Universidade Estadual de Maringá
Departamento de Estatística
Programa de Pós-graduação em Bioestatística

Larissa Bueno Fernandes

Análises de Respostas Limitadas, Aplicações do Método PBC e Recordes

Maringá – PR
2018

Larissa Bueno Fernandes

Análises de Respostas Limitadas, Aplicações do Método PBC e Recordes

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá como requisito parcial para obtenção do título de Mestre em Bioestatística.

Orientador: Prof.º Dr.º Josmar Mazucheli

Universidade Estadual de Maringá - UEM

Departamento de Estística - DES

Programa de Pós-Graduação em Bioestatística

Maringá – PR

2018

Dados Internacionais de Catalogação na Publicação (CIP)
(Biblioteca Central - UEM, Maringá, PR, Brasil)

F363a

Fernandes, Larissa Bueno
Análises de respostas limitadas, aplicações do método PBC e Recordes / Larissa Bueno Fernandes. -- Maringá, 2018.
81 f. : figs., tabs.

Orientador: Prof. Dr. Josmar Mazucheli.
Dissertação (mestrado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Estística, Programa de Pós-Graduação em Bioestatística, 2018.

1. Regressão unit-Weibull. 2. Método Cross-Fitting. 3. Recordes. I. Mazucheli, Josmar, orient. II. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Estística. Programa de Pós-Graduação em Bioestatística. III. Título.

CDD 23.ed. 519.5

Glaucia Volponi de Souza – CRB9/948

LARISSA BUENO FERNANDES

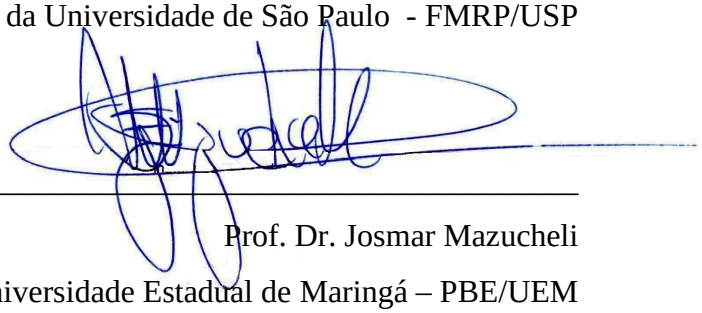
Análises de Respostas Limitadas, Aplicações do Método PBC e Recordes

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de Mestre em Bioestatística.

BANCA EXAMINADORA

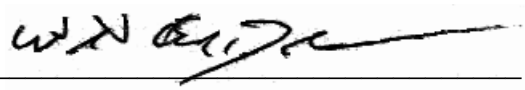
Prof. Dr. Jorge Alberto Achcar

Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo - FMRP/USP



Prof. Dr. Josmar Mazucheli

Universidade Estadual de Maringá – PBE/UEM



Prof. Dr. Emilio Augusto Coelho Barros

Universidade Tecnológica Federal do Paraná - UTFPR

Maringá, 24 de setembro de 2018.

Agradecimentos

Aos meus pais, Pedro e Rosimeri, que sempre me apoiaram em minhas escolhas e me acolheram em seus braços nos momentos em que os contratempos desta jornada me fizeram fraquejar. À minha irmã, Lays, e aos meus avós, Alcides e Genira, que juntamente aos meus pais me ofereceram todo o suporte que precisei. Palavras nunca serão suficientes para expressar toda a gratidão que sinto por vocês.

Ao meu orientador, professor Dr. Josmar Mazucheli, meu muito obrigada por compartilhar comigo seus vastos conhecimentos e por me guiar durante o desenvolvimento desta dissertação.

Agradeço também a todos os professores do PBE que contribuiram para minha formação, alguns dos quais me acompanham desde a graduação, construindo a base de todo o conhecimento que acumulei durante estes anos.

À minha amiga Luanna, que ao longo dos últimos anos se tornou uma irmã. Suas doses de humor e palavras de incentivo atuaram como injeções de ânimo para que eu chegasse até aqui. Às minhas queridas amigas Helena, Giovana e Aline, que mesmo distantes fisicamente, se fazem presentes por meio de suas palavras.

Por fim, não poderia deixar de agradecer ao Vinícius, meu colega de profissão e sócio, que embarcou comigo nessa jornada e a quem tenho admiração pela incansável busca por conhecimento.

"Life is a school of probability".
(Walter Bagehot)

Resumo

O presente trabalho permeia vários aspectos da modelagem estatística, passando pela proposta de um modelo de regressão quantílica para variáveis contínuas limitadas; a apresentação de um método recente de discriminação, que leva em conta além da qualidade do ajuste, a complexidade dos modelos; e a descrição e aplicação do esquema de recordes. O primeiro tema abordado refere-se a proposta de um novo modelo de regressão quantílica, reparametrizando a distribuição unit-Weibull, utilizando o método da máxima verossimilhança para a estimação de parâmetros. O potencial do novo modelo de regressão é demonstrado aplicando-o a três conjuntos de dados reais de diferentes áreas, comparado o ajuste obtido com aqueles obtidos pelos modelos de regressão de Kumaraswamy e Beta. Em um segundo momento, o método de discriminação PBC (*Parametric Bootstrap Cross-Fitting*), que mede o viés causado pelo mimetismo das distribuições candidatas, foi apresentado e aplicado aos conjuntos de dados referentes aos volumes de precipitações mensais da estação meteorológica convencional de Maringá - PR, observados entre 1964 e 2016, para a realização da discriminação e a comparação do mimetismo das distribuições Gama e Nakagami. Por fim, apresentou-se a caracterização frequentista da distribuição Gumbel baseada apenas nos valores de recorde, aplicando a metodologia aos mesmos conjuntos de dados de precipitações mensais avaliados no método PBC, objetivando a estimação dos parâmetros da distribuição em questão.

Palavras-chave: Regressão unit-Weibull, Método *Cross-Fitting*, Recordes.

Abstract

The present work permeates several aspects of statistical modeling, examining the proposal of a quantile regression model for limited continuous variables; and the presentation of a recent method of discrimination, which takes into account not only the quality of the adjustment, but also the models' complexity; and also the record scheme's description and application. The first topic we addressed is the proposal of a new quantile regression model, reparametrizing the unit-Weibull distribution, using the maximum likelihood method for parameter estimation. We demonstrated the potential of the new regression model by applying it to three real datasets from different areas, comparing the obtained fits with those provided by the Kumaraswamy and Beta's regression models. Secondly, we presented the PBC (Parametric Bootstrap Cross-Fitting) discrimination method, which measures the bias due the mimetism of the candidate distributions. Then, we applied it to the datasets referring to the monthly precipitation volumes of the meteorological station of Maringá - PR, observed between 1964 and 2016, in order to discriminate and compare the mimicry of the distributions Gama and Nakagami. Finally, we presented the frequentist characterization of Gumbel's distribution based only on record values, applying the methodology to the same datasets of monthly precipitation evaluated in the PBC method; with the purpose of estimating the distribution's parameters in question.

Keywords: unit-Weibull regression, *Cross-Fitting* method, Records.

Lista de ilustrações

Lista de tabelas

Sumário

Capítulo 1

Visão geral

1.1 Introdução

Como apresentado por Salsburg (2009) no clássico "Uma senhora toma chá", houve um tempo em que acreditava-se que a descrição - assim como a previsão - de todos os fenômenos poderia ser realizada por um conjunto de fórmulas matemáticas. Os desvios e as variações dos resultados destas fórmulas eram atribuídos apenas a erros de medição e acreditava-se que com o melhoramento dos métodos de medição, eles diminuiriam, mas, em vez disso, eles aumentaram. Assim, gradualmente a ciência começou a adotar um novo paradigma: o modelo estatístico.

De acordo com Bussab e Morettin (2011), quando se procede uma análise de dados, busca-se alguma regularidade ou algum modelo presente nas observações. Entretanto, além desta parte que pode ser explicada por um padrão, há também uma parte aleatória. A modelagem estatística agrega ambas as partes, não só reconhecendo a aleatoriedade inerente aos fenômenos, mas estimando o erro causado por ela. Atualmente, o número de modelos desenvolvidos é bastante amplo, assim como a quantidade de métodos de estimação de seus respectivos parâmetros e de métodos de discriminação, para a seleção do modelo mais adequado.

Neste contexto, o presente trabalho permeia vários aspectos da modelagem estatística, passando pela proposta de um modelo de regressão quantílica para variáveis

contínuas limitadas, a apresentação e a aplicação do esquema de recordes e de um método recente de discriminação, que leva em conta além da qualidade do ajuste, a complexidade dos modelos.

1.2 Objetivos

De acordo com cada um dos três temas levantados, o presente trabalho apresenta os seguintes objetivos.

- Introduzir um novo modelo de regressão considerando uma variável resposta no intervalo $(0, 1)$ com distribuição unit-Weibull, ajustando o modelo proposto a conjuntos de dados reias e comparando seus resultados com o modelo de regressão Beta e Kumaraswamy.
- Realizar a discriminação e a comparação do mimetismo das distribuições Gama e Nakagami, por meio da aplicação da variação do método PBC (*Parametric Bootstrap Cross-Fitting*), sendo tais distribuições ajustadas aos dados referentes as precipitações mensais observadas na estação meteorológica convencional de Maringá - PR.
- Caracterizar a distribuição Gumbel, com base nos valores de recorde superiores, aplicando a metodologia proposta para ajustar os recordes de precipitações mensais observadas na estação meteorológica convencional de Maringá - PR.

1.3 Organização do trabalho

No Capítulo 2, é apresentada a proposta de um modelo de regressão quantílica para variáveis respostas contínuas no intervalo limitado $(0, 1)$, com base na distribuição unit-Weibull, introduzida por Mazucheli, Menezes e Ghitany (2018). Este modelo pode ser considerado como alternativa ao modelo de regressão quantílica Kumaraswamy e ao modelo de regressão Beta. São apresentadas aplicações do modelo proposto a três conjuntos de dados reais, comparando seus resultados com outros dois modelos concorrentes - Beta e Kumaraswamy. Com base nos critérios de informação considerados,

o modelo unit-Weibull foi selecionado para o ajuste de duas das três aplicações realizadas. Apresenta-se no Apêndice A a versão em inglês, com algumas modificações, deste capítulo a qual está em processo de revisão na revista *Biometrical Journal*. Informo que por motivos de direitos autorais que este apêndice não constará na versão final desta dissertação.

Já no Capítulo 3 é apresentado o método PBC (*Parametric Bootstrap Cross-Fitting*), introduzido por Wagenmakers et al. (2004), que mede o viés causado pelo mimetismo entre duas distribuições candidatas, o que implica na consideração da avaliação da complexidade na seleção da distribuição de probabilidade. Com o intuito de explicar o comportamento dos volumes de precipitações mensais, a discriminação e a comparação do mimetismo das distribuições Gama e Nakagami foi realizada por meio da aplicação do método PBC aos dados de precipitações mensais da estação meteorológica convencional de Maringá - PR, observados entre 1964 e 2016. Na maior parte dos conjuntos de dados divididos por mês, a distribuição Nakagami se mostrou mais adequada, embora esta não seja uma distribuição muito utilizada em estudos envolvendo variáveis de natureza climatológica. Ainda, foi verificado que a Gama é funcionalmente mais complexa em relação à Nakagami, apresentando maior viés causado pelo mimetismo em todas as séries de precipitação mensal total. Apresenta-se no Apêndice B o artigo referente à este capítulo, publicado na revista Enciclopédia Biosfera.

