fato de que o resíduo proposto é assintoticamente Normal padrão distribuído em modelos de regressão para dados circulares (como demonstrado algebricamente), assim como com o resíduo quantílico no caso linear. Além disso, o resíduo quantílico circular se destaca dos demais especialmente nos cenários onde a variância circular é menor.

Para os modelos de regressão von Mises, já foi demonstrado que as extensões propostas por Souza e Paula (2002) - isto é, os resíduos d^*_i e r_i - possuem distribuição aproximadamente Normal padrão. Contudo, os estudos de simulação da Seção 2.4.1 indicam que a distribuição do

resíduo d_i^* possui menor concordância com a Normal padrão, quando comparada à distribuição do resíduo *deviance*, uma vez que sua variância se afasta de 1.

Em relação à capacidade de identificar falhas no ajuste, as aplicações mostraram que o resíduo proposto também se sobressai aos demais. A primeira aplicação mostrou que o resíduo quantílico circular sugere mais claramente a falta de especificação do modelo, quando comparado aos demais resíduos. A segunda aplicação destaca a superioridade do $r_{q_i}^*$, especialmente no caso em que a variável resposta é assimétrica.

Além dessas vantagens apontadas pelos estudos de simulação e aplicações, o resíduo proposto também se destaca dos demais em relação à simplicidade e invariabilidade de sua expressão algébrica ao considerar distintas classes de modelos de regressão para dados circulares. Ele pode ser utilizado para qualquer modelo de regressão circular-linear paramétrico que assume observações independentes, incluindo modelos nos quais mais do que um parâmetro da variável resposta é uma função de covariáveis. Também pode ser usado em modelos de regressão circular-linear não paramétricos, nos quais a função de distribuição da variável resposta pode ser estimada, como no modelo proposto por Marzio, Panzera e Taylor (2013). No geral, o resíduo quantílico circular é, portanto, uma escolha melhor para realizar a análise de diagnóstico de regressão circular do que os resíduos concorrentes.

CAPÍTULO

3

DETECÇÃO DE *OUTLIERS* EM MODELOS DE REGRESSÃO BETA INFLACIONADOS EM DOIS E TRÊS PONTOS

A regressão beta (FERRARI; CRIBARI-NETO, 2004) é frequentemente utilizada para modelar a relação entre uma variável contínua do tipo proporção e um conjunto de variáveis explicativas. Nos casos em que o suporte da variável resposta possui um ou mais pontos com massa de probabilidade positiva, o problema recai na classe dos modelos de regressão beta inflacionados em um ou mais pontos (MARTINEZ, 2008; PEREIRA; BOTTER; SANDOVAL, 2013), que também podem ser chamados de modelos de regressão beta ajustados em um ou mais pontos ou modelos de regressão beta aumentados em um ou mais pontos (GALVIS; BANDYOPADHYAY; LACHOS, 2014). A distribuição beta inflacionada considera que o componente contínuo é modelado pela distribuição beta, e o componente discreto por meio de uma distribuição discreta. Na análise de diagnóstico destes modelos, é comum que se utilize o resíduo quantílico aleatorizado (DUNN; SMYTH, 1996), o que é útil para verificar falhas na especificação do modelo. No entanto, pode falhar para identificar *outliers*, conforme observado por Pereira *et al.* (2020) no contexto dos modelos de regressão beta inflacionados em zero. O resíduo quantílico aleatorizado apresenta limitações na detecção destes pontos cujos valores se aproximam da componente discreta. O mesmo ocorre para os modelos beta inflacionados em dois e três pontos, conforme será discutido neste capítulo.

Pereira *et al.* (2020) propuseram um resíduo com boas propriedades para detecção *outliers*, sobretudo quando não são identificados pelo resíduo quantílico aleatorizado. Neste capítulo, é introduzida uma generalização do resíduo desses autores, abrangendo os modelos de regressão beta inflacionados em dois e três pontos, mas que também pode ser facilmente estendida para mais pontos de inflação. Ospina e Ferrari (2012) também propõem um resíduo para avaliação da qualidade do ajuste em modelos de regressão beta inflacionados no zero ou

um, ao qual também é proposta uma adaptação neste capítulo para o contexto dos modelos beta inflacionados truncados. São realizados estudos de simulação Monte Carlo a fim de comparar a distribuição desses resíduos nas caudas com a distribuição Normal padrão.

3.1 Distribuição beta inflacionada

Variáveis do tipo taxa, razão ou proporção são comumente estudadas em contextos como a proporção de jogos ganhos por um tenista profissional (PEREIRA, 2019), proporção de renda gasta em alimentação (FERRARI; CRIBARI-NETO, 2004) e proporção de compras realizadas via dispositivo móvel (WANG; MALTHOUSE; KRISHNAMURTHI, 2015). Nesses casos, é conveniente utilizar a distribuição beta, uma vez que estas quantidades são medidas no intervalo aberto (a,b), com $a,b \in \mathrm{R}$. No âmbito dos modelos de regressão com a distribuição beta, é interessante utilizar a reparametrização proposta por Ferrari e Cribari-Neto (2004) para $a=0$ e $b=1$ e estendida por Pereira, Botter e Sandoval (2012) para a, $b \in \mathrm{R}$, cuja densidade de probabilidade é dada por

$$f_B(y;m,f,a,b) = \frac{\mathrm{G}(f)(y-a)^{\left(\frac{m-a}{b-a}\right)f-1}(b-y)^{\left(\frac{b-m}{b-a}\right)f-1}}{\mathrm{G}\left[\left(\frac{m-a}{b-a}\right)f\right]\mathrm{G}\left[\left(\frac{b-m}{b-a}\right)f\right](b-a)^{f-1}}, \quad y \in (a,b), \tag{3.1}$$

em que $\mathrm{G}(.)$ é a função gama definida como $\mathrm{G}(t) = \int_0^{\infty} x^{t-1}e^{-x}dx$, $m \in (a,b)$ é a média da variável Y e $f > 0$ é um parâmetro de precisão.

Note que a distribuição beta é absolutamente contínua no intervalo aberto (a,b) e, portanto, não abrange os casos em que a variável pode assumir os valores a ou b com alguma probabilidade positiva. Na ecologia, por exemplo, é comum encontrar dados com uma alta proporção de zeros, como no estudo sobre a cobertura percentual de duas famílias de plantas em uma determinada região (TANG *et al.*, 2021). Para o caso $a=0$ e $b=1$, Ospina e Ferrari (2010) introduziram distribuições que contemplam a massa de probablidade em zero, um ou ambos. Para os casos em que a variável pode assumir o valor zero, os autores propõem uma mistura entre a distribuição beta e uma distribuição degenerada em zero, denominada distribuição beta inflacionada em zero (BIZ), cuja função densidade de probabilidade é definida como

$$f_{BIZ}(y;d_0,m,f) = \begin{cases} d_0, & \text{se } y=0, \\ (1-d_0)\,f_B(y;m,f), & \text{se } y \in (0,1), \end{cases} \tag{3.2}$$

em que $d_0 = P(Y=0)$, $m = \mathrm{E}(Y|Y \in (0,1))$, f é um parâmetro de precisão e $f_B(y;m,f)$ é definida como em (3.1), com $a=0$ e $b=1$. A correspondente função distribuição acumulada é dada por

$$F_{BIZ}(y;d_0,m,f) = d_0\mathrm{I}_{[0,\infty)}(y) + (1-d_0)F_B(y;m,f), \tag{3.3}$$

em que $F_B(y;m,f)$ é a função distribuição acumulada de $Y \sim Beta(0,1)$.

Para as proporções observadas no intervalo $[0,1]$, Ospina e Ferrari (2010) propuseram uma distribuição inflacionada com base em uma mistura entre a distribuição beta e a distribuição de Bernoulli, denominada distribuição beta inflacionada em zero e um (BIZU), cuja densidade de probabilidade pode ser escrita como

$$f_{BIZU}(y;d_0,d_1,m,f) = \begin{cases} d_0, & \text{se } y = 0, \\ d_1, & \text{se } y = 1, \\ (1-d_0-d_1)\,f_B(y;m,f), & \text{se } y \in (0,1), \end{cases} \tag{3.4}$$

em que d_0 e d_1 são, respectivamente, $P(Y=0)$ e $P(Y=1)$, $m = \mathrm{E}(Y|Y \in (0,1))$, f é um parâmetro de precisão e $f_B(y;m,f)$ é definida como em (3.1), com $a=0$ e $b=1$. Como conseguinte, a função de distribuição BIZU é dada por

$$F_{BIZU}(y;d_0,d_1,m,f) = d_0\,\mathrm{I}_{[0,¥)}(y) + d_1\,\mathrm{I}_{[1,¥)}(y) + (1-d_0-d_1)F_B(y;m,f), \tag{3.5}$$

em que $F_B(y;m,f)$ é a função distribuição acumulada de $Y \sim Beta(0,1)$.

Para além destes cenários, surgem ainda variáveis cujas observações não podem assumir valores em um determinado intervalo $(0,c)$, sendo $0 < c < 1$. Essas variáveis surgem, por exemplo, na análise de um determinado pagamento limitado entre dois valores, quando estudado em função do seu valor máximo. Em tal caso, tem-se como alternativa para o ajuste deste tipo de variável, a distribuição proposta por Pereira, Botter e Sandoval (2012), que consiste na mistura de uma distribuição beta com suporte no intervalo $(c,1)$ e uma distribuição trinomial que assume os valores 0, 1 e c . Seja $Y \sim BIZUT(d_0,d_1,d_c,m,f,c)$, sua $f.d.p.$ é dada por

$$f_{BIZUT}(y;d_0,d_1,d_c,m,f,c) = \begin{cases} d_0, & \text{se } y = 0, \\ d_1, & \text{se } y = 1, \\ d_c, & \text{se } y = c, \\ (1-d_0-d_1-d_c)\,f_B(y;m,f,c,1), & \text{se } y \in (c,1), \end{cases} \tag{3.6}$$

com função de distribuição descrita como

$$F_{BIZUT}(y;d_0,d_1,d_c,m,f,c) = d_0\,\mathrm{I}_{[0,¥)}(y) + d_c\,\mathrm{I}_{[c,¥)}(y) + d_1\,\mathrm{I}_{[1,¥)}(y) + (1-d_0-d_1)F_B(y;m,f,c,1). \tag{3.7}$$

Os autores nomearam a distribuição como beta inflacionada truncada (BIZUT). Neste trabalho, além desse nome, a distribuição também será referida como beta inflacionada em três pontos.

3.2 Regressão beta inflacionada

Diante destas extensões da distribuição beta, surgem os modelos de regressão beta inflacionada em zero (RBIZ) (OSPINA; FERRARI, 2012), beta inflacionada em zero e um (RBIZU)

(MARTINEZ, 2008) e beta inflacionados truncados (RBIZUT) (PEREIRA; BOTTER; SANDOVAL, 2013), quando há o interesse em estudar essas variáveis em função de algumas outras. Exemplos práticos dessas classes de modelos incluem análises da mortalidade em acidentes de trânsito (OSPINA; FERRARI, 2012), estudos sobre índices de eficiência de um determinado município (PEREIRA; CRIBARI-NETO, 2014) e dados sobre o uso de álcool em adolescentes (LIU; KONG, 2015). A classe de modelos RBIZU assume que a variável resposta do modelo segue distribuição BIZU, em que tanto a média dos valores contínuos quanto a probabilidade de assumir os valores discretos são ajustados em função de variáveis preditoras. Isto é, sejam $Y_1,\ldots,Y_n$ variáveis aleatórias independentes com $Y_i \sim BIZU(d_{i0},d_{i1},m_i,f_i),\forall i=1,\ldots,n$, o modelo RBIZU é definido por

$$\begin{cases} g_1(m_i) = h_{i1} = x_{i1}^T b_1, \\ g_2(f_i) = h_{i2} = x_{i2}^T b_2, \\ H(d_{i0},d_{i1}) = (h_0(d_{i0},d_{i1}),h_1(d_{i0},d_{i1})) = (z_{i0},z_{i1}) = \left(z_{i0}g_0^T, z_{i1}g_1^T\right), \end{cases} \tag{3.8}$$

em que $b_1 = (b_{11},b_{21},\ ...,\ b_{p_m1})^T$, $b_2 = (b_{12},b_{22},\ ...,\ b_{p_f2})^T$, $g_0 = (g_{10},g_{20},\ ...,\ g_{p_00})^T$ e $g_1 = (g_{11},g_{21},\ ...,\ g_{p_11})^T$ são os vetores de parâmetros desconhecidos, $x_{i1} = (x_{i11},x_{i21},\ ...,\ x_{ip_m1})^T$, $x_{i2} = (x_{i12},x_{i22},\ ...,\ x_{ip_f2})^T$, $z_{i0} = (z_{i10},z_{i20},\ ...,\ z_{ip_00})^T$ e $z_{i1} = (z_{i11},z_{i21},\ ...,\ z_{ip_11})^T$ são os vetores que representam os valores das variáveis independentes, g_1 e g_2 são funções de ligação estritamente monótonas e duplamente diferenciáveis de $(0,1)$ em R e de R^+ em R, respectivamente, e H é uma função de ligação bijetora duplamente diferenciável de C em R^2, em que C é um subespaço de R^2 definido como $C = \{(d_{i0},d_{i1}) : 0 < d_{i0} < 1, 0 < d_{i1} < 1 - d_{i0}\}$. Mais informações sobre uma classe de modelos RBIZU ainda mais geral do que a dada em (3.8) podem ser encontradas em (OSPINA; FERRARI, 2012).

Pereira, Botter e Sandoval (2013) propuseram uma extensão ao modelo introduzido por Martinez (2008), permitindo a existência de um terceiro ponto de inflação c_i e cuja variável resposta não assume valores no intervalo $(0,c_i)$. O modelo também permite que o suporte da variável dependente varie entre as unidades populacionais, bem como f varie em função de outras variáveis. Embora esse caso seja menos comum que o modelo com um ou dois pontos de inflação, o modelo tem aplicações em seguro desemprego (PEREIRA; BOTTER; SANDOVAL, 2012) e cartão de crédito (PEREIRA; BOTTER; SANDOVAL, 2013). Se $Y_1,...,Y_n$ são variáveis aleatórias independentes com distribuição $BIZUT(d_{i0},d_{i1},d_{ic},m_i,f_i,c_i)$, os modelos RBIZUT são definidos por

$$\begin{cases} g_1(m_i) = h_{i1} = x_{i1}^T b_1, \\ g_2(f_i) = h_{i2} = x_{i2}^T b_2, \\ H(d_{i0},d_{i1},d_{ic}) = (z_{i0},z_{i1},z_{ic}) = \left(z_{i0}g_0^T, z_{i1}g_1^T, z_{ic}g_c^T\right), \end{cases} \tag{3.9}$$

em que b_1, b_2, g_0, g_1, x_{i1}, x_{i2}, z_{i0} e z_{i1} são os mesmos vetores descritos no modelo RBIZU. Analogamente, $g_c = (g_{1c},g_{2c}\ ...,\ g_{p_cc})^T$ e $z_{ic} = (z_{i1c},z_{i2c},\ ...,\ z_{ip_cc})^T$, $g_1 : (c_i,1) \to \mathrm{R}$ e $g_2 : \mathrm{R}^+ \to$

R são funções estritamente monótonas e duplamente diferenciáveis, e $H : C \to \mathrm{R}^3$ é uma função de ligação bijetora duplamente diferenciável, em que C é um subespaço de R^3 definido como $C = \{(d_{i0}, d_{i1}, d_{ic}) : 0 < d_{i0} < 1, 0 < d_{i1} < 1 - d_{i0}, 0 < d_{ic} < 1 - d_{i0} - d_{i1}\}$.

Os modelos RBIZU e RBIZUT, definidos em (3.8) e (3.9), respectivamente, podem ser estimados pelo método da máxima verossimilhança, por meio de procedimentos numéricos de otimização. Nos estudos de simulação e nas aplicações desse capítulo, foi utilizado o método quase-Newton de Broyden-Fletcher-Goldfarb-Shanno (BFGS) (BROYDEN, 1970; FLETCHER, 1970; GOLDFARB, 1970; SHANNO, 1970). Utilizando-se às propriedades assintóticas do estimador de máxima verossimilhança, podem ser obtidos intervalos de confiança e testes de hipótese sobre os parâmetros, usando por exemplo o teste da razão se verossimilhanças ou o teste de Wald.

3.3 Detecção de *outliers*

Um aspecto importante no processo de construção do modelo de regressão consiste em identificar possíveis *outliers* dentre as observações. Essas quantidades podem ocorrer devido à erros operacionais durante a coleta, medição ou transcrição dos dados. Contudo, também podem ser observações genuínas, capazes de indicar falhas na especificação do modelo. Os *outliers* podem ser entendidos como observações mal ajustadas pelo modelo, cuja presença pode influenciar o ajuste da curva, afetando significativamente as estimativas dos parâmetros e a capacidade preditiva do modelo, ocasionando resultados imprecisos na análise.

Como resultado da necessidade de identificar *outliers* e diante de suas características, é comum a utilização dos resíduos para sua detecção. O uso do resíduo studentizado, por exemplo, é um critério utilizado para detectar um único valor discrepante na regressão linear (PAUL; FUNG, 1991). Um ponto é considerado como possível *outlier* se o valor absoluto do seu resíduo for significativamente maior do que o valor para as demais observações. Nos casos em que a distribuição de probabilidade do resíduo é conhecida, uma observação é identificada como possível *outlier* se o valor absoluto do respectivo resíduo exceder um determinado limite (PEREIRA *et al.*, 2020).

Na classe de modelos RBIZU obtém-se três valores ajustados para cada observação, sendo: estimativa da probabilidade da observação assumir o valor zero, estimativa da probabilidade da observação assumir o valor 1 e estimativa do valor esperado da variável resposta, dado que a observação assume valor no espaço contínuo $(0, 1)$. De maneira análoga, na classe de modelos RBIZUT existem quatro valores ajustados para cada observação, sendo os dois primeiros idênticos ao caso anterior, a estimativa da probabilidade da observação assumir valor c_i e a estimativa do valor esperado da variável resposta, dado que a observação assume valor no espaço contínuo $(c_i, 1)$. Nos casos em que a observação assume um dos pontos discretos, apenas a respectiva estimativa da probabilidade da observação assumir àquele valor é útil. Diferentemente,

quando a observação pertence ao intervalo de valores, todas as estimativas são importantes. Note que o número de valores ajustados de interesse depende do valor assumido pela variável, o que dificulta a definição de um único resíduo para todo o conjunto de pontos (PEREIRA *et al.*, 2020).

3.3.1 Resíduo quantílico

Com o intuito de obter resíduos contínuos no contexto dos modelos de regressão cuja variável resposta é discreta ou possui uma componente discreta, Dunn e Smyth (1996) propuseram uma randomização do resíduo quantílico. Contudo, no contexto dos modelos de regressão inflacionados no zero, Pereira *et al.* (2020) mostraram que o resíduo quantílico apresenta uma certa limitação na detecção de *outliers* cujos valores se aproximam da componente discreta. Isso ocorre pois se a variável resposta for próximo de 0, com $\mathrm{P}(Y_i \backslash= 0)$ não muito pequena e menor que $\frac{1}{2}$, então o resíduo será negativo com valor absoluto não muito alto e, portanto, pode não ser capaz de detectar alguns *outliers*. O mesmo ocorre nos modelos inflacionados no zero e um e inflacionados truncados, conforme será discutido abaixo.

Nos modelos RBIZU, o resíduo quantílico aleatorizado é definido como

$$r_i^q = \begin{cases} \mathrm{F}^{-1}\{u_{i0}\}, & \text{se } y_i = 0, \\ \mathrm{F}^{-1}\{F_{BIZU}(y_i; \widehat{d}_{i0}, \widehat{d}_{i1}, \widehat{m}_i, \widehat{f}_i)\}, & \text{se } y_i \in (0,1), \\ \mathrm{F}^{-1}\{u_{i1}\}, & \text{se } y_i = 1, \end{cases} \tag{3.10}$$

em que $F_{BIZU}(y_i; \widehat{d}_{i0}, \widehat{d}_{i1}, \widehat{m}_i, \widehat{f}_i)$ é a função distribuição definida em (3.5), e u_{i0} e u_{i1} são variáveis aleatórias com distribuição uniforme $U(0, \widehat{d}_{i0})$ e $U(1-\widehat{d}_{i1}, 1)$, respectivamente. Observe que, se $y_i \in (0,1)$, então $F_{BIZU}(y_i; \widehat{d}_{i0}, \widehat{d}_{i1}, \widehat{m}_i, \widehat{f}_i) > \widehat{d}_{i0}$ e, consequentemente, $r_i^q > \mathrm{F}^{-1}(\widehat{d}_{i0})$. Com isso, se y_i pertence a componente contínua e $\widehat{d}_{i0} > \dfrac{1}{2}$, então $r_i^q > 0$. De maneira similar, se $y_i \in (0,1)$ e $\widehat{d}_{i1} > \dfrac{1}{2}$, então $r_i^q < 0$. Pelos gráficos da Figura 3.1(a)(b), pode-se notar que o resíduo quantílico aleatorizado necessariamente não assume valores grandes e negativos (simultaneamente), quando y_i está próximo de zero e $\widehat{d}_{i0}$ não é muito pequeno. Semelhantemente, quando y_i está próximo de 1, o resíduo quantílico aleatorizado necessariamente não assume valores grandes positivos, se $\widehat{d}_{i1}$ não for consideravelmente pequeno. Nesses casos, o r_i^q pode não identificar alguns *outliers* no modelo RBIZU.

De maneira análoga, para a classe RBIZUT, o resíduo quantílico aleatorizado é calculado por

$$r_i^q = \begin{cases} \mathrm{F}^{-1}\{u_{i0}\}, & \text{se } y_i = 0, \\ \mathrm{F}^{-1}\{u_{ic}\}, & \text{se } y_i = c_i, \\ \mathrm{F}^{-1}\{F_{BIZUT}(y_i; \widehat{d}_{i0}, \widehat{d}_{i1}, \widehat{d}_{ic}, \widehat{m}_i, \widehat{f}_i, c_i)\}, & \text{se } y_i \in (c_i, 1), \\ \mathrm{F}^{-1}\{u_{i1}\}, & \text{se } y_i = 1, \end{cases} \tag{3.11}$$

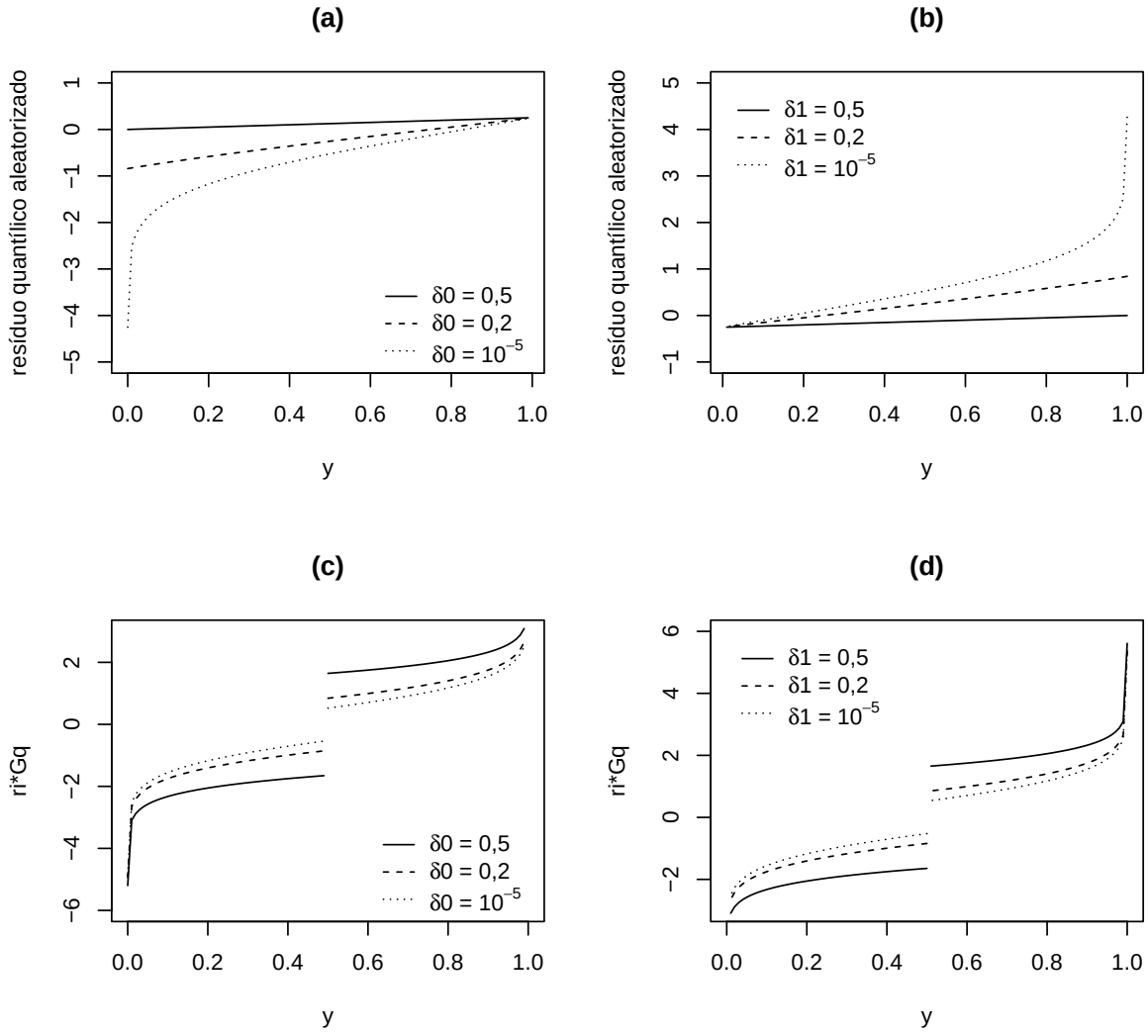

Figura 3.1 – Comportamento dos resíduos r_i^q com $d_1 = 0,4$ fixo e variando os valores de d_0 (a); r_i^q com $d_0 = 0,4$ fixo e variando os valores de d_1 (b); r_i^{*Gq} com $d_1 = 0,4$ fixo e variando os valores de d_0 (c); r_i^{*Gq} com $d_0 = 0,4$ fixo e variando os valores de d_1 (d).

em que $F_{BIZUT}(y_i; \widehat{d}_{i0}, \widehat{d}_{i1}, \widehat{d}_{ic}, \widehat{m}_i, \widehat{f}_i, c_i)$ é a função distribuição definida em (3.7), u_{i0}, u_{ic} e u_{i1} são variáveis aleatórias com distribuição uniforme $U(0, \widehat{d}_{i0})$, $U(\widehat{d}_{i0}, \widehat{d}_{i0} + \widehat{d}_{ic})$ e $U(1 - \widehat{d}_{i1}, 1)$, respectivamente. Novamente, se $y_i \in (c_i, 1)$, então $F_{BIZUT}(y_i; \widehat{d}_{i0}, \widehat{d}_{i1}, \widehat{d}_{ic}, \widehat{m}_i, \widehat{f}_i, c_i) > \widehat{d}_{i0} + \widehat{d}_{ic}$ e, consequentemente, $r_i^q > \mathrm{F}^{-1}(\widehat{d}_{i0} + \widehat{d}_{ic})$. Portanto, conforme já apontado por Pereira (2012), se y_i pertence a componente contínua e $\widehat{d}_{i0}$ e $\widehat{d}_{ic}$ não forem pequenos, r_i^q não poderá ser negativo e apresentar valor grande, em módulo, ao mesmo tempo. Desta forma, se nesse caso y_i for um *outlier* o resíduo quantílico não será capaz de detectá-lo como tal. O mesmo ocorre para os valores próximos de 1.