Por fim, o Capítulo 4 apresenta uma descrição da metodologia de estimação baseada em valores de recorde, abordando a definição da técnica e revisão da literatura a respeito do tema, além dos resultados da aplicação a um conjunto de dados real, em que o objetivo é realizar a estimação dos parâmetros da distribuição Gumbel com base nos recordes superiores de precipitação mensal, considerando a estação meteorológica de Maringá - PR. Com os resultados, foi verificado que para alguns meses, o ajuste aos dados apenas com os valores de recorde foi satisfatório. Entretanto, para outros meses considerados, o ajuste pelos recordes se mostra bastante diferente da distribuição empírica dos volumes mensais de precipitação, assim como do ajuste por meio dos dados originais. Desta forma, outras abordagens ou distribuições serão consideradas no desenvolvimento futuro do trabalho.

Capítulo 2

Modelo de regressão unit-Weibull para variáveis resposta no intervalo unitário

Resumo

Embora a distribuição Beta seja a distribuição padrão para quantificar a influência de covariáveis na média de uma variável de resposta no intervalo unitário, ela não permite quantificar suas influências nos quantis da variável de resposta. Para esta proposta, Mitnik e Baek (2013) formularam um modelo de regressão quantílica reparametrizando a distribuição de Kumaraswamy. Na mesma direção, neste trabalho propõem-se um novo modelo de regressão quantílica, reparametrizando a distribuição unit-Weibull introduzida por Mazucheli, Menezes e Ghitany (2018), utilizando o método da máxima verossimilhança para a estimação de parâmetros. O potencial do novo modelo de regressão é demonstrado aplicando-o a três conjuntos de dados reais da contabilidade, da saúde e da climatologia. O ajuste da regressão unit-Weibull foi comparado com aqueles obtidos pelos modelos de regressão de Kumaraswamy e Beta. Para a questão da seleção do modelo, os critérios de informação Akaike, Bayesiano de Schwarz e Hannan-Quinn foram considerados juntamente com a estatística de Vu-

ong. Além disso, os gráficos *half-normal* com envelope simulado foram usados para diagnósticos do modelo.

Palavras-chave: Verossimilhança, distribuição unit-Weibull, regressão quantílica, variáveis limitadas.

Abstract

Although the Beta distribution is the standard distribution for quantifying the influences of covariates on the average of a response variable on the unit interval, it is no longer useful when we are interested in quantifying their influences on the quantiles of the response variable. For this propose, Mitnik e Baek (2013) have formulated a quantile regression model by reparameterizing the Kumaraswamy distribution. In the same direction, in this paper we purpose a new quantile regression model by re-parameterizing the unit-Weibull distribution introduced by Mazucheli, Menezes e Ghitany (2018), using the maximum likelihood method to estimate the parameters. The potential of the new regression model is demonstrated by applying it to three real datasets from accounting, health and climatology. We compare the obtained fits with those provided by the Kumaraswamy and Beta regression models. For the model selection issue, the Akaike's, the Schwarz's Bayesian and the Hannan-Quinn information criteria were considered along with the Vuong's statistic. In addition, the half-normal plots with a simulated envelope were used for model diagnostics.

Keywords: Likelihood, unit-Weibull distribution, quantile regression, restricted variables.

2.1 Introdução

A análise de regressão, termo introduzido por Francis Galton no século XIX (GALTON, 1886), caracteriza-se como uma técnica amplamente utilizada em várias áreas para investigar a relação de dependência de uma variável resposta com uma ou mais variáveis preditoras. A regressão busca descobrir quais preditores são importantes,

estimar o impacto da alteração do valor de uma variável preditora sobre o valor da variável resposta e também prever valores futuros (WEISBERG, 2005).

Durante muitos anos, os modelos normais lineares foram utilizados na tentativa de descrever a maioria dos fenômenos aleatórios (PAULA, 2004), até mesmo para os quais a suposição de normalidade da variável resposta não era razoável, sendo que a transformação dos dados era uma alternativa comum para alcançar a normalidade. Entretanto, a transformação altera a relação entre a variável resposta e as variáveis preditoras pela mudança de escala, comprometendo, muitas vezes, a interpretação dos parâmetros do modelo (FERRARI; CRIBARI-NETO, 2004).

Buscando unificar os procedimentos de inferência para outras opções de distribuição da variável resposta e outras funções para ligar os parâmetros das distribuições a um preditor linear, Nelder e Baker (1972) introduziram os modelos lineares generalizados (MLG). Apesar da grande flexibilidade, os modelos lineares generalizados usuais apresentam limitações para variáveis resposta restritas a um intervalo (a, b) (BONAT; JR; ZEVIANI, 2012). A observação de variáveis desta natureza é muito comum em diversos estudos, principalmente no intervalo $(0, 1)$, como taxas e proporções.

Este tipo de variável pode ser divida em quatro categorias de acordo com Kieschnick e McCullough (2003): a primeira categoria compreende as variáveis com domínio no intervalo aberto $(0, 1)$, que podem ser modeladas por uma distribuição contínua; a segunda categoria compreende as variáveis com domínio no intervalo fechado $[0, 1]$, modelados por uma mistura discreta-contínua; a terceira e quarta categorias são referentes às extensões multivariadas das duas primeiras. Apenas o primeiro caso será considerado nesse trabalho.

Uma extensão da teoria dos modelos lineares generalizados para variáveis resposta contínuas que assumem valores no intervalo $(0, 1)$ é a regressão Beta, proposta por Ferrari e Cribari-Neto (2004), sob a suposição de que a variável pode ser caracterizada por uma distribuição Beta. Nesse modelo, os parâmetros de regressão são interpretáveis em termos da média, sendo o modelo intrinsecamente heterocedástico acomodando assimetrias (CRIBARI-NETO; ZEILEIS, 2009).

Entretanto, apesar da flexibilidade da distribuição Beta para modelar dados no domínio limitado, foram propostas outras distribuições no intervalo unitário, como a

distribuição Johnson S_B (JOHNSON, 1949), a distribuição Topp-Leone (TOPP; LEONE, 1955), a distribuição unit-Gamma (GRASSIA, 1977; TADIKAMALLA, 1981), a distribuição Kumaraswamy (KUMARASWAMY, 1980), a distribuição unit-Logistic (TADIKAMALLA; JOHNSON, 1982), a distribuição Simplex (BARNDORFF-NIELSEN; JØRGENSEN, 1991), a distribuição Beta Retangular (HAHN, 2008), entre outras.

Em seu trabalho, Kieschnick e McCullough (2003) apresentam diferentes abordagens para modelagem de variáveis dependentes limitadas, identificando além dos modelos semi-paramétricos, estimados por quasi-verossimilhança (WOOLDRIDGE, 1997), seis modelos paramétricos de acordo com sua frequência de uso, que são: normal, logístico normal aditivo, normal censurado, normal não-linear, Beta (FERRARI; CRIBARI-NETO, 2004) e Simplex (SONG; TAN, 2000). Baseada nas comparações realizadas entre tais modelos, Kieschnick e McCullough (2003) recomendam a utilização do modelo de regressão Beta ou do modelo de quasi-verossimilhança.

Porém, desde então, vários avanços e alternativas a esses modelos foram propostas na literatura e grande parte dos modelos para respostas limitadas foram estendidos para explicar o comportamento da variável resposta na presença de variáveis preditoras. Pode-se citar, como exemplo, o modelo de regressão unit-Gamma (MOUSA; EL-SHEIKH; ABDEL-FATTAH, 2016), o modelo de regressão Kumaraswamy (MITNIK; BAEK, 2013) e o modelo de regressão Beta Retangular (BAYES et al., 2012), que acomodam a heterocedasticidade e consideram as distribuições unit-Gamma, Kumaraswamy e Beta Retangular para a variável resposta, respectivamente.

Recentemente, uma nova distribuição com suporte no $(0, 1)$, que decorre de uma transformação de uma variável aleatória com distribuição Weibull de dois parâmetros foi proposta por Mazucheli, Menezes e Ghitany (2018), sendo denominada unit-Weibull.

Neste sentido, o objetivo desse trabalho é introduzir um novo modelo de regressão considerando uma variável resposta no intervalo $(0, 1)$ com distribuição unit-Weibull, ajustando o modelo proposto a conjuntos de dados reais e comparando seus resultados com o modelo de regressão Beta e Kumaraswamy.

2.2 A distribuição unit-Weibull

Pode ser mostrado que a partir da transformação $Y = e^{-X}$, na qual X segue uma distribuição exponencial generalizada de dois parâmetros (GUPTA; KUNDU, 1999), obtem-se a distribuição de Kumaraswamy (KUMARASWAMY, 1980). Da mesma transformação e tomando X como uma variável aleatória com distribuição Weibull de dois parâmetros (WEIBULL, 1951) com função de densidade de probabilidade:

$$g(x \mid \alpha, \beta) = \alpha\, \beta\, x^{\beta-1}\, e^{-\alpha\, x^{\beta}}, \tag{2.1}$$

foi proposto por Mazucheli, Menezes e Ghitany (2018) a distribuição unit-Weibull (UW). Sua função de densidade de probabilidade (f.d.p.) e função de distribuição acumulada (f.d.a.) podem ser escritas na forma:

$$f(y \mid \alpha, \beta) = \frac{1}{y}\, \alpha\, \beta\, (-\log y)^{\beta-1}\, \exp\left[-\alpha\, (-\log y)^{\beta}\right], \tag{2.2}$$

e

$$F(y \mid \alpha, \beta) = \exp\left[-\alpha\, (-\log y)^{\beta}\right], \tag{2.3}$$

em que $0 < y < 1$ e $\alpha > 0$ e $\beta > 0$ são parâmetros de forma.

Embora não seja possível obter uma expressão analítica fechada para $E(Y^k)$, inviabilizando assim um modelo de regressão para a média, a função quantílica de uma variável aleatória com distribuição unit-Weibull é escrita na forma:

$$Q(\tau \mid \alpha, \beta) = \exp\left[-\left(-\frac{\log \tau}{\alpha}\right)^{\frac{1}{\beta}}\right]. \qquad 0 < \tau < 1, \tag{2.4}$$

Para introduzir um modelo de regressão quantílica, a equação (2.4) pode ser reparametrizada em termos do τ-ésimo percentil e de $\mu = Q(\tau)$, tal que α pode ser escrito da seguinte forma:

$$\alpha = -\frac{\log\, \tau}{(-\log \mu)^{\beta}}. \tag{2.5}$$

De acordo com essa parametrização, a f.d.p. e f.d.a da distribuição unit-Weibull podem ser expressas da seguinte forma:

$$f(y \mid \mu, \beta) = \frac{\beta}{y}\left(\frac{\log \tau}{\log \mu}\right)\left(\frac{\log y}{\log \mu}^{\beta-1}\right)\tau^{\left(\frac{\log y}{\log \mu}^{\beta}\right)}, \tag{2.6}$$

e

$$F(y \mid \mu, \beta) = \tau^{\left(\frac{\log y}{\log \mu}^{\beta}\right)} \tag{2.7}$$

Assim, a partir deste ponto será considerada a notação $Y \sim UW(\mu, \beta, \tau)$ para descrever uma variável aleatória Y que segue uma distribuição unit-Weibul com o parâmetro quantílico $\mu \in (0, 1)$, o parâmetro de forma $\beta > 0$ e $\tau \in (0, 1)$ conhecido.