Um exemplo de situação em que o resíduo quantílico aleatorizado pode deixar de detectar *outliers* ocorre no pagamento de cartão de crédito. Pereira (2012) denominou de valor relativo do pagamento (VRP), a proporção do valor de uma fatura de cartão de crédito que um cliente paga em um determinado mês. Como as administradoras de cartão de crédito estabelecem um

valor mínimo para pagamento da fatura, o VRP pode assumir os valores 0 (se o cliente não paga a fatura), c_i (se ele paga o valor mínimo), 1 (se ele paga o valor total) e qualquer valor real no intervalo $(c_i, 1)$ se ele paga mais que o mínimo e menos que o total. Um cliente com problemas financeiros, tradicionalmente tem dificuldade de pagar o valor igual ou próximo ao total. Porém, para esse cliente $\widehat{d}_{i1}$ não é tão pequeno, pois comumente esse perfil de cliente contrai um empréstimo com taxa de juros mais barata para pagar a fatura do cartão de crédito que geralmente apresenta altas taxas de juros. Porém, raramente esse perfil de cliente pagará um valor próximo do valor total da fatura, já que, quando ele pega outro empréstimo, geralmente ele é capaz de pagar o valor total. Assim, um cliente com dificuldade financeira que tiver um VRP próximo de 1 deveria ser considerado uma observação *outlier*. Porém, se usarmos o resíduo quantílico aleatorizado, a observação não apresentará valor alto do resíduo, pelos motivos discutidos anteriormente.

3.3.2 Resíduo r_i^{**q}

Para os modelos de regressão beta inflacionados no zero ou um, é comum realizar a análise de resíduos dos componentes discreto e contínuo separadamente. Ospina e Ferrari (2012) apresentaram um conjunto de ferramentas de diagnóstico para avaliação da qualidade do ajuste e identificação de possíveis *outliers* no componente discreto e contínuo, separadamente. Os autores propõem uma versão do resíduo de Pearson padronizado para verificação das suposições do modelo e identificação de possíveis *outliers* no componente discreto. Já para componente contínua, os autores definem uma modificação do resíduo ponderado padronizado 2 de Espinheira, Ferrari e Cribari-Neto (2008). Usando conjuntamente as informações de ambos os componentes, os autores sugerem a utilização do resíduo quantílico aleatorizado, cujos resultados indicam bom desempenho para avaliar as suposições do modelo, mas pode falhar em identificar alguns *outliers* (PEREIRA *et al.*, 2020), conforme já discutido.

Estudos de simulação realizados por Pereira *et al.* (2020) sugerem que o resíduo proposto por Ospina e Ferrari (2012) para o componente contínuo do modelo, não é adequado para identificar *outliers* no contexto dos modelos de regressão beta inflacionados em zero. Como o resíduo dos autores foi desenvolvido para o caso dos modelos beta inflacionados em apenas um ponto, e diante da superioridade do resíduo quantílico em relação ao resíduo de Espinheira, Ferrari e Cribari-Neto (2008) (PEREIRA, 2019; PEREIRA *et al.*, 2020), define-se a seguinte extensão para as observações do componente contínuo dos modelos RBIZU e RBIZUT

$$r_i^{**q} = \frac{r_{q_i}}{\sqrt{1-\widehat{a}_i}}, \tag{3.12}$$

em que r_{q_i} é o resíduo quantílico definido em (2.41), considerando $F(y_i; \widehat{m}_i, \widehat{f}_i)$ a função distribuição acumulada de uma variável aleatória beta distribuída, e $1-\widehat{a}_i$ equivale a probabilidade de y_i pertencer a componente contínua do modelo. Ou seja, no caso dos modelos RBIZU, $1-\widehat{a}_i = 1-\widehat{d}_{i0}-\widehat{d}_{i1}$, e na classe RBIZUT, $1-\widehat{a}_i = 1-\widehat{d}_{i0}-\widehat{d}_{i1}-\widehat{d}_{ic}$. Basicamente, a diferença

do resíduo r_i^{**q} em relação ao proposto por Ospina e Ferrari (2012) é o uso do resíduo quantílico ao invés do resíduo ponderado padronizado 2 (ESPINHEIRA; FERRARI; CRIBARI-NETO, 2008). Note que o resíduo quantílico utilizado aqui não é a versão aleatorizada, sendo assim o numerador de r_i^{**q} se refere apenas ao componente contínuo.

3.3.3 Resíduo r_i^*

No contexto dos modelos de regressão inflacionados no zero, Pereira *et al.* (2020) introduziram uma classe de resíduos para identificação de *outliers* na componente contínua. O resíduo proposto é uma função de $\widehat{d}_{i0}$ e qualquer resíduo r_i utilizado em modelos de regressão para variáveis de resposta contínua, definido como

$$r_i^* = \begin{cases} \mathrm{F}^{-1}[(1-\mathrm{F}(|r_i|))(1-\widehat{d}_{i0})], & \text{se } r_i < 0, \\ \mathrm{F}^{-1}[1-(1-\mathrm{F}(|r_i|))(1-\widehat{d}_{i0})], & \text{se } r_i \geq 0, \end{cases} \tag{3.13}$$

e que pode ser reescrito como

$$r_i^* = \begin{cases} \mathrm{F}^{-1}[\mathrm{F}(r_i)(1-\widehat{d}_{i0})], & \text{se } r_i < 0, \\ \mathrm{F}^{-1}[\widehat{d}_{i0}+\mathrm{F}(r_i)(1-\widehat{d}_{i0})], & \text{se } r_i \geq 0. \end{cases} \tag{3.14}$$

Os próprios autores apontam o resíduo quantílico como uma escolha natural para r_i. Note que, nesse caso, não é necessário a utilização da aleatorização, uma vez que o resíduo r_i^* se refere apenas à integrante contínua. Nesse caso, o resíduo foi nomeado como resíduo quantílico ajustado no zero (r_i^{*q}), para os quais são válidos os seguintes teoremas:

Teorema 3.3.1. Se $r_{q_i} > 0$, então $r_i^{*q} = r_i^q$.

Teorema 3.3.2. Se r_i possui distribuição Normal padrão e d_{i0} é conhecido $\forall i$, tal que $\forall k > \mathrm{F}^{-1}\left[1-\frac{1}{2}(1-d_{i0})\right] = \mathrm{F}^{-1}\left[\frac{1}{2}+\frac{1}{2}d_{i0}\right]$,

$$\mathrm{P}(r_i^* < -k) = \mathrm{P}(r_i^* > k) = 1-\mathrm{F}(k).$$

O Teorema 3.3.1 sugere que r_i^{*q} possa ser interpretado como uma correção do resíduo quantílico aleatorizado, quando $r_{q_i} < 0$, para identificação de *outliers*. O Teorema 3.3.2 garante que, se $r_i \sim N(0,1)$ e d_{i0} é conhecido, nas caudas o resíduo r_i^* possui comportamento semelhante à uma variável aleatória com distribuição Normal. Essa característica é interessante, uma vez que facilita a definição de um limiar de classificação para o resíduo.

Embora, em geral, a distribuição de r_i não seja idêntica à Normal nas caudas, sobretudo no contexto de amostras pequenas, estudos de simulação realizados pelos autores sugeriram que o resíduo quantílico ajustado no zero possui nas caudas propriedades semelhantes à variáveis com distribuição $N(0,1)$. Bem como, as aplicações apresentadas indicaram que o resíduo r_i^{*q}

pode identificar observações *outliers* em cenários que o resíduo quantílico aleatorizado falha na identificação.

3.4 Extensão proposta

Diante das limitações do resíduo quantílico aleatorizado na identificação de *outliers* no contexto dos modelos RBIZU e RBIZUT, mencionadas na Seção 3.3.1, é conveniente definir um resíduo diferente para cada componente do modelo (PEREIRA *et al.*, 2020). Isso posto, será realizado uma extensão à classe de resíduos proposta por Pereira *et al.* (2020), adequando o seu uso para identificação de observações *outliers* no componente contínuo dos modelos de regressão beta inflacionados em dois e três pontos. Baseado nas equações (3.13) e (3.14), para a classe de modelos RBIZU e RBIZUT, propõe-se a seguinte extensão:

$$r_i^{*G} = \begin{cases} \mathrm{F}^{-1}[(1-\mathrm{F}(|r_i|))(1-\widehat{a}_i)], & \text{se } r_i < 0, \\ \mathrm{F}^{-1}[1-(1-\mathrm{F}(|r_i|))(1-\widehat{a}_i)], & \text{se } r_i \geq 0, \end{cases} \tag{3.15}$$

em que $1-\widehat{a}_i = 1-\widehat{d}_{i0}-\widehat{d}_{i1}$ no caso dos modelos RBIZU e $1-\widehat{a}_i = 1-\widehat{d}_{i0}-\widehat{d}_{i1}-\widehat{d}_{ic}$ no caso dos modelos RBIZUT. Note que o resíduo definido pela Equação (3.15) consiste em uma generalização do resíduo r_i^*, podendo, inclusive, abranger os modelos de regressão beta inflacionados em $p \in \mathrm{N}$ pontos. Por esse motivo sua notação recebe o super índice G. Se $r_i < 0$, então por propriedade da distribuição Normal padrão, tem-se que $1-\mathrm{F}(|r_i|) = \mathrm{F}(r_i)$, portanto

$$\mathrm{F}^{-1}[(1-\mathrm{F}(|r_i|))(1-\widehat{a}_i)] = \mathrm{F}^{-1}[\mathrm{F}(r_i)(1-\widehat{a}_i)]. \tag{3.16}$$

Se $r_i > 0$, então $|r_i| = r_i$ e

$$\mathrm{F}^{-1}[1-(1-\mathrm{F}(|r_i|))(1-\widehat{a}_i)] = \mathrm{F}^{-1}[\widehat{a}_i+\mathrm{F}(r_i)(1-\widehat{a}_i)]. \tag{3.17}$$

Portanto, pelas equações (3.16) e (3.17), pode-se reescrever r_i^{*G} como

$$r_i^{*G} = \begin{cases} \mathrm{F}^{-1}[\mathrm{F}(r_i)(1-\widehat{a}_i)], & \text{se } r_i < 0, \\ \mathrm{F}^{-1}[\widehat{a}_i+\mathrm{F}(r_i)(1-\widehat{a}_i)], & \text{se } r_i \geq 0. \end{cases} \tag{3.18}$$

Novamente, tem-se que o resíduo quantílico é uma escolha natural para r_i, obtendo r_i^{*Gq}. Os gráficos da Figura 3.1(c)(d) exibem o comportamento do resíduo r_i^{*Gq} considerando diferentes combinações de d_0 e d_1. Note que as limitações apresentadas pelo resíduo quantílico aleatorizado são corrigidas. Embora o Teorema 3.3.1 não seja válido nesse contexto dos modelos beta inflacionados em mais de um ponto, o Teorema 3.3.2 pode ser estendido da seguinte forma:

Teorema 3.4.1. Se r_i possui distribuição Normal padrão e a_i é conhecido $\forall i$, tal que $\forall k > \mathrm{F}^{-1}\left[1-\frac{1}{2}(1-a_i)\right]$,

$$\mathrm{P}(r_i^{*G} < -k) = \mathrm{P}(r_i^{*G} > k) = 1 - \mathrm{F}(k).$$

A demonstração deste teorema pode ser feita por meio de dois passos:

(I) $\mathrm{P}(r_i^{*G} < -k) = 1 - \mathrm{F}(k)$.

(II) $\mathrm{P}(r_i^{*G} > k) = 1 - \mathrm{F}(k)$.

Prova de (I)

Para o caso RBIZU, considere $c_i = 0, \forall i = 1, ..., n$. Note que $k > 0$ e, portanto

$$\begin{aligned}
\mathrm{P}(r_i^{*G} < -k) &= \mathrm{P}[\mathrm{F}^{-1}[\mathrm{F}(r_i)(1-a_i)] < -k,\ y_i \in (c_i, 1)] \\
&= \mathrm{P}[\mathrm{F}(r_i)(1-a_i) < \mathrm{F}(-k),\ y_i \in (c_i, 1)] \\
&= \mathrm{P}\left[\mathrm{F}(r_i) < \frac{\mathrm{F}(-k)}{1-a_i},\ y_i \in (c_i, 1)\right] \\
&= \mathrm{P}\left[r_i < \mathrm{F}^{-1}\left(\frac{\mathrm{F}(-k)}{1-a_i}\right),\ y_i \in (c_i, 1)\right] \\
&= \mathrm{F}\left[\mathrm{F}^{-1}\left(\frac{\mathrm{F}(-k)}{1-a_i}\right)\right] \mathrm{P}(y_i \in (c_i, 1)) \\
&= \left(\frac{\mathrm{F}(-k)}{1-a_i}\right)(1-a_i) \\
&= \mathrm{F}(-k) \\
&= 1 - \mathrm{F}(k).
\end{aligned}$$

Prova de (II)

De maneira análoga,

$$\begin{aligned}
\mathrm{P}(r_i^{*} > k) &= \mathrm{P}[\mathrm{F}^{-1}[a_i + \mathrm{F}(r_i)(1-a_i)] > k,\ y_i \in (c_i, 1)] \\
&= \mathrm{P}[a_i + \mathrm{F}(r_i)(1-a_i) > \mathrm{F}(k),\ y_i \in (c_i, 1)] \\
&= \mathrm{P}\left[r_i > \mathrm{F}^{-1}\left(\frac{\mathrm{F}(k)-a_i}{1-a_i}\right),\ y_i \in (c_i, 1)\right] \\
&= \mathrm{P}\left[r_i > \mathrm{F}^{-1}\left(\frac{\mathrm{F}(k)-a_i}{1-a_i}\right)\right] \mathrm{P}[y_i \in (c_i, 1)] \\
&= \left(1 - \mathrm{F}\left[\mathrm{F}^{-1}\left(\frac{\mathrm{F}(k)-a_i}{1-a_i}\right)\right]\right)(1-a_i) \\
&= \left(1 - \frac{\mathrm{F}(k)-a_i}{1-a_i}\right)(1-a_i) \\
&= 1 - a_i - \mathrm{F}(k) + a_i \\
&= 1 - \mathrm{F}(k).
\end{aligned}$$

Portanto, por (I) e (II) tem-se que o Teorema 3.4.1 é válido.

3.5 Estudos de simulação

Nessa seção são apresentados os estudos de simulação Monte Carlo para avaliar a distribuição caudal dos resíduos propostos, nos modelos RBIZU e RBIZUT em diferentes situações. Considerou-se o caso particular dos modelos de regressão logística beta inflacionada (RLBIZU) e regressão logística beta inflacionada truncada (RLBIZUT), definidos pelas Equações 3.19 e 3.20, respectivamente.

Para cada modelo, foram avaliadas diferentes condições derivadas de um cenário base, por meio de variações nos parâmetros ou na distribuição atribuída às covariáveis. Para todos os casos, foram realizadas 25000 réplicas de Monte Carlo com $n = 100$, mantendo as variáveis preditoras fixas ao longo de todas as réplicas.

De maneira semelhante ao desenvolvido por Pereira *et al.* (2020), o desempenho do resíduo r_i^{*Gq} foi comparado à extensão do resíduo proposto por Ospina e Ferrari (2012), definida pela Equação (3.12). Para cada conjunto de resíduos foram calculadas as porcentagens de resíduos menores que $-3, -2, -1$ e maiores que $1, 2, 3$, entre as 25000 réplicas. Quando y assume algum dos pontos discretos, os resíduos não são definidos e, portanto, considerou-se que eles não são menores que -1 ou maiores que 1. Com o intuito de analisar as similaridades entre o comportamento do resíduo proposto com a distribuição Normal padrão, foram calculadas as estatísticas descritivas das porcentagens mencionadas acima, comparando-as aos valores teóricos da $N(0,1)$.

3.5.1 Estudos de simulação para a classe RBIZU

Nos estudos de simulação para o caso em que os dados são observados no intervalo $[0,1]$, será considerado o modelo RLBIZU com a seguinte estrutura:

$$\begin{cases} \ln\left(\dfrac{m_i}{1-m_i}\right) = b_{11} + b_{21}x_{i21} + b_{31}x_{i31}, \\ \ln(f_i) = b_{12} + b_{22}x_{i22} + b_{32}x_{i32}, \\ \left(\ln\left(\dfrac{d_{i0}}{1-a_i}\right), \ln\left(\dfrac{d_{i1}}{1-a_i}\right)\right) = (z_{i0}, z_{i1}), \end{cases} \tag{3.19}$$

em que $1 - a_i = 1 - d_{i0} - d_{i1}$ e $(z_{i0}, z_{i1}) = (g_{10} + g_{20}z_{i20} + g_{30}z_{i30}, g_{11} + g_{21}z_{i21} + g_{31}z_{i31})$.

O cenário I foi construído com base em Pereira *et al.* (2020), em que: $b_{11} = -1,5$, $b_{21} = -1,0$, $b_{31} = 1,0$, $b_{12} = 4$, $b_{22} = b_{32} = 0$, $g_{10} = -0,5$, $g_{20} = 0,5$, $g_{30} = -1,0$, $g_{11} = -1,0$, $g_{21} = -0,5$, $g_{31} = 0,5$, resultando em $m \in (0,09,\ 0,34)$, $f_i = 54,60, \forall i = 1, ..., n$, $d_0 \in (0,13,\ 0,42)$, $d_1 \in (0,11,\ 0,31)$. As variáveis explicativas x_{i21} e x_{i31} foram geradas por meio de sorteios independentes da distribuição $U(0,1)$, e assumiu-se $x_{i22} = z_{i20} = z_{i21} = x_{i21}$ e $x_{i32} = z_{i30} = z_{i31} = x_{i31}$. A partir deste cenário base, foram construídos outros 4 cenários, cujas alterações são apresentadas de maneira resumida na Tabela 3.1. O cenário II consiste em aumentar o intervalo de valores possíveis para d_0 e, para isto, considerou-se $g_{10} = 0,3$, $g_{11} = -0,5$ e

$g_{31} = 0,2$, resultando em $d_0 \in (0,24,\ 0,59)$, $d_1 \in (0,11,\ 0,31)$. No cenário III, as covariáveis x_{i21} e x_{i31} foram geradas por meio de sorteios independentes das distribuições $N(1/2, 1/12)$ e $Gamma(3, 1/6)$, respectivamente. No cenário IV, reduziu-se o valor de f_i por meio de $b_{12} = 3$, o que resultou em $f_i = 20,08, \forall i = 1, ..., n$. Por fim, o cenário V considera o modelo sob precisão variável, por meio de $b_{12} = 4$, $b_{22} = -1/2$ $b_{32} = 1/2$, em que $f_i \in (35,58,\ 83,41)$ e $\bar{f}_i = 55,94$.

Tabela 3.1 – Alterações feitas nos cenários em relação ao cenário I, para o modelo RLBIZU.

Cenário	Alterações
II	aumentou-se o intervalo de valores de d_0
III	mudança na distribuição das variáveis preditoras
IV	redução do valor de f_i
V	f_i variável

A Tabela 3.2 apresenta os resultados das simulações dos cenários I-V. A média e a mediana das porcentagens calculadas para o resíduo r_i^{*Gq} são próximas aos valores teóricos da distribuição Normal padrão. Considerando a variabilidade amostral, mesmo os quartis e valores mínimos e máximos não são distantes dos valores de referência, em geral. Dessa forma, há indícios de que a distribuição de r_i^{*Gq} nas caudas seja próxima à distribuição Normal padrão para $n = 100$, independente do valor das variáveis preditoras. Não há diferenças significativas entre os cenários, o que sugere que a distribuição do resíduo r_i^{*Gq} nas caudas seja invariante ao aumento da probabilidade de ocorrência de zeros na amostra, à mudanças na distribuição das covariáveis e ao aumento da variabilidade dos dados. Por outro lado, ao considerar o modelo com precisão variável, as porcentagens calculadas para r_i^{*Gq} apresentam-se um pouco mais longe dos valores teóricos da distribuição Normal padrão do que no Cenário I. Portanto, assim como também observado por Pereira *et al.* (2020), quando incluídas variáveis explicativas no submodelo para o parâmetro f_i, a distribuição do resíduo nas caudas se distancia um pouco da Normal padrão. Já para o resíduo r_i^{**q}, tanto a média quanto a mediana das porcentagens calculadas ficam distantes dos valores teóricos da distribuição Normal padrão, indicando uma distribuição um pouco mais achatada do que o desejado.

3.5.2 Estudos de simulação para a classe RBIZUT

Os estudos de simulação considerando a classe de modelos de regressão beta inflacionados em três pontos foram conduzidos de maneira análoga ao caso anterior, cujo modelo RLBIZUT possui a seguinte estrutura:

Tabela 3.2 – Estatísticas descritivas para porcentagem de resíduos em cada intervalo - modelo de regressão BIZU - cenários I-V.