A Figura 1 apresenta alguns dos possíveis comportamentos da f.d.p. da distribuição unit-Weibull para valores selecionados dos parâmetros θ e β, considerando $p = 0,5$, isto é, a mediana.

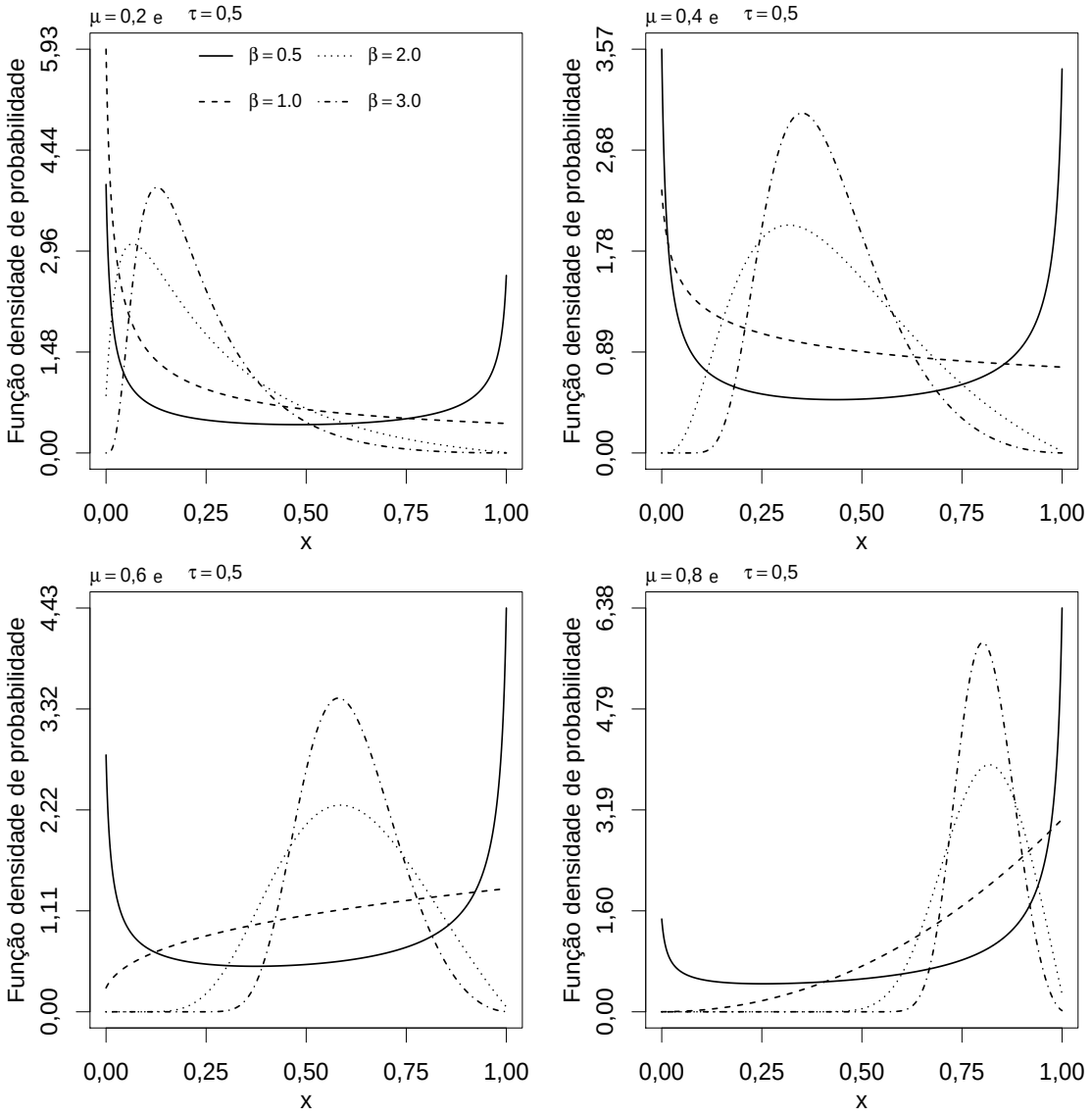


Figura 1 – Densidade reparametrizada da distribuição unit-Weibull para alguns valores de μ e β.

Nota-se na Figura 1 que a f.d.p. pode assumir diferentes formas de acordo com os valores de seus parâmetros: unimodal simétrica, unimodal assimétria à esquerda, unimodal assimétrica à direita, forma de banheira, exponencial crescente, exponencial decrescente e constante. Essa flexibilidade em sua forma faz com que a distribuição unit-Weibull torne-se uma boa alternativa para a análise de dados no intervalo unitário.

Além disso, como μ é o τ-ésimo percentil da distribuição de Y, pode-se interpretar que esse parâmetro é um parâmetro de locação no intervalo de valores da variável que está sendo modelada.

2.3 Modelo de regressão quantílica unit-Weibull

Considerando a f.d.p da distribuição unit-Weibull reparametrizada em (2.6), pode-se formular um modelo de regressão quantílica como apresentado em Mitnik e Baek (2013) e Santos e Bolfarine (2015), nos quais os autores consideraram as distribuições de Kumarasawamy e Laplace, respectivamente. Seja $Y = (Y_1, ..., Y_n)^T$ um vetor de n variáveis aleatórias, em que cada $Y_i, i = 1, ..., n$ segue a f.d.p especificada em (2.6), com parâmetro quantílico μ_i, parâmetro de forma desconhecido β e $\tau \in (0, 1)$ conhecido, isto é, $Y_i \sim UW(\mu_i, \beta, \tau)$. O modelo de regressão quantílica unit-Weibull é definido assumindo que o parâmetro quantílico de Y_i satisfaz a seguinte relação funcional:

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^T \delta, \tag{2.8}$$

em que $\delta = (\delta_1, ..., \delta_p)^T$ é um vetor de tamanho p $(p < n)$ dos coeficientes de regressão e $\boldsymbol{x}_i^T = (x_{i1}, ..., x_{ip})$ os valores observados de p variáveis independentes. A função $g(\cdot)$ é chamada de função de ligação, que relaciona o quantil μ_i com o vetor de covariáveis $\boldsymbol{x}_i^T$ e assume-se que ela é estritamente monótona e duplamente diferenciável em $(0, 1)$, sendo que $g : (0, 1) \to \mathbb{R}$.

Existem várias funções de ligação que satisfazem tais suposições, citando-se a função Logito, Probito e Log-Log Complementar. Tais funções baseiam-se na função de distribuição acumulada (f.d.a.) de determinadas distribuições, dadas pelas seguintes fórmulas:

Logito: baseada na f.d.a. da distribuição Logísitica, a função de ligação Logito é escrita na forma:

$$\text{Logito}(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right). \tag{2.9}$$

Probito: baseada na f.d.a. da distribuição Normal Padrão, a função de ligação Probito é escrita na forma:

$$\text{probito}(\mu_i) = \Phi^{-1}(\mu_i). \tag{2.10}$$

em que $\Phi(.)$ é a função de distribuição acumulada de uma variável aleatória normal padrão;

Log-Log Complementar: baseada na f.d.a. da distribuição do Valor Extremo, a função de ligação Log-Log Complementar é escrita na forma:

$$\text{clog-log}(\mu_i) = \log[-\log(1-\mu_i)]. \tag{2.11}$$

Para uma discussão das funções de ligação, pode-se consultar McCullagh e Nelder (1989). Em um primeiro momento, a função de ligação Logito será considerada nesse trabalho. Logo, tem-se que:

$$\begin{aligned} \log\left(\frac{\mu_i}{1-\mu_i}\right) &= \boldsymbol{x}_i^T\delta \\ \frac{\mu_i}{1-\mu_i} &= \exp\left(\boldsymbol{x}_i^T\delta\right) \\ \mu_i &= \frac{\exp\left(\boldsymbol{x}_i^T\delta\right)}{1+\exp\left(\boldsymbol{x}_i^T\delta\right)}. \end{aligned} \tag{2.12}$$

2.3.1 Estimação

Uma vez que δ representa o efeito das variáveis explicativas sobre o τ-ésimo percentil da variável resposta, têm-se interesse em estimar esses parâmetros. Sob uma abordagem clássica, o vetor de parâmetros desconhecido δ são estimados maximizando a função de máxima verossimilhança, que pode ser expressa como:

$$\ell(\delta,\beta \mid \boldsymbol{y},\boldsymbol{x},\tau) \propto n\log\,\beta + \beta\sum_{i=1}^{n}\log\left(\frac{\log\,y_i}{\log\,\mu_i}\right) + \log\,\tau\sum_{i=1}^{n}\log\left(\frac{\log\,y_i}{\log\,\mu_i}\right)^{\beta}, \tag{2.13}$$

em que $\mu_i = \frac{\exp\left(\boldsymbol{x}_i^T\delta\right)}{1+\exp\left(\boldsymbol{x}_i^T\delta\right)}$, assumindo a função de ligação Logito.

Não é possível derivar uma solução analítica para as estimativas de máxima verossimlhança (EMV) dos parâmetros δ e β, uma vez que essas são obtidas pela solução do sistema de equações não-lineares $U_\delta(\delta, \beta) = 0$ e $U_\beta(\delta, \beta) = 0$, que não apresentam uma solução analítica fechada. Desta forma, faz-se necessária a utilização de maximização numérica do logaritmo da função de verossimilhança, por meio de algum algoritmo de otimização como o de Newton-Raphson, quasi-Newton ou escore de Fisher. Sugere-se utilizar como um palpite inicial para δ as estimativas de mínimos quadrados ordinários deste vetor de parâmetros obtidas a partir da regressão linear das respostas transformadas $g(y_1), ..., g(y_n)$ em X, isto é, $(X^TX)^{-1}X^Tz$, em que $z = (g(y_1), ..., g(y_n))^T$.

É bem conhecido que sob condições de regularidade moderada (ver, por exemplo, Lehmann e Casella (1998)) e quando n é grande, a distribuição assintótica dos estimadores de máxima verossimilhança é tal que:

$$\begin{pmatrix} \widehat{\boldsymbol{\delta}} \\ \widehat{\beta} \end{pmatrix} \xrightarrow{D} N_{kp+1}\left[\begin{pmatrix} \boldsymbol{\delta} \\ \beta \end{pmatrix}, K^{-1}(\boldsymbol{\delta}, \beta)\right], \tag{2.14}$$

em que $\xrightarrow{D}$ denota convergência em distribuição e $K^{-1}(\boldsymbol{\delta}, \beta)$ é a inversa da matriz de informação de Fisher esperada, sendo que não há expressão fechada para a matriz $K(\boldsymbol{\delta}, \beta)$. No entanto, como mostrado por Lindsay e Li (1997), a informação de Fisher observada é um estimador consistente da matriz de informação de Fisher esperada. Portanto o comportamento assintótico permanece se $K(\boldsymbol{\delta}, \beta) = \lim_{n\to\infty} n^{-1}J(\boldsymbol{\delta}, \beta)$, em que $J(\boldsymbol{\delta}, \beta)$ denota a matriz de informação de Fisher observada.

Seja δ_r o *r*-ésimo componente de $\boldsymbol{\delta}$, então o $100 \times (1 - \gamma/2)\%$ intervalo de confiança assintótico para δ_r é dado por:

$$\widehat{\delta}_r \pm \Phi^{-1}(1 - \gamma/2)\,\mathrm{se}\left(\widehat{\delta}_r\right), \qquad r = 1, \ldots, p \tag{2.15}$$

em que $\Phi^{-1}(\cdot)$ é a função quantílica da distribuição normal padrão e $\mathrm{se}\left(\widehat{\delta}_r\right)$ é o erro padrão assintótico de $\widehat{\delta}_r$, que é obtido da raiz quadrada de (*r*-*r*)-ésimo elemento de $\mathbf{K}^{-1}(\boldsymbol{\delta}, \beta)$.