Cenário	Resíduos	Intervalo residual	Valor teórico	Mínimo	Q_1	Mediana	Média	Q_3	Máximo
I	r_i^{*Gq}	< -3	0,13	0,04	0,09	0,11	0,11	0,12	0,18
		< -2	2,28	1,80	2,16	2,26	2,26	2,36	2,62
		< -1	15,87	15,46	15,85	15,96	15,98	16,12	16,50
		> 1	15,87	15,31	15,78	15,96	15,96	16,15	16,52
		> 2	2,28	1,82	2,20	2,29	2,29	2,38	2,61
		> 3	0,13	0,06	0,10	0,11	0,11	0,12	0,19
	r_i^{**q}	< -3	0,13	0,52	0,65	0,71	0,72	0,78	0,92
		< -2	2,28	3,25	3,69	3,83	3,82	3,95	4,18
		< -1	15,87	11,33	12,19	12,58	12,52	12,80	13,46
		> 1	15,87	11,06	12,19	12,56	12,50	12,85	13,41
		> 2	2,28	3,32	3,71	3,85	3,83	3,99	4,21
		> 3	0,13	0,56	0,68	0,75	0,74	0,80	1,04
II	r_i^{*Gq}	< -3	0,13	0,03	0,07	0,09	0,09	0,11	0,15
		< -2	2,28	1,92	2,18	2,30	2,29	2,40	2,66
		< -1	15,87	13,06	15,29	15,58	15,37	15,82	16,31
		> 1	15,87	12,80	15,07	15,52	15,33	15,86	16,37
		> 2	2,28	1,90	2,21	2,30	2,30	2,43	2,70
		> 3	0,13	0,05	0,09	0,10	0,11	0,12	0,20
	r_i^{**q}	< -3	0,13	0,76	1,05	1,20	1,17	1,29	1,43
		< -2	2,28	3,43	3,97	4,14	4,10	4,27	4,57
		< -1	15,87	8,48	9,93	10,54	10,39	10,98	11,68
		> 1	15,87	8,23	9,84	10,54	10,35	10,95	11,65
		> 2	2,28	3,57	3,94	4,12	4,11	4,30	4,65
		> 3	0,13	0,79	1,05	1,20	1,19	1,33	1,50
III	r_i^{*Gq}	< -3	0,13	0,02	0,08	0,10	0,10	0,12	0,18
		< -2	2,28	1,64	2,20	2,30	2,28	2,41	2,59
		< -1	15,87	15,35	15,80	15,97	15,98	16,17	16,56
		> 1	15,87	15,00	15,79	15,97	15,96	16,14	16,59
		> 2	2,28	1,74	2,21	2,32	2,29	2,41	2,58
		> 3	0,13	0,05	0,09	0,11	0,11	0,13	0,17
	r_i^{**q}	< -3	0,13	0,42	0,67	0,75	0,73	0,79	0,91
		< -2	2,28	2,88	3,74	3,89	3,84	4,00	4,22
		< -1	15,87	11,26	12,30	12,54	12,50	12,73	13,25
		> 1	15,87	10,98	12,26	12,58	12,49	12,78	13,31
		> 2	2,28	2,98	3,73	3,90	3,84	3,99	4,36
		> 3	0,13	0,44	0,68	0,76	0,74	0,80	0,94
IV	r_i^{*Gq}	< -3	0,13	0,04	0,08	0,10	0,10	0,12	0,18
		< -2	2,28	1,82	2,14	2,24	2,25	2,36	2,59
		< -1	15,87	15,51	15,87	15,99	16,01	16,14	16,50
		> 1	15,87	15,32	15,77	15,95	15,94	16,15	16,52
		> 2	2,28	1,84	2,20	2,30	2,30	2,38	2,61
		> 3	0,13	0,06	0,10	0,11	0,11	0,12	0,19
	r_i^{**q}	< -3	0,13	0,51	0,64	0,70	0,71	0,77	0,89
		< -2	2,28	3,25	3,69	3,81	3,82	3,96	4,16
		< -1	15,87	11,34	12,21	12,62	12,54	12,82	13,47
		> 1	15,87	11,02	12,18	12,55	12,50	12,86	13,44
		> 2	2,28	3,35	3,71	3,85	3,84	3,99	4,22
		> 3	0,13	0,57	0,69	0,75	0,75	0,80	1,04
V	r_i^{*Gq}	< -3	0,13	0,00	0,03	0,06	0,07	0,10	0,20
		< -2	2,28	1,56	2,10	2,23	2,25	2,42	2,78
		< -1	15,87	15,69	16,09	16,20	16,22	16,35	16,70
		> 1	15,87	15,45	16,07	16,22	16,22	16,39	16,79
		> 2	2,28	1,48	2,08	2,24	2,24	2,44	2,75
		> 3	0,13	0,00	0,03	0,05	0,06	0,09	0,22
	r_i^{**q}	< -3	0,13	0,32	0,54	0,63	0,64	0,74	0,93
		< -2	2,28	2,93	3,70	3,87	3,89	4,09	4,42
		< -1	15,87	11,50	12,51	12,86	12,79	13,07	13,72
		> 1	15,87	11,57	12,52	12,84	12,80	13,12	13,66
		> 2	2,28	2,96	3,74	3,94	3,89	4,09	4,42
		> 3	0,13	0,25	0,55	0,65	0,65	0,74	0,97

$$\begin{cases} \ln\left(\dfrac{m_i - c_i}{1 - m_i}\right) = b_{11} + b_{21}x_{i21} + b_{31}x_{i31}, \\ \ln(f_i) = b_{12} + b_{22}x_{i22} + b_{32}x_{i32}, \\ \left(\ln\left(\dfrac{d_{i0}}{1-a_i}\right), \ln\left(\dfrac{d_{i1}}{1-a_i}\right), \ln\left(\dfrac{d_{ic}}{1-a_i}\right)\right) = (z_{i0}, z_{i1}, z_{ic}), \end{cases} \tag{3.20}$$

em que $1 - a_i = 1 - d_{i0} - d_{i1} - d_{ic}$ e $(z_{i0}, z_{i1}, z_{ic}) = (g_{10} + g_{20}z_{i20} + g_{30}z_{i30}, g_{11} + g_{21}z_{i21} + g_{31}z_{i31}, g_{1c} + g_{2c}z_{i2c} + g_{3c}z_{i3c})$.

Os cenários foram construídos de maneira semelhante ao caso RBIZU. Para o cenário I, considerou-se $c_i = 0,2$, $\forall i = 1, ..., n$, $b_{11} = -1,5$, $b_{21} = -1,0$, $b_{31} = 1,0$, $b_{12} = 4$, $b_{22} = b_{32} = 0$, $g_{10} = -0,5$, $g_{20} = 0,5$, $g_{30} = -1,0$, $g_{11} = -1,0$, $g_{21} = -0,5$,$g_{31} = 0,5$, , $g_{1c} = -0,5$, $g_{2c} = -1$,$g_{3c} = 0,5$ resultando em $m \in (0,27;\ 0,47)$, $f_i = 54,60, \forall i = 1, ..., n$, $d_0 \in (0,13;\ 0,42)$, $d_1 \in (0,11;\ 0,31)$ e $d_c \in (0,10;\ 0,32)$. As variáveis explicativas x_{i21} e x_{i31} foram geradas por meio de sorteios independentes da distribuição $U(0,1)$, e assumiu-se $x_{i22} = z_{i20} = z_{i21} = z_{i2c} = x_{i21}$ e $x_{i32} = z_{i30} = z_{i31} = z_{i3c} = x_{i31}$. A partir deste cenário base, foram construídos outros 6 cenários, cujas alterações são apresentadas de maneira resumida na Tabela 3.3. Os cenários Ia e Ib trazem modificações no terceiro ponto de inflação c_i, sendo $c_i = 0,5$, $\forall i = 1, ..., n$ e c_i gerado a partir de uma distribuição $U(1/5, 1/2)$, respectivamente. O cenário II consiste em aumentar o intervalo de valores possíveis para d_0 e, para isto, considerou-se $g_{10} = 0,3$, $g_{11} = -0,5$, $g_{31} = 0,2$, $g_{2c} = -0,5$, $g_{3c} = 0,65$, resultando em $d_0 \in (0,16;\ 0,53)$, mantidos os intervalos de d_1 e d_c como no cenário base. Nos cenários III, IV e V foram realizadas alterações idênticas ao caso RBIZU.

Tabela 3.3 – Alterações feitas nos cenários em relação ao cenário I, para o modelo RLBIZUT.

Cenário	Alterações
Ia	aumentou-se o valor da constante c
Ib	c_i variável
II	aumentou-se o intervalo de valores de d_0
III	mudança na distribuição das variáveis preditoras
IV	redução do valor de f_i
V	f_i variável

A Tabela 3.4 apresenta os resultados da simulação para os cenários I, Ia e Ib. A média e a mediana das porcentagens calculadas para o resíduo r_i^{*Gq} são próximas aos valores teóricos da distribuição Normal padrão. Considerando a variabilidade amostral, mesmo os quartis e valores mínimos e máximos não são distantes dos valores de referência, em geral. Dessa forma, há indícios de que a distribuição de r_i^{*Gq} nas caudas seja próxima à distribuição Normal padrão para $n = 100$, independente do valor das variáveis preditoras. Assim como observado nas simulações para o modelo de regressão RBIZU, a distribuição do resíduo r_i^{*Gq} nas caudas é mais próxima da distribuição Normal padrão, quando comparada à distribuição do resíduo r_i^{**q} nas caudas, para todos os intervalos e em todos os cenários. Os resultados para os dois resíduos

não sofrem alteração quando apenas modificado o valor de c_i, considerando-o constante para todas as observações. Isso se deve ao fato das estimativas do modelo também serem mantidas inalteradas. No caso em que a constante varia de acordo com as covariáveis, os resultados possuem uma inexpressiva diferença. Os resultados para os cenários II-V (Tabela 3.5) são similares ao observado no caso RBIZU.

Tabela 3.4 – Estatísticas descritivas para porcentagem de resíduos em cada intervalo - modelo de regressão BIZUT - cenários I, Ia e Ib

Cenário	Resíduos	Intervalo residual	Valor teórico	Mínimo	Q_1	Mediana	Média	Q_3	Máximo
I	r_i^{*Gq}	< -3	0,13	0,04	0,08	0,10	0,10	0,12	0,18
		< -2	2,28	1,95	2,18	2,27	2,28	2,40	2,58
		< -1	15,87	14,73	15,55	15,80	15,74	15,92	16,63
		> 1	15,87	14,79	15,52	15,68	15,67	15,91	16,38
		> 2	2,28	2,06	2,24	2,31	2,32	2,39	2,61
		> 3	0,13	0,06	0,08	0,11	0,11	0,12	0,18
	r_i^{**q}	< -3	0,13	0,78	0,96	1,04	1,04	1,13	1,24
		< -2	2,28	3,56	3,94	4,09	4,07	4,21	4,52
		< -1	15,87	9,53	10,59	11,08	10,98	11,38	11,88
		> 1	15,87	9,51	10,61	10,97	10,96	11,36	11,88
		> 2	2,28	3,63	4,00	4,10	4,10	4,22	4,48
		> 3	0,13	0,85	1,02	1,09	1,08	1,14	1,29
Ia	r_i^{*Gq}	< -3	0,13	0,04	0,08	0,10	0,10	0,12	0,18
		< -2	2,28	1,95	2,18	2,27	2,28	2,40	2,58
		< -1	15,87	14,73	15,55	15,80	15,74	15,92	16,63
		> 1	15,87	14,79	15,52	15,68	15,67	15,91	16,38
		> 2	2,28	2,06	2,24	2,31	2,32	2,39	2,61
		> 3	0,13	0,06	0,08	0,11	0,11	0,12	0,18
	r_i^{**q}	< -3	0,13	0,78	0,96	1,04	1,04	1,13	1,24
		< -2	2,28	3,56	3,94	4,09	4,07	4,21	4,52
		< -1	15,87	9,53	10,59	11,08	10,98	11,38	11,88
		> 1	15,87	9,51	10,61	10,97	10,96	11,36	11,88
		> 2	2,28	3,63	4,00	4,10	4,10	4,22	4,48
		> 3	0,13	0,85	1,02	1,09	1,08	1,14	1,29
Ib	r_i^{*Gq}	< -3	0,13	0,04	0,08	0,09	0,10	0,12	0,18
		< -2	2,28	1,91	2,19	2,26	2,28	2,39	2,58
		< -1	15,87	14,79	15,54	15,80	15,73	15,92	16,53
		> 1	15,87	14,80	15,52	15,69	15,67	15,89	16,39
		> 2	2,28	2,06	2,23	2,31	2,31	2,40	2,60
		> 3	0,13	0,05	0,09	0,11	0,11	0,12	0,19
	r_i^{**q}	< -3	0,13	0,80	0,96	1,04	1,04	1,12	1,22
		< -2	2,28	3,58	3,94	4,08	4,08	4,21	4,51
		< -1	15,87	9,50	10,56	11,02	10,98	11,38	11,84
		> 1	15,87	9,54	10,61	10,96	10,96	11,37	11,86
		> 2	2,28	3,62	4,00	4,10	4,10	4,22	4,48
		> 3	0,13	0,86	1,02	1,09	1,08	1,14	1,27

3.6 Conclusões

Neste capítulo foram apresentadas duas extensões de resíduos para verificação da qualidade do ajuste e detecção de possíveis pontos *outliers*, no contexto dos modelos de regressão

beta inflacionados em um ou mais pontos. Para a classe de resíduos r_i^{*G} foi demonstrado um resultado algébrico, que garante um comportamento caudal semelhante à distribuição $N(0,1)$, dadas algumas premissas. Por meio de estudos de simulação Monte Carlo, comparou-se a distribuição caudal desses resíduos à distribuição Normal padrão, uma vez que conhecido o seu comportamento, pode-se classificar uma observação como *outlier* caso o valor absoluto do seu respectivo resíduo seja, por exemplo e em módulo, superior a 3.

Os resultados do estudo de simulação sugerem que a extensão proposta para o resíduo de Pereira *et al.* (2020) (r_i^{*Gq}) possui distribuição caudal semelhantes à uma variável Normal padrão distribuída. Em conjunto com o Teorema 3.4.1 demonstrado, os resultados sugerem que esta classe de resíduos pode identificar *outliers* em situações que o resíduo quantílico aleatorizado não consegue detectá-los. Em comparação com a adaptação proposta ao resíduo de Ospina e Ferrari (2012), o resíduo r_i^{*Gq} parece ser uma melhor opção para identificação de *outliers* em modelos de regressão beta inflacionados em um ou mais pontos.

Tabela 3.5 – Estatísticas descritivas para porcentagem de resíduos em cada intervalo - modelo de regressão BIZUT - cenários II-V

Cenário	Resíduos	Intervalo residual	Valor teórico	Mínimo	Q_1	Mediana	Média	Q_3	Máximo
II	r_i^{*Gq}	< -3	0,13	0,02	0,07	0,09	0,09	0,11	0,15
		< -2	2,28	1,86	2,22	2,32	2,31	2,41	2,64
		< -1	15,87	11,99	13,54	14,21	14,03	14,55	15,07
		> 1	15,87	11,75	13,58	14,06	13,96	14,57	15,22
		> 2	2,28	1,78	2,21	2,35	2,34	2,46	2,75
		> 3	0,13	0,05	0,08	0,10	0,10	0,12	0,17
	r_i^{**q}	< -3	0,13	1,13	1,33	1,48	1,45	1,55	1,72
		< -2	2,28	3,52	3,99	4,10	4,09	4,22	4,68
		< -1	15,87	7,74	8,59	9,02	8,91	9,25	9,66
		> 1	15,87	7,60	8,55	8,90	8,85	9,20	9,85
		> 2	2,28	3,25	3,92	4,11	4,11	4,31	4,66
		> 3	0,13	1,03	1,41	1,51	1,49	1,59	1,79
III	r_i^{*Gq}	< -3	0,13	0,02	0,08	0,10	0,10	0,12	0,20
		< -2	2,28	1,61	2,20	2,30	2,28	2,42	2,61
		< -1	15,87	14,88	15,59	15,78	15,76	15,91	16,44
		> 1	15,87	14,54	15,49	15,75	15,71	15,95	16,27
		> 2	2,28	1,54	2,25	2,33	2,30	2,43	2,68
		> 3	0,13	0,03	0,09	0,11	0,11	0,13	0,18
	r_i^{**q}	< -3	0,13	0,62	0,98	1,05	1,03	1,10	1,25
		< -2	2,28	3,16	3,97	4,12	4,07	4,22	4,55
		< -1	15,87	9,90	10,86	11,06	11,02	11,22	11,68
		> 1	15,87	9,88	10,81	11,05	11,00	11,30	11,69
		> 2	2,28	3,07	4,00	4,14	4,10	4,29	4,54
		> 3	0,13	0,60	0,99	1,08	1,05	1,14	1,28
IV	r_i^{*Gq}	< -3	0,13	0,04	0,08	0,09	0,09	0,11	0,16
		< -2	2,28	1,95	2,16	2,26	2,27	2,36	2,56
		< -1	15,87	14,87	15,63	15,79	15,78	15,98	16,34
		> 1	15,87	14,80	15,45	15,72	15,67	15,92	16,29
		> 2	2,28	1,90	2,21	2,32	2,32	2,42	2,57
		> 3	0,13	0,04	0,10	0,12	0,11	0,13	0,18
	r_i^{**q}	< -3	0,13	0,78	0,97	1,04	1,03	1,09	1,16
		< -2	2,28	3,68	3,93	4,10	4,06	4,19	4,43
		< -1	15,87	9,72	10,69	11,04	11,00	11,33	11,94
		> 1	15,87	9,61	10,58	11,04	10,95	11,34	11,91
		> 2	2,28	3,63	3,96	4,10	4,10	4,23	4,52
		> 3	0,13	0,85	1,01	1,08	1,08	1,14	1,28
V	r_i^{*Gq}	< -3	0,13	0,00	0,02	0,04	0,05	0,08	0,15
		< -2	2,28	1,52	1,88	2,10	2,06	2,26	2,46
		< -1	15,87	13,88	14,93	15,39	15,33	15,80	16,46
		> 1	15,87	14,07	15,84	16,08	16,02	16,36	17,02
		> 2	2,28	1,50	1,98	2,12	2,11	2,25	2,56
		> 3	0,13	0,00	0,02	0,04	0,05	0,07	0,13
	r_i^{**q}	< -3	0,13	1,31	2,20	2,69	2,72	3,07	5,28
		< -2	2,28	3,22	5,20	6,00	5,97	6,62	9,00
		< -1	15,87	8,48	11,16	11,98	11,97	12,97	14,52
		> 1	15,87	8,21	11,66	12,99	12,67	13,85	14,81
		> 2	2,28	3,13	5,14	6,04	6,19	7,21	9,00
		> 3	0,13	1,23	2,11	2,46	2,69	3,19	5,00

CAPÍTULO

4

SELEÇÃO DE MODELOS VIA RESÍDUO QUANTÍLICO

A seleção de modelos é um problema comum no âmbito da análise de regressão, uma vez que existem inúmeros modelos teóricos que podem ser ajustados, variando a distribuição da variável resposta, a quantidade de variáveis preditoras e a função de ligação, por exemplo. Na prática, em geral, não existe um modelo teórico que represente a relação entre uma variável resposta e um conjunto de variáveis preditoras. Portanto, o objetivo consiste em selecionar o melhor modelo teórico dentre um conjunto específico de modelos candidatos. De um modo geral, os métodos para comparação de modelos podem ser divididos entre abordagens baseadas na informação e processos tradicionais baseados em testes de hipóteses (LEWIS; BUTLER; GILBERT, 2011). Diante disso, esse capítulo introduz três critérios de seleção de modelos por meio de testes de bondade do ajuste com o uso do resíduo quantílico. Para avaliar o desempenho relativo destes métodos, foram realizadas simulações de Monte Carlo em um contexto específico de seleção da distribuição da variável resposta em modelos aditivos generalizados para localização, escala e forma (GAMLSS). Os resultados obtidos foram comparados com dois critérios de seleção baseados na verossimilhança. Com o intuito de enfatizar a importância da análise de diagnóstico do modelo selecionado, foram consideradas duas aplicações.

4.1 GAMLSS

Rigby e Stasinopoulos (2001) e Rigby e Stasinopoulos (2005) propuseram os GAMLSS como alternativa aos modelos lineares generalizados (MLG) (NELDER; WEDDERBURN, 1972) e modelos aditivos generalizados (GAM) (HASTIE; TIBSHIRANI, 1990), onde algumas premissas para o ajuste dos modelos são alargadas. Nos GAMLSS, a suposição de distribuição para variável resposta é relaxada para uma família de distribuição geral e, portanto, não se restringe às distribuições que pertencem à família exponencial. A parte sistemática do modelo

permite modelar os parâmetros de localização, forma e escala da variável resposta, por meio de funções paramétricas, não paramétricas, lineares, não lineares ou aditivas. No caso dos MLG e GAM, parâmetros relacionados à variância, assimetria e curtose não são modeladas explicitamente em função das preditoras lineares.

Os GAMLSS consideram observações independentes para a variável resposta, as variáveis explicativas e os valores dos efeitos aleatórios. Para os casos onde não há termos aditivos em qualquer um dos parâmetros da distribuição, o modelo é chamado de GAMLSS paramétrico, e é nele que este capítulo irá focar. Para este caso, sejam $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias independentes com função densidade de probabilidade $f(y_i; \underset{\sim}{g}^i)$, em que $\underset{\sim}{g}^{iT} = (q_{i1}, q_{i2}, q_{i3}, q_{i4}) = (m_i, s_i, n_i, t_i)$. Os parâmetros m_i e s_i são caracterizados como parâmetros de localização e escala, respectivamente, enquanto que n_i e t_i, se houverem, são ditos parâmetros de forma. Para $k = 1, 2, 3, 4$, seja $g_k(.)$ uma função de ligação monótona conhecida, tal que

$$g_k(\underset{\sim}{q}_k) = \underset{\sim}{h}_k = \mathbf{X}_k \underset{\sim}{b}_k \tag{4.1}$$

em que $\underset{\sim}{q}_k$ e $\underset{\sim}{h}_k$ são vetores $n \times 1$, sendo $\underset{\sim}{q}_k^T = (q_{1k}, q_{2k}, \ldots, q_{nk})$, $\underset{\sim}{b}_k^T = (b_{1k}, b_{2k}, \ldots, b_{p_k k})$ e $\mathbf{X}_k$ é a matriz das covariáveis de dimensão $n \times p_k$, e p_k corresponde ao número de variáveis regressoras no modelo para q_k. O componente paramétrico $\mathbf{X}_k \underset{\sim}{b}_k$ pode conter termos lineares, fatores, polinômios e termos de interação para as variáveis explicativas.

As estimativas dos parâmetros nos modelos GAMLSS podem ser obtidas por meio de dois algoritmos para maximização da função de verossimilhança, ou até mesmo pela combinação de ambos. São eles: algoritmo RS (RIGBY; STASINOPOULOS, 1996) e algoritmo CG (COLE; GREEN, 1992). Nos estudos de simulação e aplicação deste capítulo, será considerado o procedimento conjunto para o ajuste dos modelos, em que utiliza-se duas vezes o algoritmo RS antes de alternar para o algoritmo CS.

No contexto computacional, vários pacotes no R possuem a estrutura de modelagem GAMLSS implementada. Há apenas uma restrição para as funções densidade de probabilidade populacionais na implementação dos modelos GAMLSS em ambiente R (RIGBY; STASINOPOULOS, 2005). Faz-se necessário que a função $f(y_i; \underset{\sim}{g}^i)$ e as suas primeiras derivadas em relação a cada um dos parâmetros de $\underset{\sim}{g}^i$ sejam computáveis. Na biblioteca *gamlss* (RIGBY; STASINOPOULOS, 2005), estão implementadas inúmeras distribuições, dentre elas contínuas, discretas, para dados binários ou de contagem, inflacionadas em um ou dois pontos e com diferentes números de parâmetros (RIGBY *et al.*, 2019). Nesse capítulo, serão consideradas as distribuições contínuas de dois parâmetros $(\underset{\sim}{g}^i = (m_i, s_i))$, com suporte nos R^+ e cuja média seja o parâmetro m_i que define a função densidade de probabilidade. São elas: Gama, Gaussiana Inversa e Weibull (terceira parametrização) (JOHNSON; KOTZ; BALAKRISHNAN, 1995). A Tabela 4.1 apresenta a função densidade de probabilidade que define a distribuição da variável Y_i para os três casos considerados e sua respectiva variância.

Como $\mathrm{E}(Y_i) = m_i$ para as três distribuições e com base nas variâncias (Tabela 4.1), nota-se

Tabela 4.1 – Notação, função densidade de probabilidade (*f.d.p.*) e variância das distribuições Gama, Gaussiana Inversa e Weibull.

Distribuição	Notação	*f.d.p.*		Variância
Gama	$Gama(m_i, s_i)$	$\frac{y_i^{1/s_i^2-1} e^{-y_i/(s_i^2 m_i)}}{(s_i^2 m_i)^{1/s_i^2}\,\Gamma(1/s_i^2)}$	(4.2)	$s_i^2 m_i^2$
G. Inversa	$GI(m_i, s_i)$	$\frac{1}{\sqrt{2p\, s_i^2 y_i^3}} \exp\left[-\frac{1}{2m_i^2 s_i^2 y_i}(y_i - m_i)^2\right]$	(4.3)	$s_i^2 m_i^3$
Weibull	$Wei(m_i, s_i)$	$\frac{s_i y_i^{s_i-1}}{[m_i/a_i]^{s_i}} \exp\left[-\left(-\frac{y_i}{m_i/a_i}\right)^{s_i}\right]$	(4.4)	$\left(\frac{m_i}{a_i}\right)^2 (w_i - a_i^2)$
Nota		$a_i = \Gamma(s_i^{-1} + 1)$ $w_i = \Gamma(2s_i^{-1} + 1)$		

que, nas distribuições Gama e Weibull o coeficiente de variação depende apenas de s_i, sendo o próprio parâmetro s_i na Gama. Na Weibull a relação entre o parâmetro s_i e o coeficiente de variação não é tão simples, dependendo da função gama, mas é possível verificar que o coeficiente de variação descresce à medida que o parâmetro aumenta, ou seja, s_i é um parâmetro de precisão nesse caso. Na distribuição Gaussiana Inversa, o coeficiente de variação aumenta à medida que s_i cresce, mantido m_i fixo.

4.2 Seleção de modelos

Um modelo GAMLSS pode ser representado por $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, I\}$, em que $\mathcal{D}$ especifica a função distribuição da variável dependente, $\mathcal{G}$ define o conjunto das funções de ligação, $\mathcal{T}$ o conjunto dos termos preditores e I representa o conjunto dos hiperparâmetros. Para o ajuste dos GAMLSS a um conjunto de dados específico é necessário que seja especificada cada uma destas componentes, das quais existem distintas possibilidades de escolha, o que permite ajustar diferentes modelos concorrentes por meio de suas inúmeras combinações. Dito isso, é necessário que sejam estabelecidos métodos de comparação entre esses modelos candidatos, a fim de selecionar àquele que melhor se ajusta aos dados.

No contexto dos modelos GAMLSS, pode-se destacar o uso de dois critérios de seleção de modelos, sendo ambos baseados na informação: critério de informação de Akaike (*AIC*) (AKAIKE, 1973) e critério de informação Bayesiano (*BIC*) (SCHWARZ, 1978). Nesta seção serão apresentadas 3 critérios baseados no resíduo quantílico, visando a seleção de modelos que se diferem apenas pela distribuição de probabilidade assumida para a variável dependente. Isso pois, em particular, resultados da inferência estatística podem ser sensíveis à especificação incorreta da distribuição teórica atribuída à variável resposta.

4.2.1 Critérios de informação

Os dois critérios de seleção de modelos baseados na função de verossimilhança mais utilizados são o *AIC* e o *BIC* (KUHA, 2004). Para evitar a seleção de um modelo altamente parametrizado e que, portanto, implicam em problemas de aplicabilidade em novos conjuntos de dados, ambos os critérios são compostos de uma penalização sobre a complexidade do modelo. O primeiro critério foi derivado por Akaike (1973), sendo definido como

$$AIC = -2\log\left(\ell(\widehat{\boldsymbol{g}};y)\right) + 2p, \tag{4.5}$$

em que $\ell(\widehat{\boldsymbol{g}};y)$ é o logaritmo da função de verossimilhança do modelo ajustado e p é o número de parâmetros do modelo. O *AIC* é um estimador de máxima verossimilhança para divergência de Kullback-Leibler (KULLBACK; LEIBLER, 1951), que mede a distância entre um modelo candidato e o modelo verdadeiro.