2.3.2 Adequação do modelo

De acordo com Paula (2004), a análise dos resíduos é constituída por um conjunto de técnicas para avaliar se a distribuição em estudo é apropriada para descrever o comportamento da variável resposta e ainda identificar a presença de possíveis pontos extremos no conjunto de dados. Nesse trabalho foram considerados os resíduos propostos por Cox e Snell (1968). Os resíduos Cox-Snell, são dados por:

$$r_{C_i} = -log\, \widehat{S}(y_i \mid \widehat{\Theta}), \qquad i = 1, \ldots, n, \tag{2.16}$$

em que $\widehat{S}(.)$ é a estimativa da função de sobrevivência baseada na estimativa de máxima verossimilhança $\widehat{\Theta}$.

Os resíduos Cox-Snell possuem a propriedade principal de que, se o modelo se ajusta aos dados, $r_i, i = 1, ..., n$ segue a distribuição exponencial padrão, com f.d.p. $f(r) = e^r$. O gráfico de r_i *versus* $-\log \widehat{S}(r_i)$ deve ser uma linha reta com inclinação unitária e intercepto zero se o ajuste do modelo é adequado, sendo $\widehat{S}(r_i)$ o estimador de *Kaplan-Meier* de $S(r_i)$. Para mais detalhes veja, por exemplo, Lee e Wang (2003, p. 215) ou Lawless (2003).

Para os resíduos Cox-Snell, o gráfico *half-normal* com envelopes simulados são uma ferramenta de diagnóstico útil (ZHAO et al., 2011). Veja, por exemplo, Atkinson (1985), Collet (2003) e Kutner et al. (2005), para mais detalhes sobre os gráficos *half-normal*.

2.4 Aplicações

Nesta seção, apresenta-se três aplicações reais para mostrar a potencialidade do modelo de regressão proposto. Para fins de comparação, além do modelo de regressão quantílica unit-Weibull, também foram considerados dois modelos de regressão alternativos comumente usados na análise de variáveis limitadas, que são resumidamente descritos a seguir:

Beta: modelo de regressão Beta introduzido por Cepeda-Cuervo (2001) e Ferrari e

Cribari-Neto (2004) tem f.d.p. dado por:

$$f(y \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, y^{\alpha-1}\,(1-y)^{\beta-1}, \tag{2.17}$$

em que $\alpha, \beta > 0$.

Kumaraswamy: modelo de regressão Kumaraswamy introduzido por Mitnik e Baek (2013) tem f.d.p. dado por:

$$f(y \mid \alpha, \beta) = \alpha\,\beta\, y^{\alpha-1}\,(1-y^{\alpha})^{\beta-1}, \tag{2.18}$$

em que $\alpha, \beta > 0$.

Para discriminar e escolher o melhor entre os modelos propostos, os critérios de informação de Akaike (AIC), Schwarz (SBC) e Hannan-Quinn (HQIC) foram considerados. Seja L_{fit} a verossimilhança do modelo ajustado, p o número de parâmetros no modelo e n o número de observações. Os critérios são dados por:

- O critério de informação de Akaike (AIC), proposto por Akaike (1974), é dado por $AIC = -2\,\log(L_{fit}) + 2\,p$;
- O critério Bayesiano de Schwarz (SBC) ou critério de informação Bayesiano (BIC), proposto por Schwarz et al. (1978), é dado por $SBC = -2\,\log(L_{fit}) + p\,\log(n)$;
- O critério de informação de Hannan-Quinn (HQIC), proposto por Hannan e Quinn (1979), é dado por $HQIC = -2\,\log\,L_{fit} + 2\,p\,\log\log(n)$.

A regra de decisão, em todos esses critérios, é favorável ao modelo com o menor valor das estatísticas (HELD; BOVÉ, 2014).

Ainda, utilizou-se o teste de proximidade de *Vuong* (VUONG, 1989), para avaliar se existe alguma diferença significativa quanto ao ajuste dos dois modelos alternativos em relação ao unit-Weibull. Esse teste, adequado para comparar modelos não encaixados (*not nested*), tem por hipótese nula que não existem diferenças significativas entre o ajuste dos modelos.

Sendo F_θ e G_γ dois modelos não encaixados, com respectivas densidades dadas por $f(y_i \mid x_i, \theta)$ e $g(y_i \mid x_i, \gamma)$, avaliadas nos EMV's, a estatística da razão das verossimilhanças para comparar ambos os modelos é dada por:

$$LR(\widehat{\theta}, \widehat{\gamma}) = \ell_f(\widehat{\theta}) - \ell_g(\widehat{\gamma}) = \sum_{i=1}^{n} \log \frac{f(y_i \mid x_i, \widehat{\theta})}{g(y_i \mid x_i, \widehat{\gamma})}. \tag{2.19}$$

Entretanto a estatística apresentada em (2.19) não segue uma distribuição qui-quadrado. Para contornar o problema, Vuong (1989) propôs uma abordagem alternativa baseada no critério de informação de Kullback-Liebler (KULLBACK; LEIBLER, 1951). Considerando a distância entre cada modelo e o processo verdadeiro que gera os dados, denominado $h_0(y_i, X_i)$, têm-se a estatística:

$$T_{LR,NN} = \frac{1}{\sqrt{n}} \frac{LR(\widehat{\theta}, \widehat{\gamma})}{\widehat{\omega}^2}, \tag{2.20}$$

em que $\widehat{\omega}^2$ é um estimador da variância de $\frac{1}{\sqrt{n}} LR(\widehat{\theta}, \widehat{\gamma})$, dado por:

$$\widehat{\omega}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\log \frac{f(y_i \mid x_i, \theta)}{g(y_i \mid x_i, \gamma)} \right)^2 - \left(\frac{1}{n} \sum_{i=1}^{n} \left(\log \frac{f(y_i \mid x_i, \theta)}{g(y_i \mid x_i, \gamma)} \right) \right)^2, \tag{2.21}$$

Quando $n \to \infty$, $T_{LR,NN} \xrightarrow{d} N(0,1)$ sob a hipótese nula. Portanto, a um nível de significância de $\alpha\%$, rejeita-se hipótese nula de equivalência das distribuição se $|T| < z_{\alpha/2}$.

As estimativas da máxima verossimilhança foram obtidas pelo algoritmo quasi-Newton, disponível no procedimento SAS/NLMIXED (SAS, 2010). O erros padrões assintóticos e intervalos de confiança foram calculados por meio da matriz de informação observada de Fisher.

2.4.1 Rentabilidade do gerenciamento de risco

O primeiro conjunto de dados considerado é apresentado por Schmit e Roth (1990) e corresponde às 73 respostas a um questionário enviado a 374 gerentes de risco de grandes organizações norte americanas. O objetivo do estudo de Schmit e

Roth (1990) foi avaliar a relação custo-eficácia com a filosofia da administração de controlar a exposição da empresa a várias perdas de propriedade e acidentes, levando em consideração características da empresa, como tamanho e tipo de indústria. A descrição das variáveis e suas codificações são apresentadas a seguir:

Firmcost: variável contínua restrita ao intervalo $(0, 1)$ referente à rentabilidade do gerenciamento de risco da empresa;

Assume: variável contínua referente à quantidade de retenção por ocorrência como percentual dos ativos totais;

Cap: variável dummy que indica se a empresa possui uma companhia de seguros cativa (1) ou não (0);

Sizelog: variável contínua referente ao logaritmo dos ativos totais;

Indcost: variável contínua referente ao risco da indústria;

Central: variável contínua referente à importância dos gerentes locais na escolha da quantidade de risco a ser mantida;

Soph: variável contínua referente ao grau de importância no uso de ferramentas analíticas.

A seguinte estrutura de regressão foi empregada:

$$\begin{aligned} \text{Logito}(\mu_i) &= \log\left(\frac{\mu_i}{1-\mu_i}\right) \\ &= \delta_0 + \delta_1\,\text{Assume}_i + \delta_2\,\text{Cap}_i + \delta_3\,\text{Sizelog}_i + \delta_4\,\text{Indcost}_i + \\ &\quad \delta_5\,\text{Central}_i + \delta_6\,\text{Soph}_i, \quad i = 1, \ldots, 73, \end{aligned} \tag{2.22}$$

em que μ_i denota a mediana nos modelos unit-Weibull e Kumaraswamy, já no modelo Beta μ_i denota a média.

Na Tabela 1, apresenta-se as estimativas dos parâmetros e seus respectivos erros padrões (E.P.) para os três modelos considerados no ajuste aos dados de rentabilidade do gerenciamento de risco de grandes organizações norte americanas.

Tabela 1 – Estimativas dos parâmetros e erros padrões para os três modelos considerados no ajuste aos dados de rentabilidade do gerenciamento de risco.

	unit-Weibull		Kumaraswamy		Beta	
Parâmetro	Est.	IC (95%)	Est.	IC (95%)	Est.	IC (95%)
δ_0	3,471	(1,289; 5,654)	2,539	(-0,500; 5,577)	1,888	(-0,410; 4,186)
δ_1	-0,008	(-0,033; 0,018)	-0,036	(-0,071; -0,002)	-0,012	(-0,039; 0,015)
δ_2	0,128	(-0,364; 0,619)	0,596	(-0,169; 1,362)	0,178	(-0,276; 0,632)
δ_3	-0,804	(-1,045; -0,564)	-0,798	(-1,114; -0,482)	-0,512	(-0,752; -0,271)
δ_4	1,439	(0,635; 2,244)	5,257	(2,443; 8,071)	1,236	(0,336; 2,137)
δ_5	-0,024	(-0,191; 0,143)	-0,028	(-0,262; 0,207)	-0,012	(-0,184; 0,159)
δ_6	-0,002	(-0,045; 0,041)	-0,027	(-0,090; 0,035)	-0,004	(-0,046; 0,038)
β	3,353	(2,728; 3,979)	0,978	(0,771; 1,186)	6,331	(4,130; 8,531)

* Est.: Estimativa; IC: Intervalo de Confiança.

A Tabela 2 apresenta os valores das estatísticas usadas como critérios de seleção dos modelos unit-Weibull, Kumaraswamy e Beta ajustados ao conjunto de dados em questão. Observa-se que os três critérios de informação avaliados indicam que o modelo de regressão unit-Weibull apresentou um melhor ajuste quando comparado aos modelos concorrentes, seguido pelo modelo Kumaraswamy, cuja regressão também é aplicada a mediana. Corroborando com os critérios de informação, a hipótese de que não existem diferenças significativas do ajuste dos modelos, em relação ao unit-Weibull, foi rejeitada ao nível de significâncias de 5%, favorecendo a escolha do mesmo.

Tabela 2 – Medidas utilizadas para discriminação entre os três modelos considerados no ajuste aos dados de rentabilidade do gerenciamento de risco.

Critério	Modelo		
	unit-Weibull	Kumaraswamy	Beta
AIC	-206,2227	-181,6534	-159,4460
SBC	-187,8990	-163,3297	-141,1223
HQIC	-198,9204	-174,3511	-152,1437
Voung	—	2,1513 (0,0157)	4,5817 (0,0000)

Pela Figura 2, que apresenta os gráficos *half-normal* dos resíduos de Cox-Snell para os três modelos concorrentes, vê-se que os ajustes foram satisfatórios, sobretudo

para o modelo unit-Weibull no qual as probabilidades empíricas se aproximam das probabilidades teóricas.