O segundo critério baseado na informação também faz uso de uma função do número de parâmetros como penalização por complexidade, no intuito de aumentar a capacidade de generalização do modelo escolhido em relação à outros bancos de dados. Schwarz (1978) define o *BIC* como

$$BIC = -2\log\left(\ell(\widehat{\boldsymbol{g}};y)\right) + p\log(n), \tag{4.6}$$

em que em que $\ell(\widehat{\boldsymbol{g}};y)$ e p são como definidos em 4.5 e n é o tamanho amostral. O *BIC* visa selecionar o modelo com máxima probabilidade a posteriori dada as observações (HUANG, 2017).

Tanto o *AIC* quanto o *BIC* são funções que devem ser minimizadas, mas apesar de serem similares em sua definição, os critérios possuem diferentes propriedades sob diferentes suposições. Enquanto o *BIC* é assintoticamente consistente, o *AIC* é assintoticamente eficiente. Isto é, para os casos em que o modelo verdadeiro está entre os modelos candidatos, é esperado que o *BIC* selecione o modelo verdadeiro à medida que o tamanho amostral cresce (dentre outras suposições). Por outro lado, quando o modelo verdadeiro não está no conjunto dos modelos candidatos, é esperado que o *AIC* selecione assintoticamente o modelo que minimize o erro quadrático médio da estimativa (VRIEZE, 2012).

O *AIC* e *BIC* são critérios versáteis, que podem ser utilizados tanto em modelos aninhados quanto não aninhados. Kieschnick e McCullough (2003), por exemplo, utilizaram os critérios para seleção da distribuição da variável resposta do tipo proporção, e Peluso, Vinciotti e Yu (2019) no contexto dos modelos GAM, embora essa abordagem receba algumas críticas em virtude das log-verossimilhanças dos modelos candidatos terem estruturas diferentes. Há também uma certa discussão com relação a regra adotada para seleção de modelos via critérios baseados na informação. Colonna e Sauleau (2013) e Fraaije *et al.* (2015) consideram diferenças superiores a 10 como altas, entre 5 e 10 como moderadas e menores que 5 como incertas sobre qual modelo escolher. Contudo, as magnitudes dos valores de *AIC* e *BIC* variam bastante conforme o conjunto

de modelos candidatos e, portanto, as quantidades 10 e 5 podem representar proporcionalidades muito diferentes.

4.2.2 Testes de especificação para distribuição da variável resposta

Muitos trabalhos versam sobre o uso de critérios de seleção de modelos baseados em estatísticas de testes, porém se restringem ao problema de seleção de variáveis. Mallows (1973) propôs uma estatística para seleção de submodelos no contexto de regressão linear múltipla. Hosmer, Jovanovic e Lemeshow (1989) e Pregibon (1980) introduziram duas extensões ao critério de Mallows para o uso em MLG, sendo a primeira baseada na estatística c^2 de Pearson e a segunda na estatística *deviance*. Lawless e Singhal (1978) consideraram situações de regressão não Normal e o utilizaram a estatística do teste da razão de verossimilhança para testar submodelos em relação ao modelo completo.

Por outro lado, testes de qualidade do ajuste são propostos para avaliar as premissas sobre a distribuição da variável dependente, nesse contexto em que assume-se que as observações são provenientes de uma mesma família de distribuições, mas com parâmetros diferentes. Dentre as medidas utilizadas, pode-se citar o coeficiente de correlação de Filliben (1975) e as estatísticas de Anderson-Darling, Shapiro-Wilk, Cramér-Von Mises e suas variadas extensões (KLAR; MEINTANIS, 2012).

Filliben (1975) propôs uma estatística para testar hipóteses compostas de normalidade, combinando o gráfico de probabilidade Normal e o coeficiente de correlação produto-momento. O autor define o gráfico de probabilidade Normal como o gráfico da estatística de ordem $\mathcal{R}_{(i)}$ *versus* alguma medida de locação $loc(\mathcal{R}_{(i)})$. De maneira similar ao que foi proposto na Seção 2.2, Filliben (1975) não encoraja o uso da média como medida de locação, mas sim a mediana. Isso pois a $E(\mathcal{R}_{(i)})$ possui três propriedades indesejadas, dentre elas a não existência de uma técnica uniforme de obtenção entre as distribuições e, em alguns casos como para a distribuição Cauchy, $E(\mathcal{R}_{(i)})$ não pode ser definida. Com isso, é esperado que o gráfico de $\mathcal{R}_{(i)}$ *versus* $M_i = loc(\mathcal{R}_{(i)}) =$ mediana $(\mathcal{R}_{(i)})$ seja aproximadamente linear, para os casos em que de fato a amostra foi gerada de uma distribuição Normal. Para medir a linearidade do gráfico, o autor propõe o uso do coeficiente de correlação produto-momento. Considerando as observações ordenadas $\mathcal{R}_{(}i)$ e as medianas da estatística de ordem M_i de uma distribuição $N(0,1)$, o coeficiente de correlação do gráfico de probabilidade Normal é definido por

$$r^F = Corr(\mathcal{R}, M) = \frac{\sum_{i=1}^{n}(\mathcal{R}_{(i)} - \bar{\mathcal{R}})(M_i - \bar{M})}{\sqrt{\sum_{i=1}^{n}(\mathcal{R}_{(i)} - \bar{\mathcal{R}})^2 \sum_{i=1}^{n}(M_i - \bar{M})^2}}. \tag{4.7}$$

Filliben (1975) ressalta a possibilidade de extensão do estatística 4.7 como critério de seleção da melhor distribuição dentre um conjunto candidato finito. Por isso, nesse capítulo, propõe-se o uso do coeficiente de correlação de Filliben como critério de seleção de modelos que se diferem pela distribuição especificada para variável resposta. Nesse caso, o modelo

selecionado será àquele com maior valor r^F, dentre o conjunto dos modelos candidatos. Essa abordagem é conveniente no contexto dos modelos de regressão, uma vez que já é frequente o uso do gráfico de probabilidade Normal para verificação das suposições do modelo (mas não como forma comparativa entre modelos concorrentes). Nesse caso, o gráfico é construído para os resíduos gerados pelo modelo ajustado e, para os modelos GAMLSS em particular, é comum o uso do resíduo quantílico.

Diante disso e devido à propriedade do resíduo quantílico ter distribuição assintótica Normal padrão, quando os parâmetros do modelo são consistentemente estimados, propõe-se também o uso da estatística do teste de Anderson-Darling (ANDERSON; DARLING, 1954) como critério de seleção de modelos, dada por

$$ADS = -n - \frac{1}{n}\sum_{i=1}^{n}[2i-1][\log(p_{(i)}) + \log(1-p_{(n-i+1)})], \tag{4.8}$$

em que $p_{(i)} = \Phi([\mathcal{R}_{(i)} - \overline{\mathcal{R}}]/s)$ e Φ é a função de distribuição acumulada da $N(0,1)$, $\overline{\mathcal{R}}$ e s são a média e o desvio padrão dos resíduos ajustados pelo modelo. Nesse caso, quanto menor a estatística do teste, entende-se que mais próxima está a distribuição do resíduo à $N(0,1)$ e, portanto, seleciona-se o modelo com menor ADS dentre o conjunto dos modelos candidatos.

De maneira análoga, o terceiro critério de seleção proposto nesse capítulo baseia-se no resíduo quantílico padronizado, introduzido por Klar e Meintanis (2012) no intuito de derivar um teste de bondade de ajuste para MLG. Nesse caso, será utilizada a estatística do teste de Anderson-Darling, definida em (4.8), em que $\mathcal{R}$ será dado por (KLAR; MEINTANIS, 2012)

$$Z_i = \frac{r_{qi} - \bar{r}_q}{s_{r_q}}, \tag{4.9}$$

sendo r_{qi} o resíduo quantílico definido em (2.41). Para diferenciar este critério de seleção ao anterior, denomina-se a estatística do teste de Anderson-Daring baseada em Z_i como $ADS^{\{2\}}$.

4.3 Estudos de simulação

Para avaliar o desempenho dos critérios de seleção apresentados na Seção 4.2, foram realizados estudos de simulação Monte Carlo. Em cada réplica, foram geradas as observações da variável resposta com base em um modelo GAMLSS com uma das três distribuições apresentadas na Tabela 4.1. Considerando todas as demais componentes do modelo especificadas corretamente, isto é, sem variar a função de ligação e o conjunto de preditoras utilizadas na construção da amostra, ajustou-se os 3 modelos GAMLSS possíveis. Ou seja, para cada réplica foram ajustados dois modelos com má especificação na distribuição da variável resposta e o modelo correto. Para cada conjunto destes 3 ajustes, foram calculados os 5 critérios de seleção, obtendo suas respectivas taxas de acerto ao longo das 10000 réplicas de Monte Carlo.

O procedimento foi dividido de acordo com a distribuição da variável resposta, onde foram considerados diferentes cenários e tamanhos amostrais ($n = 25, 50, 100$ e 500). Para as 3

distribuições, o cenário I foi construído segundo o modelo

$$\ln(m_i) = b_{01} + b_{11}x_{i1} + b_{21}x_{i2} + b_{31}x_{i3} + b_{41}x_{i4}, \tag{4.10}$$

em que $b_{01} = 1$, $x_{ik} \sim N(10,2)$ e $b_{k1} = 0,05, \forall k = 1,2,3,4$. Neste cenário, $s_i^{\mathcal{D}}$ foi considerado fixo para todas as observações e $\mathcal{D} = GA, GI$ ou WB, sendo $s_i^{GA} = 0,3$ para o modelo com resposta Gama, $s_i^{GI} = 0,11$ para o modelo com resposta Gaussiana Inversa e $s_i^{WB} = 3,7$ para o modelo com distribuição Weibull, de modo que o coeficiente de variação seja igual a $0,3$ para todos os três casos. Outros 6 cenários são derivados à partir deste. Os cenários II e III consistem em aumentar o coeficiente de variação para $0,6$ e $0,9$, respectivamente, considerando $s_i^{GA} = 0,6$, $s_i^{GI} = 0,22$ e $s_i^{WB} = 1,7$ e $s_i^{GA} = 0,9$, $s_i^{GI} = 0,33$ e $s_i^{WB} = 1,1$, na devida ordem. No cenário IV, $x_{ik} \sim Gama(10,2), \forall k = 1,2,3,4$, para as três distribuições da variável resposta. No cenário V, acrescentou-se 4 covariáveis ao modelo 4.10, sendo $x_{ik} \sim N(10,2)$ e $b_{k1} = 0,05, \forall k = 5,6,7,8$. No cenário VI, as covariáveis foram geradas de maneira correlacionada, por meio da distribuição $N_4(\underset{\sim}{10}, \mathrm{S})$, sendo

$$\underset{\sim}{10} = \begin{bmatrix} 10 \\ 10 \\ 10 \\ 10 \end{bmatrix} \quad \text{e} \quad \mathrm{S} = \begin{bmatrix} 4 & 0.8 & 0.1 & 0 \\ 0.8 & 4 & 0 & 0.1 \\ 0.1 & 0 & 4 & 0.7 \\ 0 & 0.1 & 0.7 & 4 \end{bmatrix}. \tag{4.11}$$

Por fim, o cenário VII considera $s_i^{\mathcal{D}}$ variando conforme as 4 covariáveis, de tal forma que

$$\ln\left(s_i^{\mathcal{D}}\right) = b_{02} + b_{12}x_{i1} + b_{22}x_{i2} + b_{32}x_{i3} + b_{42}x_{i4}, \tag{4.12}$$

em que $x_{ik} \sim N(10,2)$ e $b_{k2} = (-1)^{k+1}0,03, \forall k = 1,2,3,4$ e $b_{02} = \log(t)$, sendo t o valor de $s_i^{\mathcal{D}}$ no cenário I para a respectiva distribuição atribuída a variável resposta. Com isso, o valor de $s_i^{\mathcal{D}}$ no cenário VII é similar ao valor de $s_i^{\mathcal{D}}$ no cenário base, para cada uma das três distribuições se o valor das covariáveis forem todos iguais aos seus valores médios.

A Tabela 4.2 apresenta o percentual de réplicas em que cada critério seleciona o modelo corretamente especificado, isto é, cuja distribuição considerada para variável resposta de fato coincide com a distribuição geradora das observações. Os resultados são divididos por distribuição e tamanho amostral, destacando o critério com melhor desempenho em cada cenário. De maneira geral, os critérios baseados na verossimilhança obtiveram maior taxa de acerto que os critérios baseados em testes de hipótese com o resíduo quantílico. Todos os critérios têm desempenho melhor à medida que o tamanho amostral aumenta, alcançando quase 100% de acerto com $n = 500$ em alguns cenários (alguns valores são iguais a $1,00$ devido ao sistema de arrendondamento com apenas duas casas decimais).

À medida que a variabilidade da resposta aumenta (cenários I-III), a performance dos métodos diminui, no caso da distribuição Weibull e aumenta na distribuição Gaussiana Inversa. Na distribuição Gama, o desempenho dos métodos baseados na informação também cai, e dos métodos baseados no resíduo quantílico oscilam entre os tamanhos amostrais. Os cenários

com mudança de distribuição das covariáveis (cenário IV) ou com covariáveis correlacionadas (cenário VI), em geral, não possuem resultados muito diferentes ao cenário base. O acréscimo de covariáveis no modelo (cenário V) diminui o desempenho dos critérios *AIC* e *BIC* na distribuição Gama e aumenta nas demais, oscilando também conforme o tamanho amostral para os critérios propostos. No último cenário, com $s_i^{\mathcal{D}}$ também variando conforme as covariáveis, o desempenho dos critérios piora nas distribuições Gama e Weibull e, em geral, melhora na distribuição Gaussiana Inversa, embora hajam pequenas flutuações entre os tamanhos amostrais considerados.

Para as simulações cuja variável resposta possui distribuição Gama, a estatística *ADS* obteve melhores resultados que os critérios *AIC* e *BIC* em $n = 25$. Contudo, o mesmo não ocorre nos modelos com as demais distribuições. Já as estatísticas $ADS^{\{2\}}$ e r^F apresentaram, no máximo, desempenho semelhante aos demais critérios com o resíduo quantílico, sendo, em geral, piores que os demais métodos comparados.

É válido ressaltar que os resultados dos critérios *AIC* e *BIC* são idênticos, uma vez que os modelos comparados possuem a mesma quantidade de parâmetros. Em todos os cenários com resposta Gaussiana Inversa ou Weibull, e $n = 25, 50$ ou 100, esses critérios são superiores aos métodos propostos, tendo performance igual ou maior em $n = 500$. Apesar das críticas aos critérios baseados na informação, conforme discutido na Seção 4.2, os resultados indicam um bom desempenho para escolha da distribuição atribuída a variável resposta, sobretudo em comparação aos critérios propostos neste trabalho, que possuem melhor justificativa teórica para o uso.

4.4 Aplicação

Embora os resultados da seção anterior mostrem que os critérios *AIC* e *BIC* possuem bom desempenho para seleção de modelos, é válido ressaltar a importância da análise de diagnóstico envolvendo os resíduos gerados pelo modelo escolhido. Uma vez que os critérios selecionarão o melhor modelo dentre àqueles ajustados, não há garantias se as premissas do ajuste são de fato atendidas. Para ilustrar essa situação, foram consideradas duas aplicações, sendo a primeira via dados simulados, onde o verdadeiro processo de geração dos dados é conhecido, e a segunda com um banco de dados reais.

4.4.1 Aplicação com dados simulados

Para esta primeira aplicação, foram geradas 500 observações provenientes do modelo de regressão com reposta Gaussiana Inversa Generalizada (GIG), definida pela função densidade de

Tabela 4.2 – Taxa de acerto dos critérios de seleção, por tamanho amostral e distribuição atribuída à variável resposta.

Cen.	Critério	Gama				Gaussiana Inversa				Weibull			
		n				*n*				*n*			
		25	50	100	500	25	50	100	500	25	50	100	500
I	*AIC*	0,26	**0,47**	**0,66**	**0,98**	**0,61**	**0,73**	**0,86**	**1,00**	**0,77**	**0,85**	**0,93**	**1,00**
	BIC	0,26	**0,47**	**0,66**	**0,98**	**0,61**	**0,73**	**0,86**	**1,00**	**0,77**	**0,85**	**0,93**	**1,00**
	ADS	**0,30**	0,39	0,58	0,93	0,45	0,60	0,72	0,96	0,61	0,80	0,90	**1,00**
	$ADS^{\{2\}}$	0,28	0,39	0,57	0,92	0,46	0,61	0,73	0,96	0,65	0,80	0,89	**1,00**
	r^F	0,28	0,41	0,63	0,95	0,47	0,62	0,76	0,98	0,69	0,83	0,92	**1,00**
II	*AIC*	0,22	0,43	0,62	0,95	**0,74**	**0,87**	**0,96**	**1,00**	**0,73**	**0,76**	**0,81**	**0,96**
	BIC	0,22	0,43	0,62	0,95	**0,74**	**0,87**	**0,96**	**1,00**	**0,73**	**0,76**	**0,81**	**0,96**
	ADS	**0,35**	**0,45**	0,61	0,93	0,50	0,72	0,88	**1,00**	0,56	0,70	0,77	0,94
	$ADS^{\{2\}}$	0,31	0,44	0,61	0,92	0,53	0,75	0,91	**1,00**	0,60	0,71	0,76	0,93
	r^F	0,30	**0,45**	**0,64**	**0,95**	0,53	0,76	0,93	**1,00**	0,64	0,73	0,79	**0,96**
III	*AIC*	0,20	0,36	0,45	0,62	**0,83**	**0,94**	**0,99**	**1,00**	**0,61**	**0,64**	**0,64**	**0,68**
	BIC	0,20	0,36	0,45	0,62	**0,83**	**0,94**	**0,99**	**1,00**	**0,61**	**0,64**	**0,64**	**0,68**
	ADS	**0,42**	**0,43**	**0,47**	0,61	0,50	0,77	0,94	**1,00**	0,49	0,57	0,60	0,65
	$ADS^{\{2\}}$	0,33	0,40	**0,47**	0,62	0,56	0,83	0,97	**1,00**	0,56	0,61	0,60	0,65
	r^F	0,31	0,39	**0,47**	**0,64**	0,56	0,84	0,98	**1,00**	0,59	0,62	0,63	0,66
IV	*AIC*	0,27	**0,45**	**0,69**	**0,98**	**0,62**	**0,73**	**0,85**	**1,00**	**0,76**	**0,85**	**0,93**	**1,00**
	BIC	0,27	**0,45**	**0,69**	**0,98**	**0,62**	**0,73**	**0,85**	**1,00**	**0,76**	**0,85**	**0,93**	**1,00**
	ADS	**0,29**	0,40	0,57	0,92	0,44	0,60	0,73	0,96	0,62	0,80	0,90	**1,00**
	$ADS^{\{2\}}$	0,27	0,40	0,57	0,91	0,46	0,61	0,74	0,96	0,66	0,80	0,90	**1,00**
	r^F	0,26	0,42	0,63	0,94	0,47	0,62	0,76	0,98	0,69	0,83	0,92	**1,00**
V	*AIC*	0,15	**0,35**	**0,66**	**0,99**	**0,82**	**0,90**	**0,98**	**1,00**	**0,80**	**0,87**	**0,94**	**1,00**
	BIC	0,15	**0,35**	**0,66**	**0,99**	**0,82**	**0,90**	**0,98**	**1,00**	**0,80**	**0,87**	**0,94**	**1,00**
	ADS	**0,37**	**0,35**	0,55	0,94	0,23	0,57	0,86	**1,00**	0,38	0,74	0,90	**1,00**
	$ADS^{\{2\}}$	0,33	0,33	0,54	0,93	0,26	0,61	0,90	**1,00**	0,45	0,77	0,90	**1,00**
	r^F	0,33	0,34	0,59	0,96	0,24	0,61	0,92	**1,00**	0,50	0,81	0,93	**1,00**
VI	*AIC*	0,28	**0,45**	**0,68**	**0,98**	**0,61**	**0,74**	**0,86**	**1,00**	**0,77**	**0,85**	**0,93**	**1,00**
	BIC	0,28	**0,45**	**0,68**	**0,98**	**0,61**	**0,74**	**0,86**	**1,00**	**0,77**	**0,85**	**0,93**	**1,00**
	ADS	**0,30**	0,40	0,58	0,93	0,44	0,61	0,72	0,96	0,61	0,80	0,90	**1,00**
	$ADS^{\{2\}}$	0,28	0,40	0,57	0,92	0,46	0,61	0,73	0,96	0,65	0,80	0,89	**1,00**
	r^F	0,28	0,42	0,63	0,95	0,46	0,62	0,76	0,98	0,68	0,83	0,92	**1,00**
VII	*AIC*	0,05	0,25	0,53	**0,92**	**0,68**	**0,79**	**0,86**	**0,99**	**0,62**	**0,78**	**0,89**	**1,00**
	BIC	0,05	0,25	0,53	**0,92**	**0,68**	**0,79**	**0,86**	**0,99**	**0,62**	**0,78**	**0,89**	**1,00**
	ADS	**0,29**	0,31	0,52	0,89	0,36	0,65	0,77	0,97	0,48	0,72	0,85	**1,00**
	$ADS^{\{2\}}$	0,27	0,32	0,53	0,89	0,40	0,66	0,78	0,97	0,49	0,72	0,85	**1,00**
	r^F	0,27	**0,34**	**0,57**	**0,92**	0,40	0,68	0,80	**0,99**	0,50	0,74	0,88	**1,00**

probabilidade

$$f_{GIG}(y;m,s,n) = \left(\frac{c}{m}\right)^n \left[\frac{y^{n-1}}{2K_n\left(\frac{1}{s^2}\right)}\right] \exp\left[-\frac{1}{2s^2}\left(\frac{cy}{m}+\frac{m}{cy}\right)\right], \tag{4.13}$$

em que $c = \left[K_{n+1}(1/s^2)\right]\left[K_n(1/s^2)\right]^{-1}$, e $K_n(t)$ representa a função Bessel modificada de terceiro tipo e ordem n. Um caso especial inclui a distribuição hipérbole ($n = 0$), que produz uma analogia próxima a distribuição von Mises no círculo (JORGENSEN, 2012; BARNDORFF-NIELSEN, 1978; RUKHIN, 1978). As covariáveis foram geradas com base em uma distribuição $N(10,2)$ e usou-se $b_0 = 1, b_1 = b_2 = 0.05, s = 0.35, n = -9$, cujo coeficiente de variação resultante é idêntico ao considerado nos cenários bases do estudo de simulação $(CV = 0,3)$.

Suponha que um analista deseja estudar a relação entre a variável resposta Y e as duas covariáveis por meio de um modelo de regressão. Devido às características das observações geradas para variável resposta, isto é, $y \in \mathrm{R}^+$, é razoável supor que o pesquisador considere ajustar os modelos com resposta Gama, Gaussiana Inversa ou Weibull. Nesse caso, independente do critério de seleção escolhido, ao observar os valores obtidos (Tabela 4.3) o analista concluiria que o modelo com resposta Gaussiana Inversa se ajusta melhor aos dados.

Tabela 4.3 – Resultado dos critérios de seleção para os modelos ajustados com resposta Gama, Gaussiana Inversa e Weibull.

Modelos	Critérios				
	AIC	*BIC*	*ADS*	$ADS^{\{2\}}$	r^F
Gama	2220,66	2237,52	2,50	2,46	0,98
Gaussiana Inversa	**2198,45**	**2215,30**	**0,93**	**0,89**	**0,99**
Weibull	2338,27	2355,13	11,68	9,11	0,95

Ao realizar a análise de diagnóstico para o modelo selecionado, percebe-se pelo gráfico de probabilidade Normal com envelope simulado (Figura 4.1(a)) que há falhas de especificação. Diante disso, o pesquisador poderia testar outras abordagens, como variar a função de ligação para média, adicionar covariáveis no submodelo da variância ou ir em busca de novas distribuições para variável resposta. Nessa última alternativa, seria possível chegar ao modelo correto, onde não seriam apontadas falhas nas suposições do modelo ajustado (Figura 4.1(b)). Os critérios de seleção também indicariam que o modelo com resposta GIG se ajusta melhor aos dados, quando comparado com o modelo com resposta Gaussiana Inversa (Tabela 4.4). Isto é, se o modelo correto fosse considerado pelo analista desde o início do estudo, os critérios de seleção seriam suficientes para o êxito da análise. Caso contrário e se negligenciada a análise de diagnóstico, o pesquisador iria selecionar um modelo mal especificado, podendo comprometer o desempenho inferencial do seu estudo.

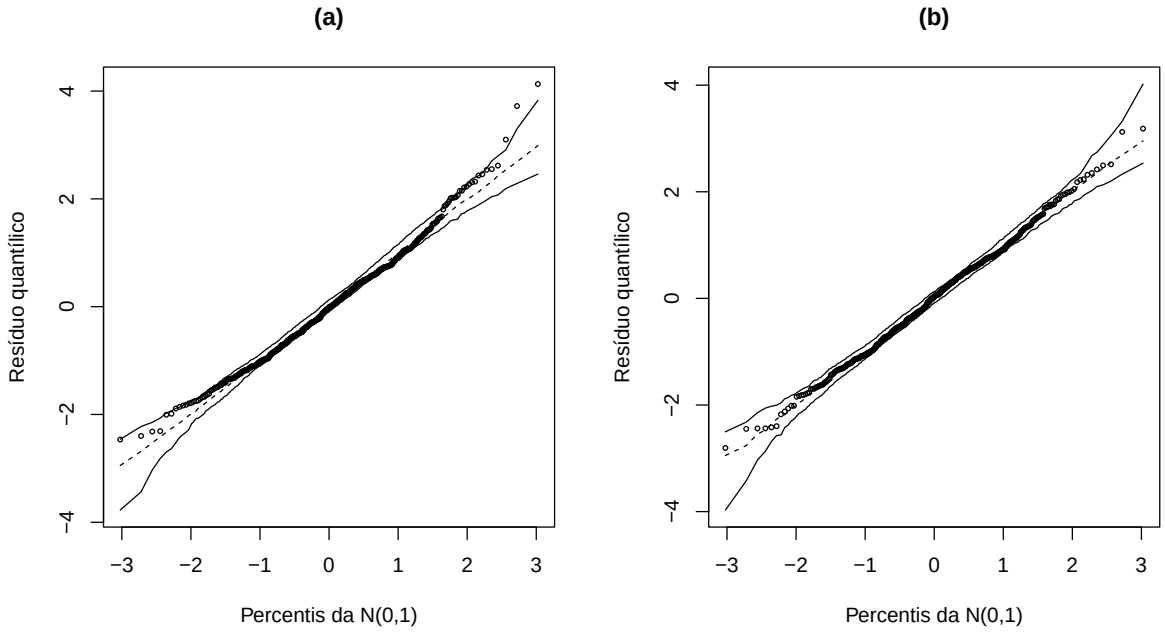

Figura 4.1 – Gráfico de probabilidade Normal com envelope simulado para os modelos com resposta Gaussiana Inversa (a) e com resposta GIG (b).