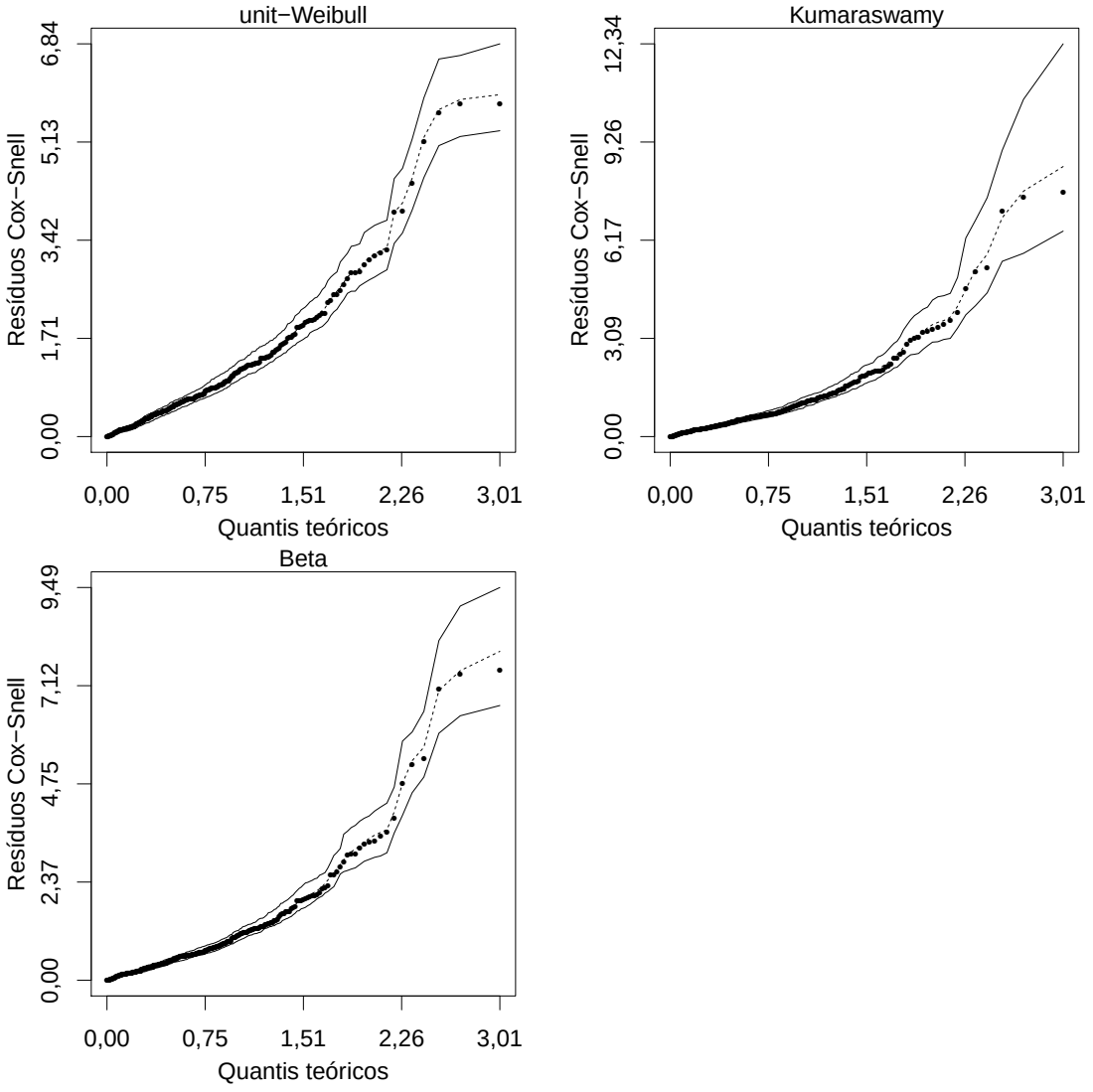


Figura 2 – Gráficos *half-normal* com envelopes simulados dos resíduos de Cox-Snell para os três modelos considerados no ajuste aos dados de rentabilidade do gerenciamento de risco.

O impacto de diferentes valores de τ nas estimativas dos parâmetros $\delta_i, i = 0, ..., 6$ são ilustrados na Figura 3, na qual observa-se variações tanto no valor da estimativa pontual, quanto a precisão das estimativas, uma vez que o tamanho do

intervalo de confiança varia de acordo com o percentil determinado. Independentemente do valor de τ fixado, a estimativa do parâmetro β e seu respectivo intervalo de confiança permanecem iguais, desta forma o mesmo não é apresentado a seguir.

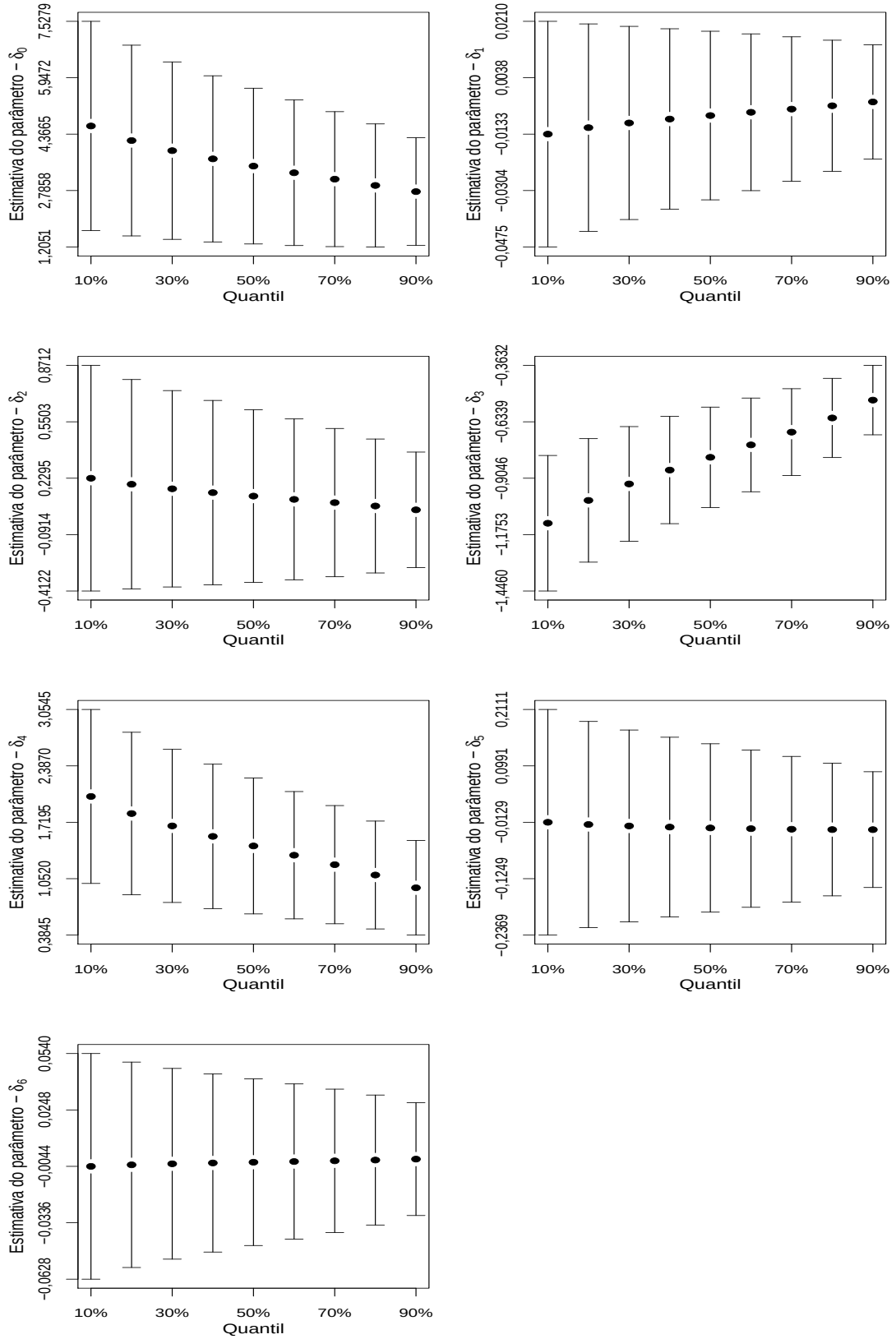


Figura 3 – Estimativas dos parâmetros e intervalos de 95% de confiança para o modelo UW e $\tau = 0.1, 0.2, \ldots, 0.8$ and 0.9 para o dados de rentabilidade do gerenciamento de risco.

2.4.2 Taxa de recuperação de células CD34+

O segundo conjunto de dados considerado corresponde a informação de 249 pacientes que concordaram com o transplante autólogo de células-tronco do sangue periférico (PBSC) após doses mieloablativas de quimioterapia entre o ano de 2003 e 2008 (ZHANG; QIU; SHI, 2016).

Uma descrição das variáveis e suas codificações são apresentadas a seguir:

CD34+: variável contínua restrita ao intervalo $(0, 1)$ referente à taxa de recuperação das células CD34+;

Sexo: variável dummy indicando se um paciente é do sexo feminino (0) ou masculino (1);

Quimio: variável dummy indicando se um paciente recebe uma quimioterapia em um protocolo de um dia (0) ou em um protocolo de 3 dias (1);

Idade: variável contínua referente à idade ajustada, que é a idade atual menos 20.

A seguinte estrutura de regressão foi empregada:

$$\begin{aligned} \text{Logito}(\mu_i) &= \log\left(\frac{\mu_i}{1-\mu_i}\right) \\ &= \delta_0 + \delta_1\,\text{Sexo}_i + \delta_2\,\text{Quimio}_i + \delta_3\,\text{Idade}_i, \quad i = 1, \ldots, 249, \end{aligned} \tag{2.23}$$

em que μ_i denota a mediana nos modelos unit-Weibull e Kumaraswamy, já no modelo Beta μ_i denota a média.

As estimativas e erros padrões dos parâmetros dos três modelos considerados no ajuste aos dados referentes a taxa de recuperação das células CD34+, de pacientes pacientes que concordaram com o transplante autólogo de células-tronco do sangue periférico, são apresentados na Tabela 3.

Tabela 3 – Estimativas dos parâmetros e erros padrões para os três modelos considerados no ajuste aos dados de taxa de recuperação de células CD34+.

	unit-Weibull		Kumaraswamy		Beta	
Parâmetro	Est.	IC (95%)	Est.	IC (95%)	Est.	IC (95%).
δ_0	0,962	(0,703; 1,221)	1,200	(0,926; 1,474)	0,999	(0,746; 1,252)
δ_1	0,017	(0,008; 0,027)	0,011	(-0,001; 0,022)	0,014	(0,004; 0,025)
δ_2	0,282	(0,089; 0,474)	0,183	(-0,042; 0,409)	0,212	(0,008; 0,415)
δ_3	0,103	(-0,082; 0,288)	0,042	(-0,145; 0,229)	0,066	(-0,118; 0,250)
β	1,680	(1,517; 1,843)	6,727	(5,837; 7,618)	11,345	(9,349; 13,340)

* Est.: Estimativa; IC: Intervalo de Confiança.

Na Tabela 4 apresenta-se a comparação do ajuste dos três modelos propostos por meio dos valores das estatísticas usadas como critérios de seleção. Assim, como para o conjunto de dados referente à rentabilidade do gerenciamento de risco, nota-se que para esta aplicação os três critérios de informação avaliados indicam que o modelo de regressão unit-Weibull apresentou um melhor ajuste quando comparado aos modelos concorrentes. Considerando um nível de 5% de significância, os resultados do teste de *Vuong* apontam que não há evidências amostrais suficientes de que os modelos Beta e unit-Weibull diferem significativamente, embora o ajuste do unit-Weibull tenha se mostrado superior a todos os demais.

Tabela 4 – Medidas utilizadas para discriminação entre os três modelos considerados no ajuste aos dados de taxa de recuperação de células CD34+.

Critério	Modelo		
	unit-Weibull	Kumaraswamy	Beta
AIC	-388,0932	-375,6599	-381,7912
SBC	-370,7109	-358,2775	-364,4089
HQIC	-381,0886	-368,6553	-374,7866
Voung	—	1,7117 (0,0435)	1,0590 (0,1448)

Para avaliar se os modelos são apropriados, na Figura 4 apresenta-se os gráficos *half-normal* com envelopes simulados para os resíduos de Cox-Snell. A Figura 4, indica um bom ajuste do modelo de regressão unit-Weibull para a taxa de recuperação de células CD34+.

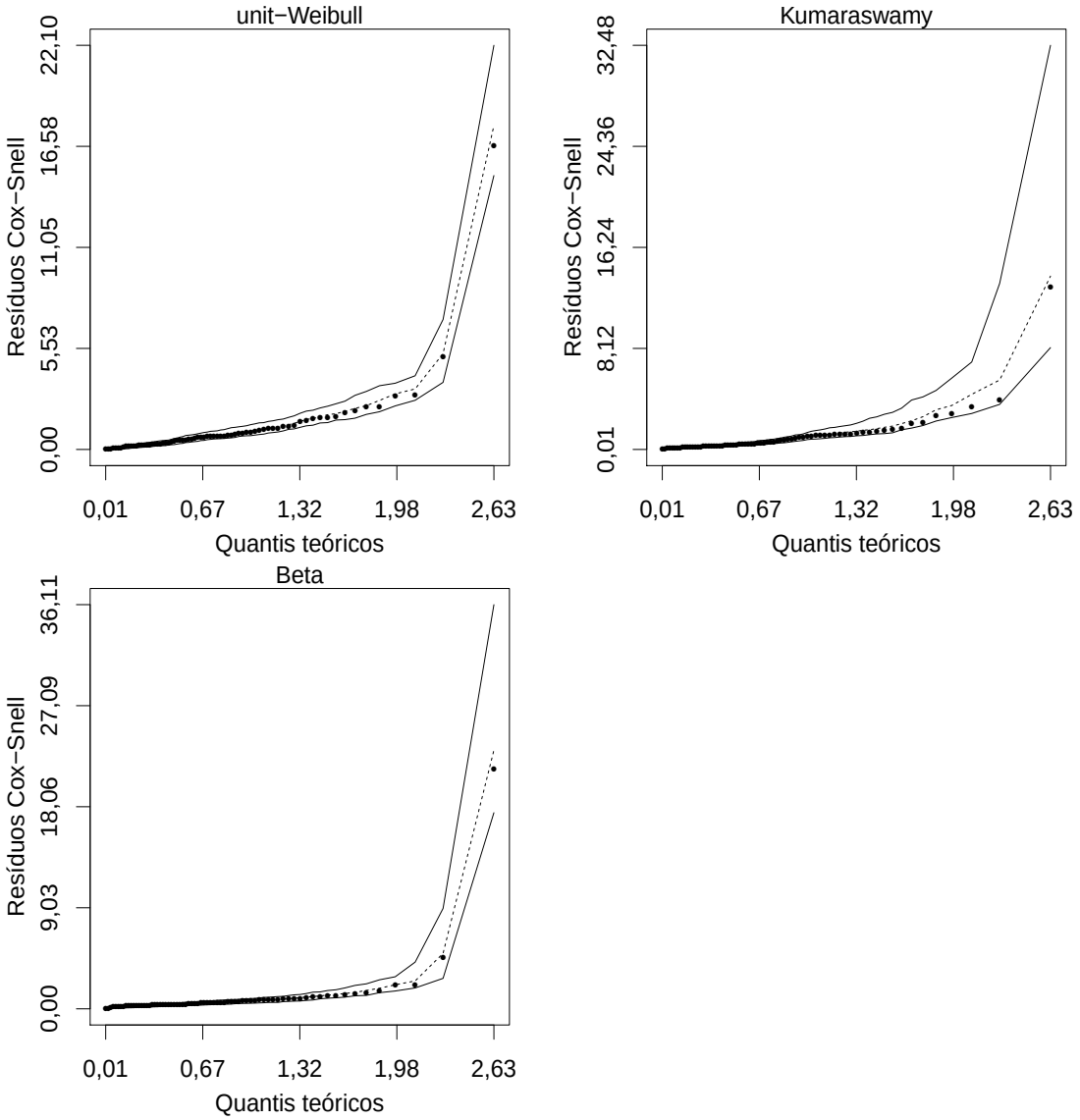


Figura 4 – Gráficos *half-normal* com envelopes simulados dos resíduos de Cox-Snell para os três modelos considerados no ajuste aos dados de taxa de recuperação de células CD34+.

O impacto de diferentes valores de τ nas estimativas dos parâmetros $\delta_i, i = 0, ..., 6$ são ilustrados na Figura 5, na qual verifica-se que o maior impacto da definição do percentil se deu no intercepto do modelo.

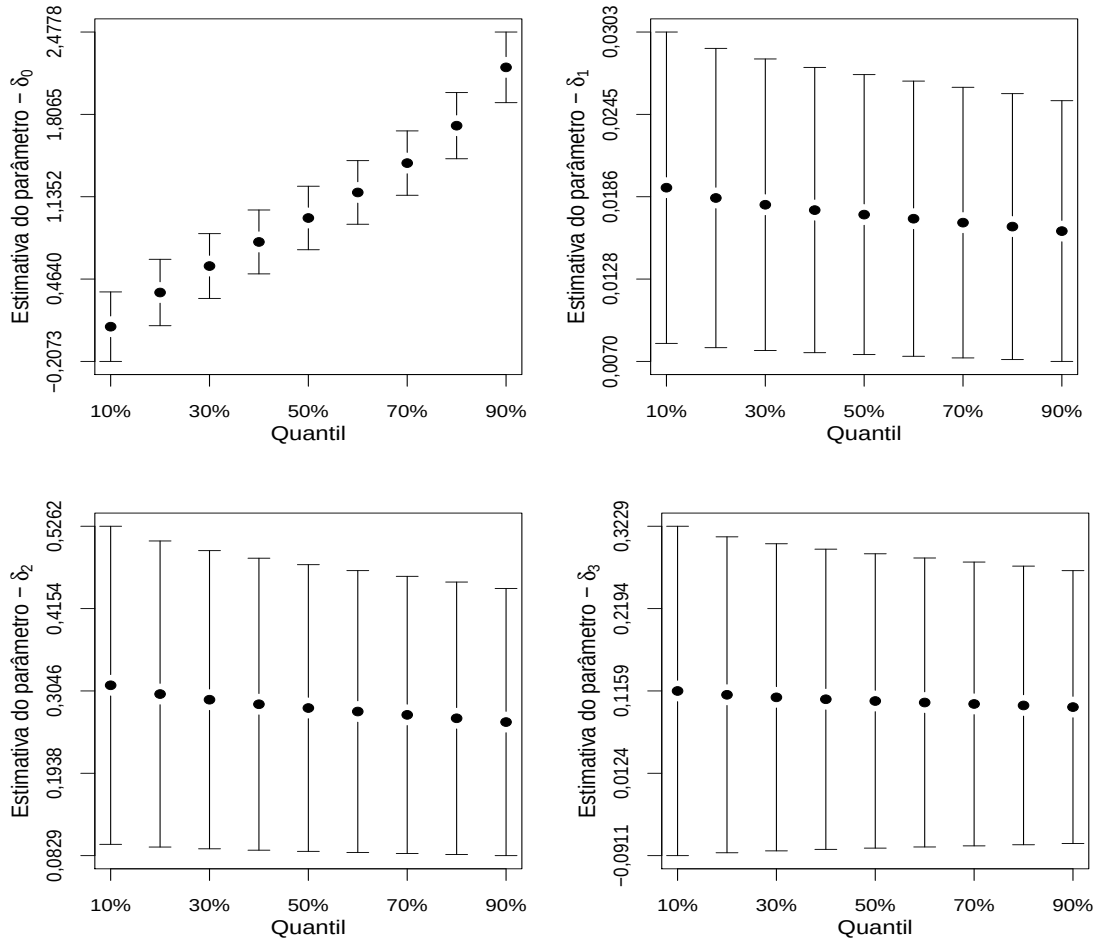


Figura 5 – Estimativas dos parâmetros e intervalos de 95% de confiança para o modelo UW e $\tau = 0.1, 0.2, \ldots, 0.8$ and 0.9 para os dados de taxa de recuperação de células CD34+.

2.4.2.1 Umidade relativa média

O terceiro conjunto de dados corresponde à 511 observações de umidade relativa média mensal da estação metereológica convencional localizada em Maringá - PR, entre os anos de 1961 e 2016. Os dados utilizados foram compilados a partir das séries históricas de precipitações mensais obtidas no Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) do Instituto Nacional de Meteorologia (INMET).

Uma descrição das variáveis e suas codificações são apresentadas a seguir:

Umidade: variável contínua restrita ao intervalo $(0, 1)$ referente à umidade relativa média mensal;

Jan: variável dummy indicando se o mês é Janeiro (1) ou não (0);

Fev: variável dummy indicando se o mês é Fevereiro (1) ou não (0);

Mar: variável dummy indicando se o mês é Março (1) ou não (0);

Abr: variável dummy indicando se o mês é Abril (1) ou não (0);

Mai: variável dummy indicando se o mês é Maio (1) ou não (0);

Jun: variável dummy indicando se o mês é Junho (1) ou não (0);

Jul: variável dummy indicando se o mês é Julho (1) ou não (0);

Ago: variável dummy indicando se o mês é Agosto (1) ou não (0);

Set: variável dummy indicando se o mês é Setembro (1) ou não (0);

Out: variável dummy indicando se o mês é Outubro (1) ou não (0);

Nov: variável dummy indicando se o mês é Novembro (1) ou não (0).

A seguinte estrutura de regressão foi empregada:

$$\begin{aligned} \text{Logito}(\mu_i) &= \log\left(\frac{\mu_i}{1-\mu_i}\right) \\ &= \delta_0 + \delta_1\,\text{Jan}_i + \delta_2\,\text{Fev}_i + \delta_3\,\text{Mar}_i + \delta_4\,\text{Abr}_i + \delta_5\,\text{Mai}_i + \delta_6\,\text{Jun}_i + \\ &\quad \delta_7\,\text{Jul}_i + \delta_8\,\text{Ago}_i + \delta_9\,\text{Set}_i + \delta_{10}\,\text{Out}_i + \delta_{11}\,\text{Nov}_i, \quad i = 1, \ldots, 511, \end{aligned} \tag{2.24}$$

em que μ_i denota a mediana nos modelos unit-Weibull e Kumaraswamy, já no modelo Beta μ_i denota a média. Assim, o mês de dezembro é tomado como base de comparação para a interpretação dos coeficientes de regressão.