Tabela 4.4 – Resultado dos critérios de seleção para os modelos ajustados com resposta Gaussiana Inversa e GIG.

Modelos	Critérios				
	AIC	BIC	ADS	$ADS^{\{2\}}$	r^F
Gaussiana Inversa	2198,45	2215,30	0,93	0,89	0,99
GIG	**2183,60**	**2204,67**	**0,38**	**0,37**	**1,00**

Sabe-se que, na prática, não há necessariamente um modelo gerador dos dados. Busca-se, portanto, àquele que melhor consegue descrever a relação entre as variáveis e satisfazer os objetivos do estudo. Contudo, devido a incontável quantidade de modelos de regressão existentes, uma vez que pode-se também variar a função de ligação, quantidade de covariáveis, submodelos para os demais parâmetros da distribuição resposta, adição de termos polinomiais, etc, os critérios de seleção apenas indicam o melhor modelo dentre o conjunto de modelos testados (com alguma taxa de acerto). Diante disso, a análise de diagnóstico é uma etapa fundamental para verificação das suposições feitas para o modelo, cujos resíduos desempenham um papel importante para verificar a adequação do modelo.

4.4.2 Aplicação com dados reais

A segunda aplicação usa dados de Hodges (1998) , correspondentes à despesa média de 45 estados ou jurisdições por internação hospitalar (American Medical Association, Pesquisa Anual de Hospitais, 1991). Os dados estão disponíveis no pacote *gamlss.data* (STASINOPOULOS;

RIGBY; De Bastiani, 2021) do software R. Como covariáveis, tem-se o tamanho populacional (censo de 1990) e a região de cada estado.

Se Y_i é a despesa média por internação hospitalar no estado i, com $i = 1, \ldots, 45$, é razoável considerar o ajuste dos modelos com resposta Gama, Gaussiana Inversa ou Weibull, uma vez que $y_i \in \mathrm{R}^+$. Suponha que um pesquisador decida modelar os dados por meio de

$$\begin{cases} \ln(m_i) = b_{01} + b_{11}x_{i1} + b_{21}\mathrm{I}_{MT}(x_{i2}) + b_{22}\mathrm{I}_{NC}(x_{i2}) + b_{23}\mathrm{I}_{NE}(x_{i2}) + b_{24}\mathrm{I}_{PA}(x_{i2}) \\ \qquad\quad + b_{25}\mathrm{I}_{SA}(x_{i2}) + b_{26}\mathrm{I}_{SC}(x_{i2}), \\ \ln(s_i) = b_{02}, \end{cases} \tag{4.14}$$

em que x_{i1} representa o tamanho populacional do *i-ésimo* estado no censo de 1990 dividido por 10000, x_{i2} é uma variável categórica para as 7 regiões que abrangem os 45 estados (MA, MT, NC, NE, PA, SA, SC) e, portanto, a estrutura do modelo ajustado considera cada nível de x_{i2} por meio de uma variável indicadora I, tendo a região MA como referência no modelo. O modelo para o parâmetro s_i foi considerado fixo e e variou-se a distribuição especificada para variável dependente conforme a Tabela 4.1. Nesse caso, independente do critério de seleção escolhido, ao observar os valores obtidos (Tabela 4.5) o analista concluirá que o modelo com resposta Weibull se ajusta melhor aos dados.

Tabela 4.5 – Resultado dos critérios de seleção para os modelos ajustados com resposta Gama, Gaussiana Inversa e Weibull, sem covariáveis para o parâmetro s.

Modelos	Critérios				
	AIC	BIC	ADS	$ADS^{\{2\}}$	r^F
Gama	768,90	785,16	3,46	3,51	0,86
Gaussiana Inversa	779,08	795,34	4,44	4,42	0,82
Weibull	**756,89**	**773,15**	**0,95**	**1,04**	**0,95**

Ao realizar a análise de diagnóstico para o modelo selecionado, percebe-se pelo gráfico de probabilidade Normal com envelope simulado (Figura 4.2(a)) que há falhas de especificação. Diante disso, o pesquisador poderá testar outras abordagens, como adicionar covariáveis no submodelo de s_i. Nesse caso, os critérios de seleção indicarão que o novo modelo se ajusta melhor aos dados, quando comparado ao modelo selecionado anteriormente (Tabela 4.6) e o gráfico de probabilidade Normal para o resíduo quantílico (Figura 4.2(b)) não indicará falhas nas premissas do ajuste.

Tabela 4.6 – Resultado dos critérios de seleção para os modelos ajustados com resposta Weibull, considerando ou não covariáveis ao submodelo para s.

Modelos	Critérios				
	AIC	BIC	ADS	$ADS^{\{2\}}$	r^F
Weibull	756,89	773,15	0,95	1,04	0,95
Weibull2	**710,08**	**738,98**	**0,52**	**0,13**	**1,00**

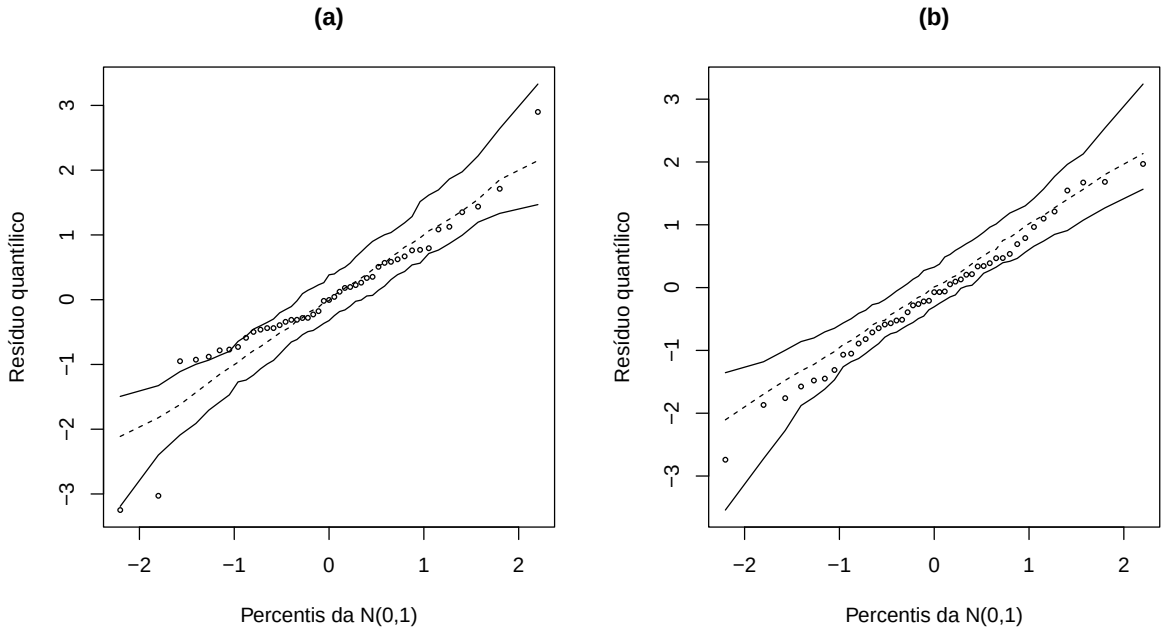

Figura 4.2 – Gráfico de probabilidade Normal com envelope simulado para os modelos com resposta Weibull, sem covariáveis para o submodelo da variância (a) e adicionando as covariáveis ao submodelo da variância (b).

A Tabela 4.7 apresenta os resultados do modelo Weibull2, cujo s_i também varia conforme as covariáveis. Observe que tanto população como região são significantes no submodelo da média tal como no do parâmetro de precisão. Estima-se que, em média, a despesa média por internação aumenta 9,12% a cada aumento de 10000 indivíduos na população de uma mesma região. A mudança da região MA (usada como referência nos modelos ajustados) para qualquer outra, implica em uma diminuição da média e da variância da variável resposta.

Por fim, os gráficos residuais (Figura 4.3) não indicam falhas nas demais suposições feitas para o ajuste, tampouco a presença de pontos influentes ou *outliers*. Os pontos encontram-se aleatoriamente dispersos em torno de zero no gráfico dos valores ajustados *versus* os resíduos (Figura 4.3(a)), o que traz indícios sobre a homoscedasticidade da variância dos resíduos. Já o gráfico dos resíduos *versus* a ordem das observações (Figura 4.3(b)), não indica correlação ou dependência residual.

4.5 Conclusões

Neste capítulo, comparou-se diferentes procedimentos de seleção de modelos no contexto de GAMLSS, dos quais três foram introduzidos neste trabalho e baseiam-se no resíduo quantílico. Embora haja uma boa justificativa teórica para esses critérios, os estudos de simulação mostraram que o desempenho dos procedimentos usuais baseados na informação são, em geral, superiores, sobretudo nos casos de amostras com $n = 50$ ou 100.

Tabela 4.7 – Modelo Weibull final (Weibull2) para os dados de despesa média por internação hospitalar.

Equação	Variável	Estimativa	Erro padrão	Valor-p	Exp(Estimativa)
m	Intercepto	8,6692	0,0184	$< 0,0001$	5820,9277
	População	0,0872	0,0101	$< 0,0001$	1,0912
	Região:MT	-0,1459	0,0282	$< 0,0001$	0,8643
	Região:NC	-0,1483	0,0110	$< 0,0001$	0,8621
	Região:NE	-0,0104	0,0227	0,6503	0,9897
	Região:PA	-0,2015	0,0123	$< 0,0001$	0,8175
	Região:AS	-0,1134	0,0068	$< 0,0001$	0,8928
	Região:SC	-0,2651	0,0032	$< 0,0001$	0,7671
s	Intercepto	-3,0060	0,4573	$< 0,0001$	0,0495
	População	4,8934	0,1940	$< 0,0001$	133,4121
	Região:MT	4,5789	0,5092	$< 0,0001$	97,4049
	Região:NC	1,9807	0,4554	0,0001	7,2476
	Região:NE	3,9946	0,5226	$< 0,0001$	54,3060
	Região:PA	4,0019	0,4878	$< 0,0001$	54,7001
	Região:AS	2,4049	0,4666	$< 0,0001$	11,0776
	Região:SC	3,9123	0,4782	$< 0,0001$	50,0126

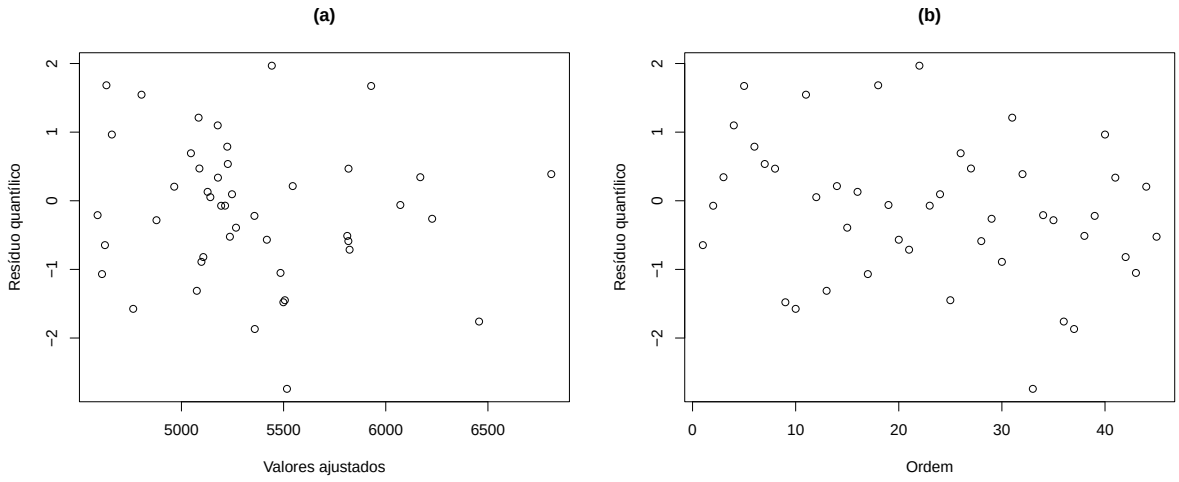

Figura 4.3 – Gráficos residuais para o modelo selecionado (Weibull2).

Contudo, é válido ressaltar que os critérios de seleção buscam o melhor modelo dentre um conjunto de modelos candidatos. Conforme discutido brevemente na Seção 4.2, os critérios *AIC* e *BIC* apresentam propriedades diferentes e que dependem se o modelo verdadeiro pertence ou não ao conjunto dos modelos candidatos (embora, na prática, nem sempre exista um modelo verdadeiro). Dito isso, foram apresentadas duas aplicações que reforçam a importância da análise de diagnóstico do modelo selecionado. No contexto dos modelos GAMLSS, o resíduo quantílico é, em geral, a ferramenta escolhida para a análise dos pressupostos. Tanto na aplicação com dados simulados, onde há a garantia de que o modelo verdadeiro não estava entre os modelos candidatos, quanto na aplicação com dados reais, os gráficos residuais indicam claramente falhas na especificação do modelo selecionado, o que poderia levar à problemas inferenciais. Embora

os critérios de seleção possuam um bom desempenho no que se propõem, eles apenas indicam o melhor modelo dentre àqueles testados pelo pesquisador. Se todos os modelos candidatos possuírem alguma falha nas premissas e a análise de diagnóstico for negligenciada, o modelo escolhido será apenas o melhor, dentre um conjunto de modelos mal especificados.

CAPÍTULO

5

CONCLUSÕES

O objetivo desse trabalho foi desenvolver extensões do resíduo quantílico, em aspectos de análise diagnóstica e seleção de modelos. Em relação à verificação das premissas para o ajuste do modelo, foram propostas extensões do resíduo quantílico no contexto dos modelos de regressão circular-linear e beta inflacionados em dois ou três pontos. Já para a seleção de modelos, estudou-se o desempenho de três critérios de seleção baseados no resíduo quantílico, para seleção da distribuição assumida para variável resposta em modelos GAMLSS.

Foi demonstrado que o resíduo quantílico circular é assintoticamente Normal padrão distribuído, assim como o resíduo quantílico no caso linear. Para amostras pequenas, há uma boa concordância entre a distribuição do resíduo quantílico circular e a distribuição Normal padrão em diferentes cenários avaliados. O resíduo também possui bom desempenho para detectar falhas no ajuste, sobretudo em relação aos demais resíduos comparados. Além dessas características, o resíduo quantílico circular é interessante pela sua simplicidade e invariabilidade da expressão algébrica ao considerar distintas classes de modelos de regressão para dados circulares.

Para identificação de *outliers* em modelos de regressão beta inflacionados em dois e três pontos, foi demonstrado que a extensão proposta possui distribuição caudal semelhante à distribuição Normal padrão, quando utilizado o resíduo quantílico em seu cálculo (r_i^{*Gq}) . Simulações de Monte Carlo sugerem que essa classe de resíduos é capaz de detectar possíveis *outliers* em situações que o resíduo quantílico aleatorizado falha.

No âmbito da seleção de modelos, os critérios introduzidos com base no resíduo quantílico não alcançaram os resultados esperados, considerando o contexto dos modelos GAMLSS analisado. Embora haja uma boa justificativa teórica para os métodos propostos, os resultados dos critérios de seleção usuais (baseados na informação) foram superiores. Contudo, as aplicações reforçam a importância da análise de diagnóstico do modelo selecionado.

As extensões do resíduo quantílico propostas nesse trabalho perpassam diferentes classes de modelos de regressão, o que indica uma numerosa quantidade de situações práticas que podem

ser beneficiadas. No contexto de dados circulares, pode-se mencionar estudos meteorológicos sob a direção do vento em relação ao nível de poluente, pesquisas sob a movimentação de animais em períodos migratórios e estudos sob o tempo de chegada de pacientes em um hospital. Já no âmbito dos modelos de regressão beta inflacionados, a extensão proposta pode ser útil para instituições financeiras ou administração federal, em análises sobre inadimplência e seguro desemprego, por exemplo. As propriedades demonstradas algebricamente e o desempenho satisfatório dos resíduos propostos em verificar as suposições do modelo ou identificar possíveis *outliers*, também indicam a versatilidade do resíduo quantílico em análises de diagnóstico. Assim, espera-se que essa tese possa motivar trabalhos aplicados na área de dados circulares ou econometria, e também trabalhos na área de Estatística, como os citados a seguir.

5.1 Trabalhos futuros

Dentre os estudos relacionados a este trabalho que podem ser desenvolvidos, destaca-se:

• Estender o resíduo quantílico circular para uso em modelos de regressão para dados cilíndricos (ABE; LEY, 2017; CREMERS; PENNINGS; LEY, 2020). Este trabalho futuro pode apresentar um modelo de regressão que utiliza funções de ligação e uma distribuição cilíndrica e avaliar o desempenho do resíduo proposto neste modelo.

• Estender o resíduo quantílico aos casos em que duas variáveis respostas são observadas simultaneamente no mesmo indivíduo. Em tal caso, com o intuito de considerar a relação intrínseca entre estas duas variáveis, surgem os modelos de regressão bivariados (OLIVEIRA; DINIZ; DURBÁN, 2018), nos quais assume-se uma estrutura de dependência entre as respostas. A extensão direta do resíduo quantílico tem limitações, portanto é interessante que se proponha uma extensão do cuja distribuição mantenha-se assintoticamente Normal padrão.

REFERÊNCIAS

ABE, T.; LEY, C. A tractable, parsimonious and flexible model for cylindrical data, with applications. **Econometrics and Statistics**, Elsevier, v. 4, p. 91–104, 2017. Citado na página 104.

ABE, T.; PEWSEY, A. Sine-skewed circular distributions. **Statistical Papers**, Springer, v. 52, n. 3, p. 683–707, 2011. Citado nas páginas 39 e 40.

ABUZAID, A.; ALLAHHAM, N. Simple circular regression model assuming wrapped cauchy error. **Pakistan Journal of Statistics**, v. 31, n. 4, p. 385–398, 2015. Citado na página 45.

ABUZAID, A.; HUSSIN, A.; MOHAMED, I. Identifying single outlier in linear circular regression model based on circular distance. **Journal of Applied Probability and Statistics**, v. 3, n. 1, p. 107–117, 2008. Citado na página 47.

____. Detection of outliers in simple circular regression models using the mean circular error statistic. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 83, n. 2, p. 269–277, 2013. Citado na página 46.

ABUZAID, A. H. Identifying density-based local outliers in medical multivariate circular data. **Statistics in Medicine**, Wiley Online Library, v. 39, n. 21, p. 2793–2798, 2020. Citado na página 28.

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. **Second International Symposium of Information Theory**, In B.N. Petrov and C. Csaki (Eds.). Budapest: Akademiai Kiado, p. 267–281, 1973. Citado nas páginas 89 e 90.

ANDERSON, T. W.; DARLING, D. A. A test of goodness of fit. **Journal of the American statistical association**, Taylor & Francis, v. 49, n. 268, p. 765–769, 1954. Citado nas páginas 54 e 92.

ANHOLETO, T.; SANDOVAL, M. C.; BOTTER, D. A. Adjusted pearson residuals in beta regression models. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 84, n. 5, p. 999–1014, 2014. Citado nas páginas 52 e 55.

ARTES, R. Hypothesis tests for covariance analysis models for circular data. **Communications in Statistics—Theory and Methods**, Taylor & Francis, v. 37, n. 10, p. 1632–1640, 2008. Citado na página 43.

ATKINSON, A. C. Two graphical displays for outlying and influential observations in regression. **Biometrika**, Oxford University Press, v. 68, n. 1, p. 13–20, 1981. Citado nas páginas 52 e 63.

AZZALINI, A. A class of distributions which includes the normal ones. **Scandinavian journal of statistics**, JSTOR, p. 171–178, 1985. Citado na página 39.

BARNDORFF-NIELSEN, O. Hyperbolic distributions and distributions on hyperbolae. **Scandinavian Journal of statistics**, JSTOR, p. 151–157, 1978. Citado na página 96.

BATSCHELET, E. Statistical methods for the analysis of problems in animal orientation and certain biological rhythms. American Institute of Biological Sciences, 1965. Citado na página 38.

BEST, D.; FISHER, N. I. Efficient simulation of the von mises distribution. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 28, n. 2, p. 152–157, 1979. Citado na página 38.

BORGIOLI, C.; MARTELLI, L.; PORRI, F.; MARCHETTI, G.; SCAPINI, F. *et al.* Orientation in talitrus saltator (montagu): trends in intrapopulation variability related to environmental and intrinsic factors. **Journal of Experimental Marine Biology and Ecology**, Elsevier, v. 238, n. 1, p. 29–47, 1999. Citado nas páginas 43 e 44.

BROYDEN, C. G. The convergence of a class of double-rank minimization algorithms 1. general considerations. **IMA Journal of Applied Mathematics**, Oxford University Press, v. 6, n. 1, p. 76–90, 1970. Citado na página 73.

CAO, L.; LI, D.; ZHANG, E.; ZHANG, Z.; SUN, H. A statistical cohomogeneity one metric on the upper plane with constant negative curvature. **Advances in Mathematical Physics**, Hindawi, v. 2014, 2014. Citado na página 41.

COLE, T. J.; GREEN, P. J. Smoothing reference centile curves: the lms method and penalized likelihood. **Statistics in medicine**, Wiley Online Library, v. 11, n. 10, p. 1305–1319, 1992. Citado na página 88.

COLONNA, M.; SAULEAU, E.-A. How to interpret and choose a bayesian spatial model and a poisson regression model in the context of describing small area cancer risks variations. **Revue d'epidemiologie et de sante publique**, Elsevier, v. 61, n. 6, p. 559–567, 2013. Citado na página 90.

COX, D. R.; SNELL, E. J. A general definition of residuals. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, v. 30, n. 2, p. 248–275, 1968. Citado na página 46.

CREMERS, J.; PENNINGS, H. J.; LEY, C. Regression models for cylindrical data in psychology. **Multivariate Behavioral Research**, Taylor & Francis, v. 55, n. 6, p. 910–925, 2020. Citado na página 104.

D'ELIA, A. A statistical model for orientation mechanism. **Statistical Methods and Applications**, Springer, v. 10, n. 1-3, p. 157–174, 2001. Citado na página 47.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996. Citado nas páginas 23, 24, 49, 69 e 74.

ESPINHEIRA, P. L.; FERRARI, S. L.; CRIBARI-NETO, F. On beta regression residuals. **Journal of Applied Statistics**, Taylor & Francis, v. 35, n. 4, p. 407–419, 2008. Citado nas páginas 48, 76 e 77.

FENG, C.; LI, L.; SADEGHPOUR, A. A comparison of residual diagnosis tools for diagnosing regression models for count data. **BMC Medical Research Methodology**, BioMed Central, v. 20, p. 175, 2020. Citado na página 49.

FENG, C.; SADEGHPOUR, A.; LI, L. Randomized quantile residuals: an omnibus model diagnostic tool with unified reference distribution. **arXiv preprint arXiv:1708.08527**, 2017. Citado na página 23.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of Applied Statistics**, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004. Citado nas páginas 69 e 70.

FILLIBEN, J. J. The probability plot correlation coefficient test for normality. **Technometrics**, Taylor & Francis, v. 17, n. 1, p. 111–117, 1975. Citado na página 91.

FISHER, N. I.; LEE, A. J. Regression models for an angular response. **Biometrics**, JSTOR, v. 48, n. 3, p. 665–677, 1992. Citado nas páginas 24, 42, 43, 44, 45 e 46.

FITZGERALD, T. M.; TAYLOR, P. D. Migratory orientation of juvenile yellow-rumped warblers (dendroica coronata) following stopover: sources of variation and the importance of geographic origins. **Behavioral Ecology and Sociobiology**, Springer, v. 62, n. 9, p. 1499–1508, 2008. Citado na página 43.

FLETCHER, R. A new approach to variable metric algorithms. **The computer journal**, Oxford University Press, v. 13, n. 3, p. 317–322, 1970. Citado na página 73.

FRAAIJE, R. G.; BRAAK, C. J. ter; VERDUYN, B.; BREEMAN, L. B.; VERHOEVEN, J. T.; SOONS, M. B. Early plant recruitment stages set the template for the development of vegetation patterns along a hydrological gradient. **Functional Ecology**, Wiley Online Library, v. 29, n. 7, p. 971–980, 2015. Citado na página 90.

GABARDA, S.; CRISTÓBAL, G. No-reference image quality assessment through the von mises distribution. **JOSA A**, Optical Society of America, v. 29, n. 10, p. 2058–2066, 2012. Citado na página 38.

GALVIS, D. M.; BANDYOPADHYAY, D.; LACHOS, V. H. Augmented mixed beta regression models for periodontal proportion data. **Statistics in medicine**, Wiley Online Library, v. 33, n. 21, p. 3759–3771, 2014. Citado na página 69.