A Tabela 5 apresenta as estimativas e erros padrões dos parâmetros dos três modelos concorrentes no ajuste da umidade relativa média mensal da estação metereológica de Maringá - Paraná.

Tabela 5 – Estimativas dos parâmetros e erros padrões para os três modelos considerados no ajuste aos dados de umidade relativa média, da estação meteorológica convencional de Maringá – PR, de 1964 a 2016.

	unit-Weibull		Kumaraswamy		Beta	
Parâmetro	Est.	IC (95%)	Est.	IC (95%)	Est.	IC (95%)
δ_0	0,908	(0,831; 0,984)	0,938	(0,851; 1,025)	0,927	(0,842; 1,012)
δ_1	0,153	(0,048; 0,258)	0,143	(0,015; 0,271)	0,162	(0,040; 0,284)
δ_2	0,154	(0,049; 0,260)	0,128	(0,001; 0,256)	0,168	(0,046; 0,290)
δ_3	0,026	(-0,081; 0,132)	-0,054	(-0,174; 0,065)	-0,017	(-0,137; 0,102)
δ_4	-0,059	(-0,167; 0,049)	-0,075	(-0,194; 0,045)	-0,062	(-0,182; 0,058)
δ_5	-0,066	(-0,176; 0,044)	0,084	(-0,045; 0,212)	0,043	(0,080; 0,166)
δ_6	0,078	(-0,030. 0,187)	0,066	(-0,061; 0,194)	0,067	(-0,057; 0,019)
δ_7	-0,210	(-0,321; -0,098)	-0,206	(-0,323; -0,088)	-0,230	(-0,351; -0,110)
δ_8	-0,536	(-0,650; -0,422)	-0,452	(-0,561; -0,343)	-0,536	(-0,654; -0,419)
δ_9	-0,443	(-0,556; -0,331)	-0,285	(-0,398; -0,172)	-0,402	(-0,520; -0,284)
δ_{10}	-0,237	(-0,348; -0,126)	-0,266	(-0,381; -0,152)	-0,272	(-0,391; -0,152)
δ_{11}	-0,240	(-0,350; -0,129)	-0,195	(-0,311; -0,078)	-0,257	(-0,376; -0,138)
β	4,578	(4,281; 4,875)	11,964	(11,150; 12,777)	57,118	(50,167; 64,070)

* Est.: Estimativa; IC: Intervalo de Confiança.

Os resultados dos critérios de informação calculados a partir do ajuste dos modelos de regressão unit-Weibull, Kumaraswamy e Beta são apresentados na Tabela 6. Ao contrário do observado nas demais aplicações realizadas nesse trabalho, a distribuição Beta foi indicada como mais adequada para o ajuste aos dados sob os três critérios de informação considerados, seguida pela distribuição unit-Weibull. Entretanto os dados amostrais não apresentam evidências suficientes de que o ajuste do modelo unit-Weibull difere significativamente do ajuste do modelo Beta, ao nível de 5% de significância.

Tabela 6 – Medidas utilizadas para discriminação entre os três modelos considerados no ajuste aos dados de umidade relativa média, da estação meteorológica convencional de Maringá – PR, de 1964 a 2016.

Critério	Modelo		
	unit-Weibull	Kumaraswamy	Beta
AIC	-1390,6670	-1337,9757	-1402,0236
BIC	-1335,5942	-1282,9029	-1346,9508
HQIC	-1369,0766	-1316,3853	-1380,4332
Voung	—	2,1382 (0,0163)	-0,7312 (0,7677)

A partir dos gráficos *half-normal* com envelopes simulados mostrados na Figura 6, pode-se concluir que o modelo de regressão unit-Weibull teve um ajuste satisfatório.

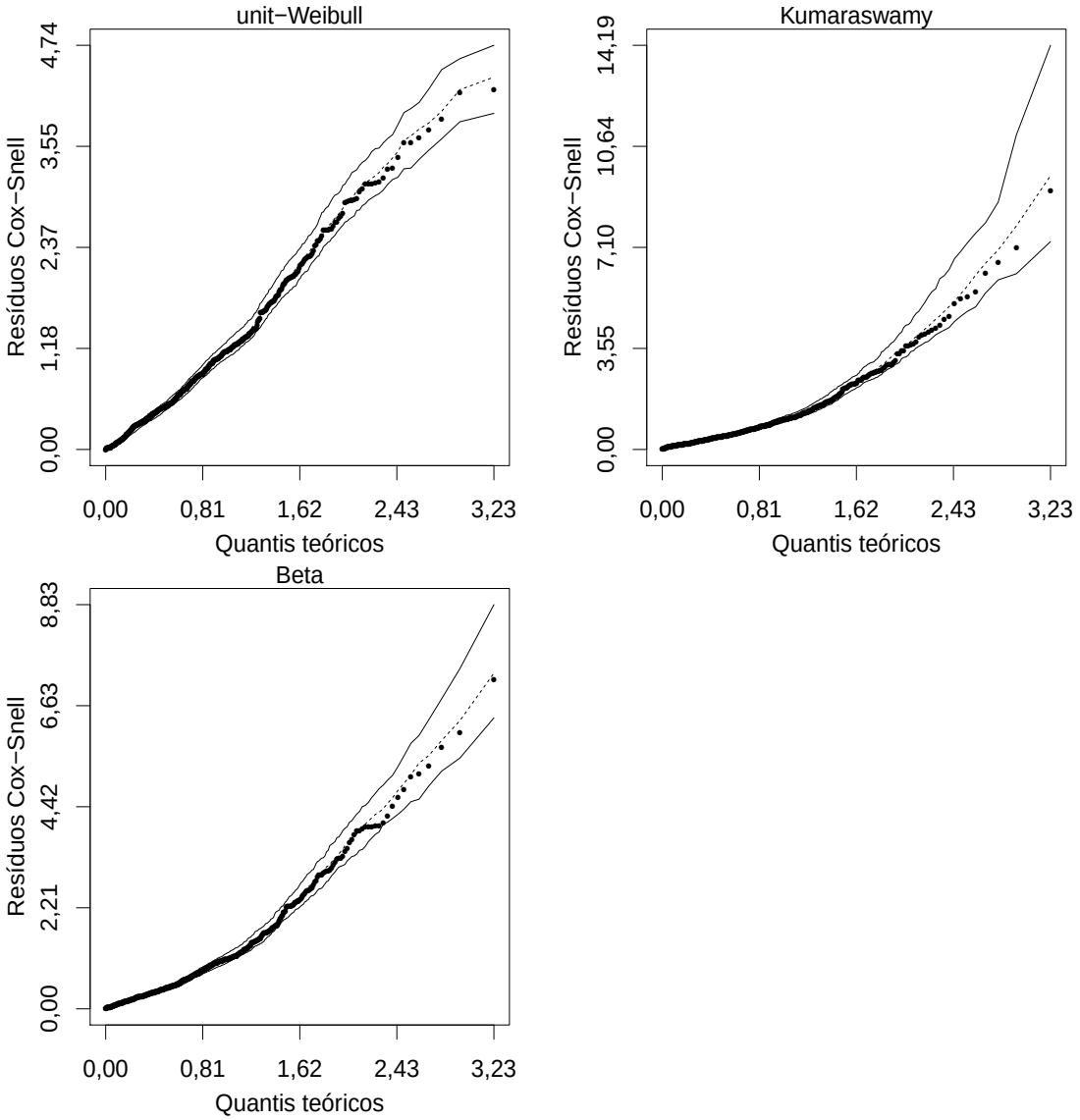


Figura 6 – Gráficos *half-normal* com envelopes simulados dos resíduos de Cox-Snell para os três modelos considerados no ajuste aos dados de umidade relativa média, da estação meteorológica convencional de Maringá – PR, de 1964 a 2016.

Por fim, o impacto de diferentes valores de τ nas estimativas dos parâmetros $\delta_i, i = 0, ..., 6$ são ilustrados na Figura 7, na qual nota-se que o impacto ocorre sobretudo no intercepto dos modelos ajustados.

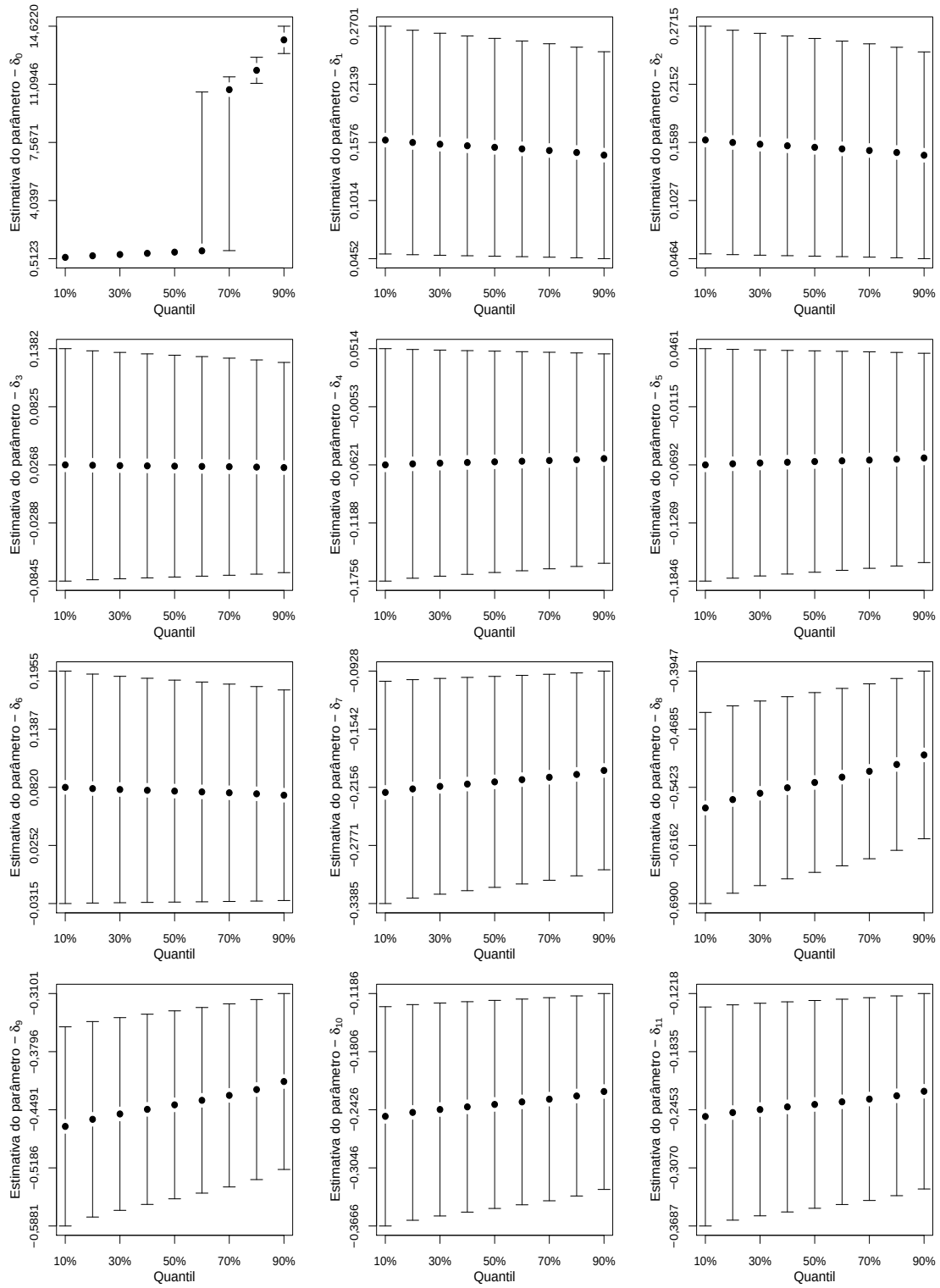


Figura 7 – Estimativas dos parâmetros e intervalos de 95% de confiança para o modelo UW e $\tau = 0.1, 0.2, \ldots, 0.8$ and 0.9 para o dados de dados de umidade relativa média, da estação meteorológica convencional de Maringá – PR, de 1964 a 2016.