GATTO, R.; JAMMALAMADAKA, S. R. The generalized von mises distribution. **Statistical Methodology**, Elsevier, v. 4, n. 3, p. 341–353, 2007. Citado na página 37.

GILL, J.; HANGARTNER, D. Circular data in political science and how to handle it. **Political Analysis**, JSTOR, v. 18, n. 3, p. 316–336, 2010. Citado nas páginas 43 e 44.

GOLDFARB, D. A family of variable-metric methods derived by variational means. **Mathematics of computation**, v. 24, n. 109, p. 23–26, 1970. Citado na página 73.

GOULD, A. L. A regression technique for angular variates. **Biometrics**, JSTOR, v. 25, n. 4, p. 683–700, 1969. Citado nas páginas 24, 28 e 42.

GREEN, P. J. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 46, n. 2, p. 149–170, 1984. Citado na página 45.

GREENWOOD, J. A.; DURAND, D. *et al.* The distribution of length and components of the sum of *n* random unit vectors. **The Annals of Mathematical Statistics**, Institute of Mathematical Statistics, v. 26, n. 2, p. 233–246, 1955. Citado na página 38.

GUMBEL, E.; GREENWOOD, J. A.; DURAND, D. The circular normal distribution: Theory and tables. **Journal of the American Statistical Association**, Taylor & Francis, v. 48, n. 261, p. 131–152, 1953. Citado na página 38.

HALL, D. B.; SHEN, J. Marginal projected multivariate linear models for clustered angular data. **Australian & New Zealand Journal of Statistics**, Wiley Online Library, v. 57, n. 2, p. 241–257, 2015. Citado na página 43.

HARRIS, R. A.; JOHNSON, J. A. Characterization of flake orientation in flakeboard by the von mises probability distribution function. **Wood and Fiber Science**, v. 14, n. 4, p. 254–266, 2007. Citado na página 38.

HASTIE, T.; TIBSHIRANI, R. Generalized additive models london chapman and hall. **Inc**, 1990. Citado na página 87.

HENSZ, C. M. Environmental factors in migratory route decisions: a case study on greenlandic arctic terns (sterna paradisaea). **Animal Migration**, De Gruyter, v. 1, n. open-issue, 2015. Citado na página 44.

HIDAYAH, S. N. **Outlier detection in cylindrical data/Nurul Hidayah Sadikon**. Tese (Doutorado) — University of Malaya, 2018. Citado na página 45.

HODGES, J. S. Some algebra and geometry for hierarchical models, applied to diagnostics. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 60, n. 3, p. 497–536, 1998. Citado na página 97.

HOSMER, D. W.; JOVANOVIC, B.; LEMESHOW, S. Best subsets logistic regression. **Biometrics**, JSTOR, p. 1265–1270, 1989. Citado na página 91.

HUANG, P.-H. Asymptotics of aic, bic, and rmsea for model selection in structural equation modeling. **Psychometrika**, Springer, v. 82, n. 2, p. 407–426, 2017. Citado na página 90.

HUSSIN, A.; FIELLER, N.; STILLMAN, E. Linear regression model for circular variables with application to directional data. **Journal of Applied Science and Technology**, v. 9, n. 1, p. 1–6, 2004. Citado na página 47.

IBRAHIM, S. **Some outlier problems in a circular regression model/Safwati binti Ibrahim**. Tese (Doutorado) — University of Malaya, 2013. Citado na página 47.

JAMMALAMADAKA, S. R.; LUND, U. J. The effect of wind direction on ozone levels: a case study. **Environmental and Ecological Statistics**, Springer, v. 13, n. 3, p. 287–298, 2006. Citado na página 43.

JAMMALAMADAKA, S. R.; SENGUPTA, A. **Topics in circular statistics**. [S.l.]: world scientific, 2001. v. 5. Citado nas páginas 35, 36, 37, 39, 41 e 47.

JHA, J.; BISWAS, A. Multiple circular–circular regression. **Statistical Modelling**, SAGE Publications Sage India: New Delhi, India, v. 17, n. 3, p. 142–171, 2017. Citado nas páginas 45 e 46.

JOHNSON, N. L.; KOTZ, S.; BALAKRISHNAN, N. **Continuous univariate distributions, volume 2**. [S.l.]: John wiley & sons, 1995. v. 289. Citado na página 88.

JOHNSON, R. A.; WEHRLY, T. E. Some angular-linear distributions and related regression models. **Journal of the American Statistical Association**, Taylor & Francis, v. 73, n. 363, p. 602–606, 1978. Citado na página 42.

JORGENSEN, B. **Statistical properties of the generalized inverse Gaussian distribution**. [S.l.]: Springer Science & Business Media, 2012. v. 9. Citado na página 96.

KATO, S.; SHIMIZU, K.; SHIEH, G. S. A circular–circular regression model. **Statistica Sinica**, JSTOR, v. 18, n. 2, p. 633–645, 2008. Citado na página 45.

KENT, J. T.; TYLER, D. E. Maximum likelihood estimation for the wrapped cauchy distribution. **Journal of Applied Statistics**, Taylor & Francis, v. 15, n. 2, p. 247–254, 1988. Citado nas páginas 38 e 41.

KIESCHNICK, R.; MCCULLOUGH, B. D. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. **Statistical modelling**, Sage Publications Sage CA: Thousand Oaks, CA, v. 3, n. 3, p. 193–213, 2003. Citado na página 90.

KIM, S. **Inverse circular regression with possibly asymmetric error distribution**. [S.l.]: University of California, Riverside, 2009. Citado na página 45.

KIM, S.; RIFAT, M. M. I. Diagnostic analysis of a circular-circular regression model using asymmetric or asymmetric bi-modal circular errors. **Communications in Statistics-Theory and Methods**, Taylor & Francis, p. 1–11, 2019. Citado nas páginas 43 e 45.

KIM, S.; SENGUPTA, A. A three-parameter generalized von mises distribution. **Statistical Papers**, Springer, v. 54, n. 3, p. 685–693, 2012. Citado nas páginas 37 e 43.

____. Multivariate-multiple circular regression. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 87, n. 7, p. 1277–1291, 2016. Citado na página 43.

____. Regressions involving circular variables: An overview. In: SPRINGER. **Platinum Jubilee International Conference on Applications of Statistics**. [S.l.], 2018. p. 25–33. Citado na página 45.

KLAR, B.; MEINTANIS, S. G. Specification tests for the response distribution in generalized linear models. **Computational Statistics**, Springer, v. 27, n. 2, p. 251–267, 2012. Citado nas páginas 91 e 92.

KUHA, J. Aic and bic: Comparisons of assumptions and performance. **Sociological methods & research**, Sage Publications Sage CA: Thousand Oaks, CA, v. 33, n. 2, p. 188–229, 2004. Citado na página 90.

KULLBACK, S.; LEIBLER, R. A. On information and sufficiency. **The annals of mathematical statistics**, JSTOR, v. 22, n. 1, p. 79–86, 1951. Citado na página 90.

LANG, M. N.; SCHLOSSER, L.; HOTHORN, T.; MAYR, G. J.; STAUFFER, R.; ZEILEIS, A. Circular regression trees and forests with an application to probabilistic wind direction forecasting. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 69, n. 5, p. 1357–1374, 2020. Citado na página 27.

LAWLESS, J.; SINGHAL, K. Efficient screening of nonnormal regression models. **Biometrics**, JSTOR, p. 318–327, 1978. Citado na página 91.

LAYCOCK, P. Optimal design: regression models for directions. **Biometrika**, Oxford University Press, v. 62, n. 2, p. 305–311, 1975. Citado na página 42.

LEMONTE, A. J.; MORENO-ARENAS, G. On residuals in generalized johnson sb regressions. **Applied Mathematical Modelling**, Elsevier, v. 67, p. 62–73, 2019. Citado nas páginas 23 e 49.

LEVY, P. L'addition des variables aléatoires définies sur une circonférence. **Bulletin de la Société mathématique de France**, v. 67, p. 1–41, 1939. Citado na página 40.

LEWIS, F.; BUTLER, A.; GILBERT, L. A unified approach to model selection using the likelihood ratio test. **Methods in Ecology and Evolution**, Wiley Online Library, v. 2, n. 2, p. 155–162, 2011. Citado nas páginas 25 e 87.

LIU, F.; KONG, Y. zoib: An r package for bayesian inference for beta regression and zero/one inflated beta regression. **R J.**, v. 7, n. 2, p. 34, 2015. Citado na página 72.

LIU, S.; MA, T.; SENGUPTA, A.; SHIMIZU, K.; WANG, M.-Z. Influence diagnostics in possibly asymmetric circular-linear multivariate regression models. **Sankhya B**, Springer, v. 79, n. 1, p. 76–93, 2016. Citado nas páginas 24 e 46.

LUND, U. Least circular distance regression for directional data. **Journal of Applied Statistics**, Taylor & Francis, v. 26, n. 6, p. 723–733, 1999. Citado na página 42.

LUND, U.; AGOSTINELLI, C.; AGOSTINELLI, M. C. Package 'circular'. **Repository CRAN**, 2017. Citado na página 50.

LUSCHI, P.; ÅKESSON, S.; BRODERICK, A. C.; GLEN, F.; GODLEY, B. J.; PAPI, F.; HAYS, G. C. Testing the navigational abilities of ocean migrants: displacement experiments on green sea turtles (chelonia mydas). **Behavioral Ecology and Sociobiology**, Springer, v. 50, n. 6, p. 528–534, 2001. Citado na página 30.

MAITRA, S. **Applications of Circular Distributions and Spatial Point Processes to the Analysis of Periodontal Data.** Tese (Doutorado), 2012. Citado na página 47.

MAITRA, S.; BRAUN, T. M. **Analysis of Periodontal Data using Circular Statistics**. [S.l.], 2012. Citado na página 44.

MALLOWS, C. Some comments on cp. **Technometrics**, JSTOR, v. 15, n. 4, p. 661, 1973. Citado na página 91.

MARCHETTI, G. M.; SCAPINI, F. Use of multiple regression models in the study of sandhopper orientation under natural conditions. **Estuarine, Coastal and Shelf Science**, Elsevier, v. 58, p. 207–215, 2003. Citado na página 65.

MARDIA, K. Distribution theory for the von mises-fisher distribution and its application. In: **A Modern Course on Statistical Distributions in Scientific Work**. [S.l.]: Springer, 1975. p. 113–130. Citado na página 38.

MARDIA, K. V. **Statistics of directional data**. [S.l.]: Academic press, 1972. Citado nas páginas 13, 28, 30, 32, 33, 36, 37 e 46.

MARDIA, K. V.; JUPP, P. E. **Directional statistics**. [S.l.]: John Wiley & Sons, 2009. v. 494. Citado nas páginas 34, 35 e 50.

MARTINEZ, R. O. **Modelos de regressão beta inflacionados**. Tese (Doutorado) — Universidade de São Paulo, 2008. Citado nas páginas 69 e 72.

MARUOTTI, A.; PUNZO, A.; MASTRANTONIO, G.; LAGONA, F. A time-dependent extension of the projected normal regression model for longitudinal circular data based on a hidden markov heterogeneity structure. **Stochastic Environmental Research and Risk Assessment**, Springer, v. 30, n. 6, p. 1725–1740, 2016. Citado na página 47.

MARZIO, M. D.; PANZERA, A.; TAYLOR, C. C. Non-parametric regression for circular responses. **Scandinavian Journal of Statistics**, Wiley Online Library, v. 40, n. 2, p. 238–255, 2013. Citado na página 68.

MEDEIROS, F. M.; FERRARI, S. L.; LEMONTE, A. J. Improved inference in dispersion models. **Applied Mathematical Modelling**, Elsevier, v. 51, p. 317–328, 2017. Citado na página 44.

MISES, R. V. Uber die "ganzzahligkeit" der atomgewicht und verwandte fragen. **Physikal**, v. 19, p. 490–500, 1918. Citado na página 38.

MIYATA, Y.; SHIOHAMA, T.; ABE, T. Identifiability of asymmetric circular and cylindrical distributions. **arXiv preprint arXiv:1908.09114**, 2019. Citado na página 39.

MUTWIRI, R. M. **Statistical distributions and modelling of GPS-Telemetry elephant movement data including the e ect of covariates**. Tese (Doutorado) — University of KwaZulu Natal, South Africa, 2015. Citado na página 47.

MUTWIRI, R. M. Application of multiple circular-linear regression models to animal movement data with covariates. 2016. Citado na página 47.

NELDER, J. A.; MEAD, R. A simplex method for function minimization. **The computer journal**, Oxford University Press, v. 7, n. 4, p. 308–313, 1965. Citado na página 45.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. **Journal of the Royal Statistical Society: Series A (General)**, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972. Citado nas páginas 47 e 87.

OLIVEIRA, W. L. de; DINIZ, C. A. R.; DURBÁN, M. A class of bivariate regression models for discrete and/or continuous responses. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, p. 1–25, 2018. Citado na página 104.

OSPINA, R.; FERRARI, S. L. Inflated beta distributions. **Statistical Papers**, Springer, v. 51, n. 1, p. 111, 2010. Citado nas páginas 24, 70 e 71.

____. A general class of zero-or-one inflated beta regression models. **Computational Statistics & Data Analysis**, Elsevier, v. 56, n. 6, p. 1609–1623, 2012. Citado nas páginas 24, 69, 71, 72, 76, 77, 80 e 85.

OTIENO, B. S. **An alternative estimate of preferred direction for circular data**. Tese (Doutorado) — Virginia Tech, 2002. Citado na página 33.

PAUL, S. R.; FUNG, K. Y. A generalized extreme studentized residual multiple-outlier-detection procedure in linear regression. **Technometrics**, Taylor & Francis, v. 33, n. 3, p. 339–348, 1991. Citado na página 73.

PAULA, G. A. Influence diagnostics in proper dispersion models. **Australian Journal of Statistics**, Wiley Online Library, v. 38, n. 3, p. 307–316, 1996. Citado na página 47.

PELUSO, A.; VINCIOTTI, V.; YU, K. Discrete weibull generalized additive model: An application to count fertility data. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 68, n. 3, p. 565–583, 2019. Citado na página 90.

PEREIRA, G. H. A. On quantile residuals in beta regression. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 48, n. 1, p. 302–316, 2019. Citado nas páginas 23, 49, 52, 55, 70 e 76.

PEREIRA, G. H. A.; BOTTER, D. A.; SANDOVAL, M. C. The truncated inflated beta distribution. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 41, n. 5, p. 907–919, 2012. Citado nas páginas 24, 70, 71 e 72.

____. A regression model for special proportions. **Statistical Modelling**, Sage Publications Sage India: New Delhi, India, v. 13, n. 2, p. 125–151, 2013. Citado nas páginas 24, 69 e 72.

PEREIRA, G. H. A.; SCUDILIO, J.; SANTOS-NETO, M.; BOTTER, D. A.; SANDOVAL, M. C. A class of residuals for outlier identification in zero adjusted regression models. **Journal of Applied Statistics**, Taylor Francis, v. 47, n. 10, p. 1833–1847, 2020. Citado nas páginas 24, 69, 73, 74, 76, 77, 78, 80, 81 e 85.

PEREIRA, G. H. d. A. **Modelos de regressão beta inflacionados truncados**. Tese (Doutorado) — Universidade de São Paulo, 2012. Citado na página 75.

PEREIRA, T. L.; CRIBARI-NETO, F. Detecting model misspecification in inflated beta regressions. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 43, n. 3, p. 631–656, 2014. Citado na página 72.

PEWSEY, A.; GARCÍA-PORTUGUÉS, E. Recent advances in directional statistics. **arXiv preprint arXiv:2005.06889**, 2020. Citado na página 45.

PEWSEY, A.; LEWIS, T.; JONES, M. The wrapped t family of circular distributions. **Australian & New Zealand Journal of Statistics**, Wiley Online Library, v. 49, n. 1, p. 79–91, 2007. Citado na página 37.

PREGIBON, D. J. Data analytic methods for generalized linear models. 1980. Citado na página 91.

PURVES, R. D. Optimum numerical integration methods for estimation of area-under-the-curve (AUC) and area-under-the-moment-curve (AUMC). **Journal of Pharmacokinetics and Biopharmaceutics**, Springer, v. 20, n. 3, p. 211–226, 1992. Citado na página 53.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2021. Disponível em: <https://www.R-project.org/>. Citado na página 45.

RAD, N. N.; BEKKER, A.; ARASHI, M. A unified model for skewed circular data. In: IEEE. **2020 IEEE 23rd International Conference on Information Fusion (FUSION)**. [S.l.], 2020. p. 1–6. Citado na página 40.

RAMBLI, A.; MOHAMED, I.; ABUZAID, A. H.; HUSSIN, A. G. Identification of influential observations in circular regression model. In: **Proceedings of the Regional Conference on Statistical Sciences 2010 (RCSS'10)**. [S.l.: s.n.], 2010. p. 195–203. Citado na página 46.

RAMLEE, I.; IBRAHIM, S.; LEOW, W.; YUSOFF, M. A review of detecting outlier in a circular regression model. **MS&E**, v. 767, n. 1, p. 012048, 2020. Citado na página 46.

RAO, J. **Some contributions to the analysis of circular data**. Tese (Doutorado) — Indian Statistical Institute, Kolkata, 1969. Citado na página 46.

RIGBY, R. A.; STASINOPOULOS, D. A semi-parametric additive model for variance heterogeneity. **Statistics and Computing**, Springer, v. 6, n. 1, p. 57–65, 1996. Citado na página 88.

RIGBY, R. A.; STASINOPOULOS, D. M. The gamlss project: a flexible approach to statistical modelling. In: UNIVERSITY OF SOUTHERN DENMARK. **New trends in statistical modelling: Proceedings of the 16th international workshop on statistical modelling**. [S.l.], 2001. v. 337, p. 345. Citado na página 87.

____. Generalized additive models for location, scale and shape. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 54, n. 3, p. 507–554, 2005. Citado nas páginas 87 e 88.

RIGBY, R. A.; STASINOPOULOS, M. D.; HELLER, G. Z.; BASTIANI, F. D. **Distributions for modeling location, scale, and shape: Using GAMLSS in R**. [S.l.]: CRC press, 2019. Citado na página 88.

RIVEST, L.-P. A decentred predictor for circular-circular regression. **Biometrika**, Oxford University Press, v. 84, n. 3, p. 717–726, 1997. Citado na página 42.

RODRÍGUEZ, C. E.; NÚÑEZ-ANTONIO, G.; ESCARELA, G. A bayesian mixture model for clustering circular data. **Computational Statistics & Data Analysis**, Elsevier, v. 143, p. 106842, 2020. Citado na página 27.

RUKHIN, A. Strongly symmetrical families and statistical analysis of their parameters. **Journal of Soviet Mathematics**, Springer, v. 9, n. 6, p. 886–910, 1978. Citado na página 96.

SAAVEDRA-NIEVES, P.; CRUJEIRAS, R. M. **HDiR: Directional Highest Density Regions**. [S.l.], 2021. R package version 1.1.1. Disponível em: <https://CRAN.R-project.org/package=HDiR>. Citado na página 65.

SARMA, Y.; JAMMALAMADAKA, S. Circular regression. In: **Statistical Science and Data Analysis. Proceedings of the Third Pacific Area Statistical Conference**. [S.l.: s.n.], 1993. p. 109–128. Citado na página 42.

SCAPINI, F.; ALOIA, A.; BOUSLAMA, M. F.; CHELAZZI, L.; COLOMBINI, I.; ELGTARI, M.; FALLACI, M.; MARCHETTI, G. M. Multiple regression analysis of the sources of variation in orientation of two sympatric sandhoppers, talitrus saltator and talorchestia brito, from an exposed mediterranean beach. **Behavioral Ecology and Sociobiology**, Springer, v. 51, n. 5, p. 403–414, 2002. Citado nas páginas 64 e 65.

SCHULGASSER, K. Fibre orientation in machine-made paper. **Journal of materials science**, Springer, v. 20, n. 3, p. 859–866, 1985. Citado na página 40.

SCHWARZ, G. Estimating the dimension of a model. **The annals of statistics**, JSTOR, p. 461–464, 1978. Citado nas páginas 89 e 90.

SCUDILIO, J.; PEREIRA, G. H. Adjusted quantile residual for generalized linear models. **Computational Statistics**, Springer, v. 35, n. 1, p. 399–421, 2020. Citado nas páginas 23 e 49.

SEN, P. K.; SINGER, J. M.; LIMA, A. C. P. de. **From finite sample to asymptotic methods in statistics**. [S.l.]: Cambridge University Press, 2010. v. 28. Citado na página 121.

SENGUPTA, A.; KIM, S. Statistical inference for homologous gene pairs between two circular genomes: a new circular–circular regression model. **Statistical Methods & Applications**, Springer, v. 25, n. 3, p. 421–432, 2015. Citado na página 43.

SENGUPTA, A.; KIM, S.; ARNOLD, B. C. Inverse circular–circular regression. **Journal of Multivariate Analysis**, Elsevier, v. 119, p. 200–208, 2013. Citado nas páginas 43 e 45.

SENGUPTA, A.; UGWUOWO, F. I. Asymmetric circular-linear multivariate regression models with applications to environmental data. **Environmental and Ecological Statistics**, Springer, v. 13, n. 3, p. 299–309, 2006. Citado na página 45.

SHANNO, D. F. Conditioning of quasi-newton methods for function minimization. **Mathematics of computation**, v. 24, n. 111, p. 647–656, 1970. Citado na página 73.

SOUZA, F.; PAULA, G. **Influência local e análise de resıduos em modelos de regressao von Mises**. Tese (Doutorado) — Tese de Doutorado. IME/USP, Sao Paulo, 1999. Citado na página 44.

SOUZA, F. A. D.; PAULA, G. A. Theory & methods: Deviance residuals for an angular response. **Australian & New Zealand Journal of Statistics**, Wiley Online Library, v. 44, n. 3, p. 345–356, 2002. Citado nas páginas 47, 48, 49, 54, 55, 58 e 67.

SOUZA, F. A. M. de. **Influência local e análise de resíduos em modelos de regressão Von Mises**. Tese (Doutorado) — Instituto de Matemática e Estatística da Universidade de São Paulo, 1999. Citado na página 55.

STASINOPOULOS, M.; RIGBY, B.; De Bastiani, F. **gamlss.data: Data for Generalised Additive Models for Location Scale and Shape**. [S.l.], 2021. R package version 6.0-2. Disponível em: <https://CRAN.R-project.org/package=gamlss.data>. Citado na página 98.

STEPHENS, M. A. Use of the von mises distribution to analyse continuous proportions. **Biometrika**, Oxford University Press, v. 69, n. 1, p. 197–203, 1982. Citado na página 38.

TANG, B.; FRYE, H. A.; GELFAND, A. E.; JR, J. A. S. Zero-inflated beta distribution regression modeling. **arXiv preprint arXiv:2112.07249**, 2021. Citado na página 70.

UMBACH, D.; JAMMALAMADAKA, S. R. Building asymmetry into circular distributions. **Statistics & probability letters**, Elsevier, v. 79, n. 5, p. 659–663, 2009. Citado nas páginas 37, 38, 39 e 40.

VRIEZE, S. I. Model selection and psychological theory: a discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic). **Psychological methods**, American Psychological Association, v. 17, n. 2, p. 228, 2012. Citado na página 90.

WANG, R. J.-H.; MALTHOUSE, E. C.; KRISHNAMURTHI, L. On the go: How mobile shopping affects customer purchase behavior. **Journal of retailing**, Elsevier, v. 91, n. 2, p. 217–234, 2015. Citado na página 70.

WATSON, G. S.; WILLIAMS, E. J. On the construction of significance tests on the circle and the sphere. **Biometrika**, JSTOR, v. 43, n. 3/4, p. 344–352, 1956. Citado na página 38.

APÊNDICE

A

DEMONSTRAÇÕES E TABELAS DO CAPÍTULO 2

A.1 Função densidade de probabilidade da *wrapped* Cauchy

Pela equação (2.21), tem-se que a *f.d.p* de uma variável *wrapped* Cauchy distribuída é dada por

$$g(q) = \frac{1}{2p}\frac{1-e^{-2g}}{1+e^{-2g}-2e^{-g}\cos(q-m)}, \quad 0 \leq q < 2p.$$

Sabe-se também que

(a) $\sinh(x) = \dfrac{e^x - e^{-x}}{2} = \dfrac{e^{2x}-1}{2e^x} = \dfrac{1-e^{-2x}}{2e^{-x}},$

(b) $\cosh(x) = \dfrac{e^x + e^{-x}}{2} = \dfrac{e^{2x}+1}{2e^x} = \dfrac{1+e^{-2x}}{2e^{-x}},$

em que sinh e cosh denotam as funções seno hiperbólico e cosseno hiperbólico, respectivamente.

Portanto, por (a) e (b) tem-se que

(i) $1-e^{-2g} = 2e^{-g}\ \sinh(g),$

(ii) $1+e^{-2g} = 2e^{-g}\ \cosh(g).$

Substituindo (i) e (ii) na equação (2.21), obtém-se

$$\begin{aligned} g(q) &= \frac{1}{2p}\frac{1-e^{-2g}}{1+e^{-2g}-2e^{-g}\cos(q-m)} \\ &= \frac{1}{2p}\frac{2e^{-g}\ \sinh(g)}{2e^{-g}\ \cosh(g)-2e^{-g}\cos(q-m)} \\ &= \frac{1}{2p}\frac{\sinh(g)}{\cosh(g)-\cos(q-m)}. \end{aligned}$$

A.2 Resíduo *deviance* para o modelo von Mises

Para a obtenção do resíduo *deviance*, como definido em (2.34), faz-se necessário o cálculo de $\ell_i(q_i|\tilde{m}_i,\hat{k})$ e $\ell_i(q_i|\hat{m}_i,\hat{k})$, a priori. Por (2.27), para o modelo de regressão von Mises, tem-se

(a)

$$\begin{aligned} \ell_i(q_i|\tilde{m}_i,\hat{k}) &= -\ln[2p\ I_0(\hat{k})]+\hat{k}\cos(q_i-\tilde{m}_i) \\ &= -\ln[2p\ I_0(\hat{k})]+\hat{k}\cos(q_i-q_i) \\ &= -\ln[2p\ I_0(\hat{k})]+\hat{k}\cos(0) \\ &= -\ln[2p\ I_0(\hat{k})]+\hat{k}, \end{aligned}$$

(b)

$$\ell_i(q_i|\hat{m}_i,\hat{k}) = -\ln[2p\ I_0(\hat{k})]+\hat{k}\cos(q_i-\hat{m}_i).$$

Portanto, por (a) e (b),

$$\begin{aligned} \ell_i(q_i|\tilde{m}_i,\hat{k})-\ell_i(q_i|\hat{m}_i,\hat{k}) &= -\ln[2p\ I_0(\hat{k})]+\hat{k}-\Big(-\ln[2p\ I_0(\hat{k})]+\hat{k}\cos(q_i-\hat{m}_i)\Big) \\ &= -\ln[2p\ I_0(\hat{k})]+\hat{k}+\ln[2p\ I_0(\hat{k})]-\hat{k}\cos(q_i-\hat{m}_i) \\ &= \hat{k}-\hat{k}\cos(q_i-\hat{m}_i) \\ &= \hat{k}\Big(1-\cos(q_i-\hat{m}_i)\Big). \end{aligned}$$

Logo, o resíduo *deviance* é dado por

$$\begin{aligned} d_i &= sgn(q_i-\hat{m}_i)\sqrt{2}\Big(\ell_i(q_i|\tilde{m}_i,\hat{k})-\ell_i(q_i|\hat{m}_i,\hat{k})\Big)^{1/2} \\ &= sgn(q_i-\hat{m}_i)\sqrt{2}\left(\hat{k}\Big(1-\cos(q_i-\hat{m}_i)\Big)\right)^{1/2} \\ &= sgn(q_i-\hat{m}_i)\sqrt{2\hat{k}}\Big(1-\cos(q_i-\hat{m}_i)\Big)^{1/2}. \end{aligned}$$

A.3 Resíduo *deviance* para o modelo *sine-skewed* von Mises

De forma análoga, para o modelo de regressão *sine-skewed* von Mises, tem-se por (2.29)

(a)

$$\begin{aligned}
\ell_i(q_i|\tilde{m}_i,\hat{k},\hat{l}\,) &= -\ln[2p\ I_0(\hat{k})]+\hat{k}\cos(q_i-\tilde{m}_i)+\ln\left[1+\hat{l}\,\mathrm{sen}\,(q_i-\tilde{m}_i)\right]\\
&= -\ln[2p\ I_0(\hat{k})]+\hat{k}\cos(q_i-q_i)+\ln\left[1+\hat{l}\,\mathrm{sen}\,(q_i-q_i)\right]\\
&= -\ln[2p\ I_0(\hat{k})]+\hat{k}\cos(0)+\ln\left[1+\hat{l}\,\mathrm{sen}\,(0)\right]\\
&= -\ln[2p\ I_0(\hat{k})]+\hat{k}+\ln(1)\\
&= -\ln[2p\ I_0(\hat{k})]+\hat{k},
\end{aligned}$$

(b)

$$\ell_i(q_i|\hat{m}_i,\hat{k},\hat{l}\,) = -\ln[2p\ I_0(\hat{k})]+\hat{k}\cos(q_i-\hat{m}_i)+\ln\left[1+\hat{l}\,\mathrm{sen}\,(q_i-\hat{m}_i)\right].$$

Portanto, por (a) e (b),

$$\begin{aligned}
\ell_i(q_i|\tilde{m}_i,\hat{k},\hat{l}\,)-\ell_i(q_i|\hat{m}_i,\hat{k},\hat{l}\,) &= -\ln[2p\ I_0(\hat{k})]+\hat{k}\\
&- \left(-\ln[2p\ I_0(\hat{k})]+\hat{k}\cos(q_i-\hat{m}_i)+\ln\left[1+\hat{l}\,\mathrm{sen}\,(q_i-\hat{m}_i)\right]\right)\\
&= \hat{k}-\hat{k}\cos(q_i-\hat{m}_i)-\ln\left[1+\hat{l}\,\mathrm{sen}\,(q_i-\hat{m}_i)\right]\\
&= \hat{k}\left(1-\cos(q_i-\hat{m}_i)\right)-\ln\left[1+\hat{l}\,\mathrm{sen}\,(q_i-\hat{m}_i)\right].
\end{aligned}$$

Logo, o resíduo *deviance* é dado por

$$\begin{aligned}
d_i &= sgn(q_i-\hat{m}_i)\sqrt{2}\left(\ell_i(q_i|\tilde{m}_i,\hat{k},\hat{l}\,)-\ell_i(q_i|\hat{m}_i,\hat{k},\hat{l}\,)\right)^{1/2}\\
&= sgn(q_i-\hat{m}_i)\sqrt{2}\left(\hat{k}\left(1-\cos(q_i-\hat{m}_i)\right)-\ln\left[1+\hat{l}\,\mathrm{sen}\,(q_i-\hat{m}_i)\right]\right)^{1/2}.
\end{aligned}$$

A.4 Resíduo *deviance* para o modelo *wrapped* Cauchy

Semelhantemente, para o modelo *wrapped* Cauchy, tem-se por (2.31)

(a)

$$\begin{aligned}
\ell_i(q_i|\tilde{m}_i,\hat{g}) &= -\ln(2p)+\ln\left[\sinh(\hat{g})\right]-\ln[\cosh(\hat{g})-\cos(q_i-\tilde{m}_i)]\\
&= -\ln(2p)+\ln\left[\sinh(\hat{g})\right]-\ln[\cosh(\hat{g})-\cos(q_i-q_i)]\\
&= -\ln(2p)+\ln\left[\sinh(\hat{g})\right]-\ln[\cosh(\hat{g})-\cos(0)]\\
&= -\ln(2p)+\ln\left[\sinh(\hat{g})\right]-\ln[\cosh(\hat{g})-1],
\end{aligned}$$

(b)

$$\ell_i(q_i|\hat{m}_i,\hat{g}) = -\ln(2p) + \ln\left[\sinh(\hat{g})\right] - \ln[\cosh(\hat{g}) - \cos(q_i - \hat{m}_i)].$$

Portanto, por (a) e (b),

$$\begin{aligned}\ell_i(q_i|\tilde{m}_i,\hat{g}) - \ell_i(q_i|\hat{m}_i,\hat{g}) &= -\ln(2p) + \ln\left[\sinh(\hat{g})\right] - \ln[\cosh(\hat{g}) - 1] \\ &- \left(-\ln(2p) + \ln\left[\sinh(\hat{g})\right] - \ln[\cosh(\hat{g}) - \cos(q_i - \hat{m}_i)]\right) \\ &= -\ln[\cosh(\hat{g}) - 1] + \ln[\cosh(\hat{g}) - \cos(q_i - \hat{m}_i)] \\ &= \ln\left[\frac{\cosh(\hat{g}) - \cos(q_i - \hat{m}_i)}{\cosh(\hat{g}) - 1}\right].\end{aligned}$$

Logo, o resíduo *deviance* é dado por

$$\begin{aligned}d_i &= sgn(q_i - \hat{m}_i)\sqrt{2}\Big(\ell_i(q_i|\tilde{m}_i,\hat{k}) - \ell_i(q_i|\hat{m}_i,\hat{k})\Big)^{1/2} \\ &= sgn(q_i - \hat{m}_i)\sqrt{2}\left(\ln\left[\frac{\cosh(\hat{g}) - \cos(q_i - \hat{m}_i)}{\cosh(\hat{g}) - 1}\right]\right)^{1/2}.\end{aligned}$$

A.5 Demonstração do teorema 2.3.1

Nesse apêndice, será demonstrado o Teorema 2.3.1 utilizando os quatro passos seguintes:

(I) $F^*(q_i;t_i) \sim U(0,1)$;

(II) $F^*(q_i;\hat{t}_i) \xrightarrow{D} U(0,1)$;

(III) $\hat{F}^*(q_i;\hat{t}_i) \xrightarrow{D} U(0,1)$;

(IV) $r^*_{q_i} \xrightarrow{D} N(0,1)$.

Prova de (I)

Para $0 \leq k \leq 1$,

$$P(F^*(q_i;t_i) \leq k) = P(\text{Md}_i - p \leq q_i \leq l),$$

em que l é tal que $\int_{\text{Md}_i-p}^{l} f(q_i)dq_i = k$. Como

$$P(\text{Md}_i - p \leq q_i \leq l) = \int_{\text{Md}_i-p}^{l} f(q_i)dq_i,$$

então

$$P(\text{Md}_i - p \leq q_i \leq l) = P(F^*(q_i;t_i) \leq k) = k$$

e, portanto,

$$F^*(q_i; t_i) \sim U(0,1).$$

Prova de (II)

Como t_i é consistentemente estimado, $\hat{t}_i \xrightarrow{P} t_i$ e $F^*(q_i; t_i) \sim U(0,1)$ (passo (I)), e utilizando uma modificação do Teorema de Slutsky (SEN; SINGER; LIMA, 2010, page 203),

$$F^*(q_i; \hat{t}_i) \xrightarrow{D} U(0,1).$$

Prova de (III)

Como Md_i é consistentemente estimado, $\widehat{\mathrm{Md}}_i \xrightarrow{P} \mathrm{Md}_i$ e $F^*(q_i; \hat{t}_i) \xrightarrow{D} U(0,1)$ (passo (II)), e utilizando a mesma modificação do Teorema de Slutsky em (II),

$$\hat{F}^*(q_i; \hat{t}_i) \xrightarrow{D} U(0;1).$$

Prova de (IV)

Como $\hat{F}^*(q_i; \hat{t}_i) \xrightarrow{D} U(0,1)$ (passo III), pelo Teorema de Sverdrup (SEN; SINGER; LIMA, 2010, page 180),

$$\mathrm{F}^{-1}\{\hat{F}^*(q_i; \hat{t}_i)\} \xrightarrow{D} \mathrm{F}^{-1}\{U(0,1)\},$$

e então

$$r^*_{q_i} \xrightarrow{D} N(0,1).$$

A.6 Tabelas dos estudos de simulação no modelo von Mises

Tabela A.1 – Resultados da simulação para $r_{q_i}^*$, r_i e d_i, considerando o cenário II com $n = 20$ - modelo de regressão von Mises.

i	m_i	Média			Variância			Assimetria			Curtose			Estatística A-D		
		$r_{q_i}^*$	r_i	d_i	$r_{q_i}^*$	r_i	d_i	$r_{q_i}^*$	r_i	d_i	$r_{q_i}^*$	r_i	d_i	$r_{q_i}^*$	r_i	d_i
1	1,87	0,02	−0,05	−0,05	1,01	1,02	1,09	0,05	−0,17	−0,18	2,83	2,81	2,82	1,57	5,81	9,50
2	−0,68	0,00	0,00	0,00	0,86	0,87	0,93	0,02	0,02	0,02	2,86	2,90	2,91	6,00	5,59	1,64
3	−0,91	0,03	0,03	0,03	1,01	1,02	1,09	0,02	0,02	0,03	2,81	2,84	2,85	2,74	2,83	5,85
4	0,84	−0,01	−0,01	−0,01	1,00	1,01	1,08	−0,03	−0,04	−0,04	2,76	2,80	2,80	0,82	0,86	3,94
5	−1,23	0,00	0,00	0,00	1,03	1,04	1,11	0,01	0,02	0,02	2,80	2,82	2,83	1,35	1,48	5,51
6	−0,61	0,00	0,00	0,00	1,00	1,01	1,08	0,03	0,03	0,04	2,83	2,87	2,88	0,69	0,72	3,24
7	−0,02	−0,01	−0,01	−0,01	0,95	0,96	1,03	−0,02	−0,02	−0,02	2,83	2,88	2,89	0,64	0,49	0,97
8	0,83	−0,03	−0,03	−0,03	1,01	1,02	1,09	−0,04	−0,04	−0,04	2,77	2,82	2,83	3,24	3,31	7,09
9	1,14	0,01	0,01	0,01	1,04	1,05	1,12	−0,08	−0,09	−0,10	2,83	2,85	2,87	3,22	3,33	8,10
10	1,26	−0,01	−0,02	−0,02	1,02	1,02	1,09	−0,03	−0,08	−0,08	2,78	2,77	2,79	2,10	2,49	6,52
11	0,26	0,00	0,00	0,00	1,08	1,10	1,17	0,04	0,04	0,04	2,76	2,80	2,80	4,77	5,14	12,02
12	−1,75	−0,01	0,03	0,04	0,99	0,99	1,06	−0,02	0,17	0,18	3,02	2,96	2,98	0,44	2,23	3,69
13	−0,63	0,01	0,01	0,01	1,04	1,05	1,12	0,01	0,01	0,01	2,72	2,76	2,76	2,64	2,82	7,95
14	0,01	0,01	0,01	0,01	0,87	0,88	0,93	−0,03	−0,03	−0,03	2,83	2,87	2,88	5,60	5,19	1,69
15	−0,19	−0,02	−0,02	−0,02	0,96	0,97	1,03	0,02	0,02	0,02	2,83	2,85	2,86	2,07	1,96	2,91
16	−1,37	0,02	0,03	0,03	1,05	1,06	1,13	0,02	0,06	0,06	2,96	2,93	2,93	2,56	3,19	7,56
17	1,62	0,02	−0,01	−0,01	1,01	1,02	1,09	0,03	−0,11	−0,12	2,96	2,90	2,92	1,13	0,97	3,90
18	0,87	−0,01	−0,01	−0,01	1,00	1,01	1,08	−0,01	−0,01	−0,02	2,75	2,79	2,80	1,79	1,81	5,29
19	−1,07	0,00	0,00	0,00	1,02	1,03	1,10	0,10	0,11	0,11	2,82	2,85	2,85	1,58	1,68	5,38
20	0,92	0,01	0,01	0,01	1,06	1,07	1,14	−0,05	−0,05	−0,05	2,85	2,89	2,91	2,46	2,71	7,72
Média		0,00	0,00	0,00	1,00	1,01	1,08	0,00	−0,01	−0,01	2,83	2,85	2,86	2,37	2,73	5,52
Desv. Pad.		0,02	0,02	0,02	0,06	0,06	0,06	0,04	0,07	0,08	0,07	0,05	0,06	1,58	1,63	2,85

Tabela A.2 – Resultados da simulação para $r^*_{q_i}$, r_i e d_i, considerando o cenário III com $n = 20$ - modelo de regressão von Mises.

i	m_i	Média			Variância			Assimetria			Curtose			Estatística A-D		
		$r^*_{q_i}$	r_i	d_i	$r^*_{q_i}$	r_i	d_i	$r^*_{q_i}$	r_i	d_i	$r^*_{q_i}$	r_i	d_i	$r^*_{q_i}$	r_i	d_i
1	1,88	0,00	−0,24	−0,26	1,06	1,01	1,18	−0,02	−0,25	−0,27	2,86	2,49	2,46	2,47	122,58	156,96
2	−0,83	0,00	0,02	0,02	0,87	0,87	1,02	0,02	0,15	0,15	3,12	3,04	3,03	8,89	9,13	1,96
3	−1,07	0,03	0,10	0,10	1,02	1,01	1,18	−0,08	0,17	0,17	3,00	2,79	2,78	3,01	18,42	32,30
4	1,00	−0,01	−0,05	−0,06	1,03	1,02	1,19	−0,03	−0,19	−0,19	3,04	2,87	2,85	0,58	6,25	18,29
5	−1,38	0,01	0,10	0,11	1,04	1,03	1,20	−0,01	0,18	0,19	2,83	2,70	2,67	1,33	21,10	39,69
6	−0,75	−0,01	0,02	0,03	0,98	0,98	1,15	−0,04	0,11	0,12	3,01	2,92	2,90	0,90	1,86	8,82
7	−0,02	0,01	0,00	0,00	0,90	0,90	1,05	−0,01	−0,04	−0,04	3,18	2,97	2,96	4,78	4,19	1,32
8	1,00	−0,01	−0,07	−0,08	1,02	1,01	1,18	0,04	−0,17	−0,17	3,01	2,83	2,82	0,90	10,23	22,31
9	1,30	0,00	−0,07	−0,08	1,03	1,03	1,20	−0,06	−0,22	−0,23	2,85	2,68	2,67	1,38	13,58	30,88
10	1,41	−0,03	−0,12	−0,13	1,04	1,03	1,20	−0,08	−0,25	−0,26	2,91	2,65	2,63	2,75	30,02	50,70
11	0,33	−0,01	−0,02	−0,02	1,06	1,06	1,23	0,04	0,00	0,00	2,89	2,78	2,76	2,54	3,04	20,81
12	−1,79	0,00	0,20	0,22	1,03	0,99	1,16	−0,02	0,25	0,26	2,87	2,54	2,52	1,04	86,62	112,06
13	−0,78	0,01	0,05	0,05	1,03	1,03	1,20	−0,07	0,07	0,08	2,91	2,76	2,75	2,30	5,92	21,48
14	0,01	0,01	0,01	0,02	0,82	0,81	0,94	−0,04	−0,05	−0,05	3,33	3,09	3,09	18,45	17,66	3,27
15	−0,24	−0,02	−0,02	−0,02	0,91	0,90	1,06	0,02	0,02	0,03	3,10	2,95	2,94	5,59	4,76	2,55
16	−1,51	0,02	0,15	0,16	1,08	1,06	1,23	−0,02	0,21	0,22	2,97	2,67	2,64	3,27	47,66	72,74
17	1,70	0,01	−0,16	−0,18	1,04	1,01	1,18	−0,01	−0,26	−0,27	2,95	2,60	2,57	1,47	56,45	79,25
18	1,04	−0,03	−0,07	−0,08	1,02	1,01	1,18	0,01	−0,15	−0,16	3,00	2,75	2,74	2,21	10,22	24,53
19	−1,23	0,01	0,08	0,09	1,04	1,03	1,20	0,03	0,24	0,25	2,97	2,76	2,74	1,01	14,95	31,18
20	1,08	0,01	−0,05	−0,06	1,03	1,03	1,20	0,04	−0,15	−0,15	2,89	2,79	2,77	1,38	6,65	21,20
Média		0,00	−0,01	−0,01	1,00	0,99	1,16	−0,01	−0,02	−0,02	2,98	2,78	2,77	3,31	24,56	37,61
Desv. Pad.		0,01	0,11	0,11	0,07	0,07	0,08	0,04	0,18	0,19	0,12	0,16	0,17	4,08	31,45	40,04

Tabela A.3 – Resultados da simulação para $r^*_{q_i}$, r_i e d_i, considerando o cenário IV com $n = 20$ - modelo de regressão von Mises.

i	m_i	Média			Variância			Assimetria			Curtose			Estatística A-D		
		$r^*_{q_i}$	r_i	d_i	$r^*_{q_i}$	r_i	d_i	$r^*_{q_i}$	r_i	d_i	$r^*_{q_i}$	r_i	d_i	$r^*_{q_i}$	r_i	d_i
1	1,19	0,02	0,01	0,01	0,94	0,94	1,01	−0,01	−0,03	−0,03	2,88	2,87	2,88	1,98	1,67	1,27
2	−1,64	−0,03	−0,01	−0,01	0,95	0,96	1,02	−0,05	0,11	0,12	3,02	2,98	3,01	3,39	1,38	1,15
3	−0,72	0,00	0,00	0,00	0,96	0,97	1,03	0,03	0,03	0,04	2,73	2,77	2,77	0,92	0,79	1,97
4	−0,81	0,00	0,00	0,00	0,91	0,92	0,98	−0,01	0,00	−0,01	2,77	2,80	2,81	2,04	1,76	0,84
5	0,39	−0,03	−0,03	−0,03	1,00	1,01	1,07	0,01	0,01	0,01	2,82	2,86	2,86	2,80	2,82	5,28
6	−0,74	0,02	0,02	0,02	0,99	1,00	1,07	0,01	0,01	0,01	2,84	2,88	2,88	0,89	0,90	3,04
7	−0,22	0,01	0,01	0,01	1,08	1,09	1,17	−0,01	−0,01	−0,01	2,76	2,80	2,80	4,45	4,83	11,47
8	0,51	−0,01	−0,01	−0,01	1,05	1,06	1,13	0,02	0,02	0,02	2,70	2,74	2,74	3,63	3,87	9,67
9	0,26	0,00	0,00	0,00	0,80	0,81	0,86	0,03	0,04	0,04	2,85	2,88	2,88	12,85	12,27	5,89
10	−0,80	0,02	0,02	0,02	1,02	1,03	1,10	0,00	0,00	0,00	2,88	2,93	2,94	1,41	1,52	4,68
11	1,03	−0,01	−0,01	−0,01	1,00	1,01	1,08	−0,05	−0,06	−0,06	2,81	2,85	2,86	0,80	0,86	3,64
12	0,01	0,00	0,00	0,00	1,06	1,07	1,14	−0,01	−0,01	−0,01	2,71	2,74	2,75	4,24	4,47	10,67
13	0,31	0,03	0,03	0,03	1,07	1,08	1,15	0,02	0,02	0,02	2,80	2,84	2,84	5,05	5,33	11,39
14	0,13	0,00	0,00	0,00	0,99	1,00	1,07	−0,03	−0,03	−0,03	2,78	2,80	2,81	0,49	0,51	2,97
15	0,02	−0,01	−0,01	−0,01	1,10	1,11	1,18	−0,05	−0,05	−0,05	2,69	2,73	2,73	7,32	7,75	15,85
16	0,60	−0,02	−0,02	−0,02	1,07	1,09	1,16	0,04	0,04	0,04	2,73	2,77	2,77	4,92	5,27	11,83
17	0,65	0,00	0,00	0,00	1,03	1,04	1,11	0,00	0,00	−0,01	2,75	2,79	2,79	1,94	2,08	6,66
18	0,44	−0,01	−0,01	−0,01	0,92	0,93	0,99	0,02	0,02	0,02	2,83	2,86	2,86	2,18	1,93	1,01
19	0,12	0,02	0,02	0,02	1,08	1,09	1,17	−0,04	−0,04	−0,04	2,74	2,78	2,79	6,32	6,66	13,85
20	0,42	−0,01	−0,01	−0,01	0,98	0,99	1,06	−0,04	−0,05	−0,05	2,73	2,76	2,76	1,33	1,29	3,92
Média		0,00	0,00	0,00	1,00	1,01	1,08	−0,01	0,00	0,00	2,79	2,82	2,83	3,45	3,40	6,35
Desv. Pad.		0,02	0,01	0,01	0,07	0,07	0,08	0,03	0,04	0,04	0,08	0,07	0,07	2,94	2,98	4,75

Tabela A.4 – Resultados da simulação para $r^*_{q_i}$, r_i e d_i, considerando o cenário V com $n = 20$ - modelo de regressão von Mises.

i	m_i	Média			Variância			Assimetria			Curtose			Estatística A-D		
		$r^*_{q_i}$	r_i	d_i	$r^*_{q_i}$	r_i	d_i	$r^*_{q_i}$	r_i	d_i	$r^*_{q_i}$	r_i	d_i	$r^*_{q_i}$	r_i	d_i
1	2,18	−0,03	−0,13	−0,14	1,06	1,05	1,13	0,02	−0,29	−0,30	2,84	2,74	2,73	4,77	37,78	44,84
2	−0,95	−0,01	−0,03	−0,04	0,81	0,80	0,86	0,02	0,06	0,06	2,94	2,99	2,99	15,43	19,75	12,20
3	−1,55	0,07	0,05	0,05	1,01	1,01	1,09	−0,06	0,10	0,10	2,89	2,89	2,88	16,82	4,90	8,00
4	1,05	−0,05	0,02	0,02	0,98	0,96	1,03	0,03	−0,08	−0,08	2,75	2,81	2,79	8,08	3,08	4,93
5	−1,77	0,02	0,02	0,02	1,05	1,05	1,12	−0,09	0,15	0,15	2,89	2,89	2,88	3,05	2,72	7,28
6	−0,93	0,01	−0,02	−0,02	0,97	0,95	1,02	0,02	0,06	0,06	2,91	2,93	2,93	0,90	1,67	1,77
7	−0,45	0,04	0,01	0,01	0,89	0,85	0,91	−0,09	−0,06	−0,06	2,79	2,78	2,76	8,29	8,36	3,69
8	1,37	−0,06	0,00	0,00	1,03	1,01	1,09	0,04	−0,13	−0,13	2,83	2,84	2,83	12,63	2,72	7,04
9	1,53	−0,01	0,02	0,03	1,04	1,03	1,11	0,08	−0,15	−0,15	2,86	2,85	2,85	4,50	6,52	12,14
10	1,75	−0,05	−0,02	−0,02	1,03	1,03	1,11	0,05	−0,20	−0,20	2,75	2,76	2,75	9,93	5,95	11,85
11	0,32	−0,01	0,01	0,01	1,08	1,01	1,09	0,07	0,03	0,03	2,71	2,72	2,69	6,77	3,77	9,17
12	−2,17	−0,01	0,12	0,13	1,04	1,03	1,11	−0,06	0,30	0,31	2,81	2,70	2,70	1,99	33,55	39,07
13	−1,21	0,06	0,01	0,01	1,02	1,01	1,09	−0,07	0,01	0,01	2,78	2,79	2,77	14,15	3,31	8,01
14	0,23	0,04	0,04	0,05	0,79	0,75	0,80	−0,10	−0,10	−0,10	2,90	2,82	2,81	21,34	30,76	20,83
15	−0,75	0,02	−0,02	−0,02	0,91	0,88	0,95	−0,03	0,00	0,00	2,78	2,81	2,80	3,32	5,20	2,19
16	−1,91	0,03	0,06	0,07	1,07	1,07	1,15	−0,10	0,21	0,22	2,94	2,91	2,91	5,96	9,80	15,37
17	1,97	0,00	−0,05	−0,05	1,05	1,06	1,13	0,05	−0,25	−0,26	2,83	2,82	2,82	3,34	7,90	13,70
18	1,17	−0,07	0,00	0,00	0,98	0,97	1,04	0,05	−0,06	−0,06	2,72	2,77	2,75	16,79	3,19	5,61
19	−1,72	0,03	0,02	0,02	1,06	1,06	1,13	−0,04	0,19	0,19	2,93	2,89	2,88	5,28	4,22	9,55
20	1,40	−0,04	0,02	0,02	1,06	1,05	1,13	0,02	−0,15	−0,16	2,98	2,98	2,98	7,14	3,98	9,13
Média		0,00	0,01	0,01	1,00	0,98	1,05	−0,01	−0,02	−0,02	2,84	2,83	2,82	8,52	9,96	12,32
Desv. Pad.		0,04	0,05	0,05	0,08	0,09	0,10	0,06	0,16	0,16	0,08	0,08	0,09	5,79	11,14	11,15

A.7 Tabelas dos estudos de simulação no modelo *sine-skewed* von Mises

Tabela A.5 – Resultados da simulação para $r^*_{q_i}$, considerando o cenário II com $n = 20$ - modelo de regressão *sine-skewed* von Mises.

i	m_i	Média	Variância	Assimetria	Curtose	Estatística A-D
1	1,87	0,04	0,95	0,00	2,73	5,19
2	−0,68	−0,01	0,86	0,01	2,81	6,47
3	−0,91	0,03	1,00	0,00	2,75	2,83
4	0,84	−0,02	1,00	−0,04	2,73	1,28
5	−1,23	0,00	1,02	−0,03	2,72	1,50
6	−0,61	−0,01	0,99	0,01	2,76	1,10
7	−0,02	−0,02	0,94	−0,03	2,82	1,80
8	0,83	−0,04	1,01	−0,05	2,75	4,90
9	1,14	0,01	1,04	−0,07	2,80	3,51
10	1,26	−0,01	1,01	−0,04	2,70	2,16
11	0,26	−0,02	1,07	0,02	2,75	5,07
12	−1,75	0,04	0,93	0,02	2,86	5,91
13	−0,63	0,00	1,04	0,00	2,70	2,51
14	0,01	−0,01	0,86	−0,02	2,81	5,62
15	−0,19	−0,04	0,96	0,03	2,82	4,20
16	−1,37	0,04	1,02	0,00	2,83	5,41
17	1,62	0,03	0,99	0,01	2,81	3,22
18	0,87	−0,02	1,00	−0,02	2,73	2,55
19	−1,07	0,00	1,01	0,07	2,73	1,52
20	0,92	0,00	1,06	−0,07	2,83	2,20
Média		0,00	0,99	−0,01	2,77	3,45
Desv. Pad.		0,03	0,06	0,04	0,05	1,75

Tabela A.6 – Resultados da simulação para $r_{q_i}^*$, considerando o cenário III com $n = 20$ - modelo de regressão *sine-skewed* von Mises.

i	m_i	Média	Variância	Assimetria	Curtose	Estatística A-D
1	1,88	0,02	0,99	−0,03	3,22	3,05
2	−0,83	−0,01	0,85	−0,03	3,16	10,99
3	−1,07	0,02	1,02	−0,04	3,03	1,87
4	1,00	0,00	1,03	−0,01	3,02	0,34
5	−1,38	0,02	1,02	−0,08	2,88	2,35
6	−0,75	−0,04	0,99	−0,11	3,31	3,37
7	−0,02	−0,04	0,90	0,00	3,13	8,81
8	1,00	−0,01	1,00	0,03	2,90	0,97
9	1,30	0,01	1,04	−0,19	3,36	2,53
10	1,41	0,01	1,02	−0,07	2,90	2,11
11	0,33	−0,03	1,05	0,05	2,89	3,67
12	−1,79	0,05	0,95	−0,14	2,97	10,32
13	−0,78	−0,02	1,03	−0,03	2,89	1,94
14	0,01	0,00	0,84	−0,01	3,55	11,98
15	−0,24	−0,07	0,92	0,01	3,24	16,38
16	−1,51	0,05	1,02	−0,10	2,84	8,89
17	1,70	0,03	0,97	−0,05	2,90	2,55
18	1,04	−0,01	1,02	0,00	2,90	1,09
19	−1,23	0,00	1,02	−0,02	2,86	0,78
20	1,08	0,02	1,02	−0,04	2,83	2,44
Média		0,00	0,98	−0,04	3,04	4,82
Desv. Pad.		0,03	0,06	0,06	0,21	4,61

Tabela A.7 – Resultados da simulação para $r_{q_i}^*$, considerando o cenário IV com $n = 20$ - modelo de regressão *sine-skewed* von Mises.

i	m_i	Média	Variância	Assimetria	Curtose	Estatística A-D
1	1,19	0,05	0,89	−0,04	2,74	8,92
2	−1,64	0,06	0,78	0,00	2,87	26,37
3	−0,72	0,01	0,94	0,02	2,68	1,15
4	−0,81	0,00	0,90	−0,01	2,72	2,45
5	0,39	−0,04	0,99	0,02	2,81	5,62
6	−0,74	0,02	0,99	0,00	2,81	0,96
7	−0,22	−0,01	1,08	0,00	2,77	3,83
8	0,51	−0,02	1,05	0,02	2,67	4,38
9	0,26	−0,03	0,79	0,06	2,81	16,27
10	−0,80	0,03	1,02	−0,01	2,83	2,31
11	1,03	0,01	0,97	−0,08	2,78	1,70
12	0,01	−0,01	1,06	−0,02	2,70	4,20
13	0,31	0,02	1,06	0,02	2,77	3,71
14	0,13	0,00	0,99	−0,03	2,73	0,49
15	0,02	−0,03	1,09	−0,04	2,68	7,76
16	0,60	−0,02	1,07	0,03	2,69	4,82
17	0,65	0,00	1,02	−0,01	2,70	2,10
18	0,44	−0,01	0,92	0,01	2,77	1,73
19	0,12	0,00	1,07	−0,03	2,70	4,69
20	0,42	−0,01	0,98	−0,04	2,67	1,66
Média		0,00	0,98	−0,01	2,75	5,26
Desv. Pad.		0,02	0,09	0,03	0,06	6,14

Tabela A.8 – Resultados da simulação para $r_{q_i}^*$, considerando o cenário V com $n = 20$ - modelo de regressão *sine-skewed* von Mises.

i	m_i	Média	Variância	Assimetria	Curtose	Estatística A-D
1	1,88	0,13	0,99	−0,01	2,67	44,34
2	−0,83	−0,02	0,84	0,07	2,83	10,98
3	−1,07	0,04	0,99	0,06	2,77	4,17
4	1,00	−0,01	1,01	0,00	2,77	1,26
5	−1,38	0,06	1,01	0,01	2,72	10,49
6	−0,75	−0,04	0,96	0,07	2,80	5,46
7	−0,02	−0,08	0,87	0,06	2,86	22,58
8	1,00	−0,03	1,01	−0,01	2,85	4,02
9	1,30	0,06	1,04	0,00	2,78	10,81
10	1,41	0,04	1,02	−0,02	2,66	7,17
11	0,33	−0,07	1,04	0,10	2,82	17,37
12	−1,79	0,15	0,93	−0,01	2,81	57,24
13	−0,78	−0,02	1,01	0,06	2,72	2,90
14	0,01	−0,06	0,78	0,07	2,83	26,81
15	−0,24	−0,09	0,88	0,11	2,86	31,04
16	−1,51	0,12	1,01	0,00	2,73	32,82
17	1,70	0,11	0,99	−0,01	2,68	31,41
18	1,04	−0,01	0,99	0,01	2,75	2,14
19	−1,23	0,04	1,01	0,11	2,74	4,14
20	1,08	0,01	1,05	−0,03	2,88	2,39
Média		0,02	0,97	0,03	2,78	16,48
Desv. Pad.		0,07	0,07	0,05	0,07	16,01

Tabela A.9 – Resultados da simulação para $r_{q_i}^*$, considerando o cenário VI com $n = 20$ - modelo de regressão *sine-skewed* von Mises.

i	m_i	Média	Variância	Assimetria	Curtose	Estatística A-D
1	1,88	−0,08	0,98	0,07	2,73	19,76
2	−0,83	0,00	0,88	0,01	2,83	4,31
3	−1,07	0,02	1,01	0,02	2,76	2,09
4	1,00	0,01	0,99	−0,03	2,76	0,76
5	−1,38	−0,04	1,04	−0,02	2,71	6,17
6	−0,75	0,00	1,00	0,01	2,79	0,48
7	−0,02	0,04	0,88	−0,04	2,88	8,42
8	1,00	−0,02	0,99	−0,04	2,76	1,27
9	1,30	−0,01	1,04	−0,06	2,74	3,45
10	1,41	−0,04	1,03	−0,01	2,71	7,19
11	0,33	0,04	1,04	0,00	2,80	5,81
12	−1,79	−0,08	0,98	−0,02	2,83	14,76
13	−0,78	0,02	1,03	−0,02	2,70	3,86
14	0,01	0,05	0,79	−0,01	2,85	22,61
15	−0,24	0,01	0,90	0,00	2,89	3,27
16	−1,51	−0,02	1,06	−0,03	3,04	3,56
17	1,70	−0,05	1,01	0,06	2,77	9,30
18	1,04	−0,01	1,00	−0,02	2,74	1,45
19	−1,23	−0,03	1,04	0,07	2,75	4,76
20	1,08	0,01	1,05	−0,07	2,81	2,45
Média		−0,01	0,99	−0,01	2,79	6,29
Desv. Pad.		0,04	0,07	0,04	0,08	6,14

Tabela A.10 – Resultados da simulação para $r^*_{q_i}$, considerando o cenário VII com $n = 20$ - modelo de regressão *sine-skewed* von Mises.

i	m_i	Média	Variância	Assimetria	Curtose	Estatística A-D
1	2,18	0,05	1,02	−0,13	2,73	11,19
2	−0,95	−0,01	0,85	−0,01	2,86	8,50
3	−1,55	0,07	1,01	−0,05	2,91	14,19
4	1,05	−0,04	0,98	0,00	2,74	4,04
5	−1,77	0,04	1,02	−0,05	2,77	5,04
6	−0,93	0,00	0,98	0,00	2,83	0,34
7	−0,45	−0,01	0,89	−0,04	2,86	3,81
8	1,37	−0,04	1,02	0,01	2,66	5,85
9	1,53	0,03	1,01	0,00	2,73	3,53
10	1,75	0,00	1,01	−0,08	2,72	2,27
11	0,32	−0,06	1,08	0,04	2,73	14,70
12	−2,17	0,03	0,99	−0,03	2,70	4,35
13	−1,21	0,04	1,04	−0,08	2,70	9,41
14	0,23	−0,01	0,79	0,03	2,77	14,75
15	−0,75	−0,03	0,92	−0,06	3,07	3,35
16	−1,91	0,05	1,03	−0,09	2,88	9,49
17	1,97	0,06	0,99	−0,06	2,73	10,26
18	1,17	−0,04	0,98	0,02	2,70	5,21
19	−1,72	0,04	1,04	0,01	2,78	4,65
20	1,40	0,00	1,05	−0,05	2,93	1,73
Média		0,01	0,98	−0,03	2,79	6,83
Desv. Pad.		0,04	0,07	0,05	0,10	4,43

A.8 Tabelas dos estudos de simulação no modelo *wrapped* Cauchy

Tabela A.11 – Resultados da simulação para $r^*_{q_i}$ e d_i, considerando o cenário II com $n = 20$ - modelo de regressão *wrapped* Cauchy.

i	m_i	Média		Variância		Assimetria		Curtose		Estatística A-D	
		$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i
1	1,87	−0,02	−0,23	1,13	2,75	0,01	−0,06	2,86	2,20	9,49	746,06
2	−0,68	0,02	0,09	1,02	2,50	−0,01	0,05	3,29	2,59	28,32	440,21
3	−0,91	−0,02	0,05	1,10	2,75	0,00	0,04	2,87	2,24	5,83	649,93
4	0,84	0,01	−0,07	1,05	2,60	0,03	−0,06	3,07	2,38	6,32	525,94
5	−1,23	−0,01	0,09	1,10	2,74	0,00	0,04	2,87	2,26	6,69	642,66
6	−0,61	−0,01	0,05	1,08	2,67	−0,01	0,05	3,07	2,36	6,84	560,47
7	−0,02	0,01	0,02	1,05	2,62	0,02	−0,01	3,05	2,38	6,26	532,51
8	0,83	0,00	−0,06	1,09	2,67	−0,10	−0,08	3,05	2,32	5,88	579,12
9	1,14	−0,01	−0,11	1,11	2,74	−0,05	−0,08	2,94	2,20	5,80	675,96
10	1,26	−0,01	−0,14	1,06	2,63	0,04	−0,07	2,99	2,29	4,27	599,41
11	0,26	0,02	0,02	1,13	2,82	−0,04	−0,04	2,80	2,16	10,90	726,08
12	−1,75	−0,02	0,15	1,04	2,58	0,00	0,11	3,05	2,32	5,13	565,54
13	−0,63	0,00	0,06	1,09	2,76	−0,01	0,03	2,80	2,20	5,47	675,50
14	0,01	−0,02	−0,03	1,05	2,56	0,02	0,01	3,36	2,58	21,65	448,06
15	−0,19	−0,03	−0,02	1,08	2,67	0,01	0,03	3,11	2,38	8,83	554,26
16	−1,37	0,00	0,12	1,09	2,70	0,00	0,07	2,93	2,25	4,65	639,28
17	1,62	−0,01	−0,17	1,07	2,64	0,03	−0,07	2,94	2,27	4,51	630,83
18	0,87	0,01	−0,07	1,10	2,75	0,03	−0,03	2,83	2,21	5,60	670,14
19	−1,07	0,00	0,08	1,07	2,70	0,00	0,05	2,84	2,26	4,37	618,64
20	0,92	0,00	−0,09	1,10	2,71	0,09	−0,01	3,04	2,30	3,93	621,38
Média		0,00	−0,01	1,08	2,68	0,00	0,00	2,99	2,31	8,04	605,10
Desv. Pad.		0,01	0,10	0,03	0,08	0,04	0,06	0,15	0,12	6,16	81,00

Tabela A.12 – Resultados da simulação para $r^*_{q_i}$ e d_i, considerando o cenário III com $n = 20$ - modelo de regressão *wrapped* Cauchy.

i	m_i	Média		Variância		Assimetria		Curtose		Estatística A-D	
		$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i
1	1,88	−0,01	−0,42	1,17	2,38	0,01	0,06	2,74	1,93	13,71	973,55
2	−0,83	0,01	0,15	1,03	2,17	0,00	0,04	3,34	2,38	26,41	386,07
3	−1,07	−0,02	0,15	1,12	2,42	−0,05	0,03	2,89	2,05	8,18	604,92
4	1,00	0,00	−0,17	1,05	2,26	0,02	−0,04	3,08	2,19	7,00	481,06
5	−1,38	0,01	0,25	1,11	2,38	−0,01	−0,03	2,77	2,02	7,94	686,63
6	−0,75	0,01	0,12	1,09	2,32	0,04	0,03	3,06	2,20	9,32	479,72
7	−0,02	0,00	0,00	1,03	2,21	−0,02	−0,01	3,23	2,33	15,52	369,08
8	1,00	0,01	−0,12	1,06	2,30	−0,03	−0,03	3,01	2,17	6,15	480,83
9	1,30	0,00	−0,21	1,10	2,38	−0,01	−0,02	2,84	2,00	5,95	642,40
10	1,41	0,00	−0,22	1,07	2,31	0,02	−0,01	2,88	2,07	3,84	595,67
11	0,33	0,02	−0,01	1,13	2,47	−0,03	−0,01	2,84	2,03	8,89	595,49
12	−1,79	−0,01	0,31	1,09	2,31	−0,01	0,01	2,86	1,98	4,11	720,52
13	−0,78	0,00	0,13	1,10	2,40	−0,01	0,01	2,91	2,07	5,24	572,33
14	0,01	−0,01	−0,03	1,00	2,10	−0,03	0,00	3,53	2,60	47,88	286,30
15	−0,24	−0,03	0,01	1,05	2,26	−0,05	0,02	3,19	2,32	16,59	394,74
16	−1,51	0,01	0,28	1,13	2,39	−0,01	0,00	2,82	2,00	8,59	723,84
17	1,70	−0,02	−0,32	1,10	2,34	0,00	0,02	2,82	1,99	6,41	750,77
18	1,04	0,00	−0,17	1,10	2,38	0,05	0,02	2,91	2,06	5,45	603,75
19	−1,23	0,01	0,23	1,11	2,37	−0,03	0,01	2,91	2,06	6,72	635,49
20	1,08	−0,03	−0,21	1,08	2,33	0,02	0,01	2,99	2,12	5,77	575,52
Média		0,00	−0,01	1,09	2,32	−0,01	0,00	2,98	2,13	10,98	577,93
Desv. Pad.		0,01	0,21	0,04	0,09	0,03	0,03	0,21	0,17	10,21	158,42

Tabela A.13 – Resultados da simulação para $r^*_{q_i}$ e d_i, considerando o cenário IV com $n = 20$ - modelo de regressão *wrapped* Cauchy.

i	m_i	Média		Variância		Assimetria		Curtose		Estatística A-D	
		$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i
1	1,88	0,00	−0,43	1,09	1,53	−0,05	0,20	2,99	2,06	5,30	647,40
2	−0,83	0,06	0,10	1,05	1,52	−0,03	0,02	3,29	2,30	35,21	146,64
3	−1,07	0,03	0,14	1,18	1,74	−0,04	−0,04	2,82	1,93	17,22	327,31
4	1,00	−0,02	−0,18	1,08	1,58	0,06	0,07	3,09	2,15	11,93	247,82
5	−1,38	0,04	0,20	1,14	1,68	−0,05	−0,05	2,87	1,94	13,67	347,92
6	−0,75	0,03	0,08	1,11	1,64	−0,07	0,00	3,17	2,09	13,35	213,79
7	−0,02	0,00	−0,06	1,07	1,58	0,00	0,05	3,19	2,22	16,76	164,16
8	1,00	0,01	−0,16	1,09	1,61	0,02	0,09	3,03	2,14	11,87	245,39
9	1,30	−0,01	−0,27	1,16	1,68	0,01	0,14	2,83	1,95	11,27	440,54
10	1,41	−0,03	−0,30	1,12	1,62	−0,01	0,12	2,88	2,01	9,67	437,74
11	0,33	−0,04	−0,12	1,18	1,77	0,01	0,11	2,71	1,91	20,81	350,36
12	−1,79	0,03	0,32	1,05	1,56	−0,06	−0,10	2,97	2,02	5,93	426,64
13	−0,78	0,01	0,07	1,18	1,75	−0,02	0,01	2,81	1,91	14,45	303,32
14	0,01	0,02	−0,02	0,98	1,44	−0,01	0,02	3,38	2,42	40,01	96,62
15	−0,24	0,02	−0,03	1,12	1,65	0,00	0,04	3,06	2,16	17,97	191,02
16	−1,51	0,03	0,22	1,13	1,68	0,01	−0,06	2,91	1,96	10,97	358,96
17	1,70	0,00	−0,40	1,10	1,52	0,04	0,17	2,99	2,07	6,35	565,83
18	1,04	−0,03	−0,25	1,16	1,67	0,13	0,16	2,95	2,02	16,19	403,41
19	−1,23	0,04	0,18	1,16	1,71	−0,07	−0,04	2,86	1,96	17,34	335,62
20	1,08	−0,03	−0,24	1,11	1,64	0,03	0,14	2,90	2,04	10,19	366,41
Média		0,01	−0,06	1,11	1,63	−0,01	0,05	2,98	2,06	15,32	330,85
Desv. Pad.		0,03	0,22	0,05	0,09	0,05	0,09	0,17	0,14	8,73	137,81

Tabela A.14 – Resultados da simulação para $r^*_{q_i}$ e d_i, considerando o cenário V com $n = 20$ - modelo de regressão *wrapped* Cauchy.

i	m_i	Média		Variância		Assimetria		Curtose		Estatística A-D	
		$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i
1	1,19	−0,02	−0,13	1,08	2,66	−0,04	−0,05	3,06	2,39	14,16	567,23
2	−1,64	0,02	0,17	1,05	2,61	0,01	0,10	2,97	2,34	9,46	584,08
3	−0,72	−0,01	0,04	1,06	2,65	0,00	0,02	2,98	2,35	5,86	558,70
4	−0,81	0,00	0,05	1,04	2,56	0,02	0,03	3,19	2,47	11,37	480,70
5	0,39	−0,01	−0,06	1,09	2,71	0,08	0,01	2,93	2,30	6,86	610,39
6	−0,74	0,00	0,07	1,08	2,66	−0,01	0,04	3,06	2,34	5,59	569,53
7	−0,22	0,01	0,03	1,11	2,78	0,03	0,01	2,79	2,19	7,05	690,96
8	0,51	−0,01	−0,04	1,12	2,75	−0,09	−0,06	2,99	2,21	5,82	665,85
9	0,26	0,00	−0,03	0,98	2,40	0,00	−0,06	3,47	2,75	49,66	369,80
10	−0,80	−0,03	0,02	1,05	2,61	0,01	0,06	3,03	2,35	5,70	536,07
11	1,03	0,02	−0,04	1,10	2,72	−0,06	−0,08	2,97	2,31	7,71	603,50
12	0,01	−0,01	−0,02	1,08	2,72	0,00	−0,01	2,89	2,21	4,44	646,23
13	0,31	0,00	−0,01	1,11	2,79	−0,02	−0,02	2,73	2,16	8,24	714,20
14	0,13	0,00	−0,04	1,10	2,72	0,10	0,01	3,09	2,37	9,49	581,10
15	0,02	−0,02	−0,04	1,15	2,85	0,01	0,01	2,82	2,16	11,96	746,60
16	0,60	0,00	−0,05	1,12	2,80	−0,03	−0,02	2,84	2,18	7,63	706,04
17	0,65	−0,01	−0,07	1,09	2,71	0,03	−0,01	2,93	2,25	5,10	635,21
18	0,44	0,01	−0,02	1,02	2,53	−0,02	−0,04	3,21	2,53	17,98	446,48
19	0,12	0,01	0,00	1,10	2,79	0,00	−0,01	2,68	2,14	8,49	719,52
20	0,42	0,01	−0,04	1,08	2,67	0,04	−0,01	3,10	2,38	5,15	555,69
Média		0,00	−0,01	1,08	2,68	0,00	0,00	2,99	2,32	10,39	599,39
Desv. Pad.		0,01	0,06	0,04	0,11	0,04	0,04	0,18	0,15	9,85	96,61

Tabela A.15 – Resultados da simulação para $r^*_{q_i}$ e d_i, considerando o cenário VI com $n = 20$ - modelo de regressão *wrapped* Cauchy.

i	m_i	Média		Variância		Assimetria		Curtose		Estatística A-D	
		$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i	$r^*_{q_i}$	d_i
1	2,18	−0,02	−0,32	1,17	2,83	0,00	−0,05	2,68	2,02	17,32	971,39
2	−0,95	0,03	0,11	1,03	2,46	−0,01	0,08	3,31	2,64	33,13	417,74
3	−1,55	0,00	0,12	1,12	2,75	−0,01	0,08	2,79	2,20	7,95	682,57
4	1,05	−0,01	−0,10	1,02	2,52	0,05	−0,06	3,16	2,48	12,12	470,18
5	−1,77	0,01	0,18	1,13	2,75	−0,02	0,07	2,77	2,16	9,54	735,46
6	−0,93	0,00	0,08	1,07	2,59	−0,02	0,06	3,11	2,43	10,51	507,48
7	−0,45	0,01	0,04	1,01	2,42	0,01	0,01	3,33	2,67	26,46	376,70
8	1,37	0,00	−0,10	1,09	2,67	−0,10	−0,11	3,05	2,30	4,93	593,47
9	1,53	−0,01	−0,15	1,11	2,74	−0,01	−0,08	2,89	2,18	6,50	702,21
10	1,75	−0,02	−0,20	1,08	2,65	0,03	−0,08	2,93	2,19	4,23	682,81
11	0,32	0,02	0,01	1,10	2,71	−0,02	−0,04	2,89	2,28	6,16	609,40
12	−2,17	−0,01	0,25	1,11	2,68	−0,04	0,10	2,83	2,09	7,97	778,72
13	−1,21	0,00	0,09	1,10	2,70	−0,01	0,08	2,81	2,24	6,03	634,41
14	0,23	−0,01	−0,04	0,96	2,27	0,02	−0,02	3,88	3,03	77,72	293,94
15	−0,75	−0,03	0,01	1,07	2,54	0,02	0,07	3,27	2,56	19,93	445,22
16	−1,91	0,02	0,20	1,14	2,77	0,00	0,08	2,77	2,12	12,21	776,70
17	1,97	−0,02	−0,24	1,10	2,70	0,04	−0,07	2,82	2,13	6,63	773,43
18	1,17	0,00	−0,10	1,08	2,69	0,05	−0,03	2,90	2,28	4,84	614,43
19	−1,72	0,02	0,15	1,11	2,73	0,01	0,08	2,73	2,17	8,72	696,76
20	1,40	−0,01	−0,16	1,10	2,68	0,10	−0,03	2,99	2,30	4,90	634,55
Média		0,00	−0,01	1,08	2,64	0,00	0,01	3,00	2,32	14,39	619,88
Desv. Pad.		0,02	0,16	0,05	0,14	0,04	0,07	0,29	0,25	16,77	162,86


ufSCar