2.5 Considerações finais

Nesse trabalho, foi proposto um novo modelo de regressão quantílica, considerando uma variável resposta no intervalo $(0, 1)$ que segue uma distribuição unit-Weibull, introduzida recentemente por Mazucheli, Menezes e Ghitany (2018). Para isso, a distribução unit-Weibull foi reparametrizada em termos do τ-ésimo percentil $\mu = Q(\tau)$, sendo que a regressão foi proposta para este novo parâmetro, permitindo ligar qualquer percentil da distribuição às covariáveis.

Uma vez que a f.d.p. da unit-Weibull pode assumir diferentes formas de acordo com os valores de seus parâmetros, essa distribuição torna-se uma boa alternativa para o ajuste de variáveis resposta no intervalo unitário.

Foram analisados também três conjuntos de dados reais de diferentes áreas para fins ilustrativos. Para dois dos conjuntos de dados, que referem-se a rentabilidade do gerenciamento de risco e à taxa de recuperação de células CD34+, o modelo de regressão unit-Weibull foi selecianada entre outros dois modelos concorrentes, de acordo com os três critérios de informação avaliados. Já para o conjunto de dados referente à umidade relativa mensal de Maringá - PR, o modelo de regressão selecionado sob os critérios utilizados foi o modelo Beta, embora o ajuste desses não tenha sido significativamente diferente do ajuste do modelo unit-Weibull.

Por fim, uma extensão futura desse trabalho poderia ser dedicada à elaboração e à aplicação de um modelo de mistura entre as distribuições binomial e unit-Weibull, a fim de acomodar valores no intervalo fechado $[0, 1]$, muito comum em problemas práticos envolvendo respostas contínuas limitadas, o que não é contemplado pelo modelo apresentado nesse trabalho. Outra proposta a ser considerada é a inclusão de efeitos aleatórios no modelo.

Capítulo 3

Aplicações do método PBC para avaliação da complexidade das distribuições Gama e Nakagami na análise de dados de precipitação

Resumo

Em geral, a seleção da distribuição de probabilidade com o melhor ajuste não considera a complexidade das distribuições rivais. O método PBC (*Parametric Bootstrap Cross-Fitting*) mede o viés causado pelo mimetismo das distribuições candidatas, levando em consideração a avaliação da complexidade das mesmas. Considerando a importância de explicar o comportamento dos volumes de precipitações mensais e da possibilidade do uso do método PBC para comparação do ajuste de duas distribuições concorrentes, objetivou-se neste trabalho realizar a discriminação e a comparação do mimetismo das distribuições Gama e Nakagami, aplicando o método PBC aos dados de precipitações mensais da estação meteorológica convencional de Maringá - PR, obser-

vados entre 1964 e 2016. Embora Nakagami não seja uma distribuição muito utilizada em estudos envolvendo variáveis de natureza climatológica, os resultados referentes a aplicação do método PBC indicam que para a maior parte das séries históricas mensais, a distribuição mais apropriada para a descrição do comportamento da precipitação mensal é a Nakagami quando comparada a Gama. Ainda, foi verificado que a Gama é funcionalmente mais complexa em relação a Nakagami, apresentando maior viés causado pelo mimetismo em todas as séries de precipitação mensal total.

Palavras-chave: Climatologia; Distribuição de Probabilidade; Método *Cross-Fitting*.

Abstract

In general, the selection of the probability distribution with the best fit does not take into account the distributions' complexity. The PBC (Parametric Bootstrap Cross-Fitting) method quantifies the bias due to the mimicry of the candidate distributions, taking into account the evaluation of the models' complexity. Considering the importance of explaining the behavior of the monthly rainfall volumes and the possibility of using the PBC method to compare the fit of two competing distributions, the objective of this work is to discriminate and compare the mimicry of the Gamma and Nakagami distributions, by applying the PBC method to the monthly precipitation data of the Maringá - PR meteorological station, observed between 1964 and 2016. Although Nakagami is not a very usual distribution in studies involving climatological variables, results referring to the PBC method indicated that for most of the monthly historical series, Nakagami is the most appropriate distribution for the description of monthly precipitation behavior when compared to Gamma. In addition, it was observed that the Gamma distribution was shown to be functionally more complex than the Nakagami, presenting greater measures of bias due to mimicry for all datasets.

Keywords: Climatology; Cross-Fitting Method; Probability Distribution.

3.1 Introdução

A caracterização da distribuição da precipitação constitui um importante instrumento de apoio para estudos e desenvolvimento de atividades econômicas relacionadas. O estudo da distribuição de variáveis climáticas determina padrões de ocorrência permitindo uma previsão razoável do comportamento climático pluvial de uma região (MARTIN et al., 2015). Tanto a intensidade quanto a frequência com que os fenômenos climáticos ocorrem influenciam diretamente na sustentabilidade ambiental e econômica dos setores do agronegócio (BEYRUTH, 2008).

No estudo de variáveis climatológicos (volume de precipitação, velocidade do vento, temperatura, entre outras), em geral, existe o interesse na estimação da probabilidade de uma variável aleatória X exceder um valor x_T qualquer, ou seja $P(X \geq x_T)$. Em climatologia, a estimação desta quantidade é o foco principal do que é conhecido como análise de frequências. O objetivo da análise de frequências é relacionar a magnitude de eventos extremos com a frequência de ocorrência por intermédio de uma distribuição de probabilidade apropriada (CHOW; MAIDMENT; MAYS, 2013).

Na literatura, várias distribuições de probabilidade são usadas no estudo de variáveis climatológicas. Dentre estas citam-se: Gama, Weibull, Log-Normal, Log-Logística, Inversa-Gaussiana e Valor-Extremo (COE; STERN, 1982). Dado um conjunto de observações, a seleção de uma distribuição é, sem dúvida, o passo mais crítico, e aquele que apresenta as maiores incertezas (MEYLAN; FAVRE; MUSY, 2011).

Outra distribuição que pode ser aplicada para dados climatológicos é a Nakagami cujas aplicações concentram-se, principalmente, na área de engenharia de comunicações (NAKAGAMI, 1960). Se comparada com outras distribuições, a Nakagami é considerada genérica, de grande flexibilidade e simplicidade matemática. O número de aplicações na área climatológica é ainda pequeno e pode-se citar os trabalhos recentes dos autores Schwartz, Godwin e Giles (2013), Singh e Sarkar (2013) e Mazucheli e Emanuelli (2015), sendo que apenas neste último a distribuição foi ajustada a dados referentes ao volume de precipitação.

Em geral, várias distribuições concorrentes podem ser propostas para a explicação de um mesmo fenômeno. A escolha da distribuição mais adequada deve-se pautar na comparação de medidas que avaliam a qualidade do ajuste, que são denominadas de

goodness-of-fit (GOF), como o valor da estatística do teste de *Kolmogorov-Smirnov* (KS). Estas medidas quantificam o quão bem a distribuição ajusta-se a um conjunto de observações, e podem ser usadas para comparar o ajuste de distribuições concorrentes (SCHULTHEIS; NAIDU, 2014). No entanto, existe uma tendência para a seleção de distribuições complexas, mesmo que distribuições mais simples sejam mais parcimoniosas (PITT; MYUNG, 2002).

Para contornar este problema surgiu o conceito de mimetismo, o qual quantifica o quão bem os modelos são capazes de imitar uns aos outros, ou seja, a habilidade de cada distribuição fornecer bons ajustes aos dados obtidos pela outra distribuição (SCHULTHEIS; NAIDU, 2014). Para quantificar o mimetismo, existe o procedimento denominado de PBC (*Parametric Bootstrap Cross-Fitting*) que gera duas distribuições a partir das subtrações de uma especificada medida de GOF, esperadas sob cada uma das distribuições concorrentes, sendo tais distribuições obtidas por meio do método *Bootstrap* , introduzido por Efron (1992). Entretanto, ao se realizar uma busca na literatura, não foram encontradas aplicações do método PBC para a discriminação entre distribuições usadas no estudo de variáveis climatológicas.

No trabalho em que propõe o método PBC, Wagenmakers et al. (2004) apresentam duas versões do mesmo, que diferem-se quanto a forma de gerar os dados, sendo a primeira indicada para avaliar o ajuste e mimetismo da distribuição para um banco de dados especificado e a segunda indicada para extrair conclusões mais genéricas sobre o ajuste das distribuições em questão.

Conforme Wagenmakers et al. (2004), dado o valor de alguma estatística de GOF (como a KS), a versão do PBC que depende de um determinado conjunto de dados é aplicado adotando os passos a seguir:

1. Gerar uma amostra *Bootstrap* não paramétrica, denotada por x^* dos dados originais x;
2. Ajustar as distribuições A e B a amostra x^*, obtendo os EMV dos vetores de parâmetros $\widehat{\theta}_A^*$ e $\widehat{\theta}_B^*$, respectivamente;
3. Simular dados segundo a distribuição A ($D(\widehat{\theta}_A^*)$) e a distribuição B ($D(\widehat{\theta}_B^*)$), por meio do *Bootstrap* paramétrico;

4. Ajustar as distribuições A e B a $D(\widehat{\theta}^*_A)$, e obter a diferença de GOF ($\Delta GOF^*_{AB} \mid A = GOF^*_A - GOF^*_B$) entre os dados simulados pela distribuição A;

5. Ajustar as distribuições A e B a $D(\widehat{\theta}^*_B)$, e obter a diferença de GOF ($\Delta GOF^*_{AB} \mid B = GOF^*_A - GOF^*_B$) entre os dados simulados pela distribuição B;

6. Repetir os passos anteriores M vezes.

Desta forma, a aplicação do PBC resulta em um vetor de diferenças de GOF sob a distribuição A e uma sob a distribuição B. Schultheis e Singhaniya (2013) apontam que a utilidade do PBC depende essencialmente de uma utilização adequada das distribuições de diferença de GOF (denotadas por ΔGOF^*_{AB}) geradas para selecionar entre os modelos concorrentes.

A quantidade $\Delta GOF^*_{AB} \mid A$ permite medir o quão provável é obter a diferença GOF observada (δ_{AB}) ajustando as distribuições aos dados observados se a variável de interesse segue a distribuição A. A quantidade $\Delta GOF^*_{AB} \mid B$ permite medir o quão provável é obter se a distribuição B é o modelo de geração (SCHULTHEIS; NAIDU, 2014). Assim, a probabilidade de que a variável de interesse segue a distribuição A, em vez da distribuição B, dada a diferença observada de GOF, pode ser quantificada pela razão entre as estimativas da altura das distribuições $\Delta GOF^*_{AB} \mid A$ e $\Delta GOF^*_{AB} \mid B$ no valor δ_{AB}:

$$P(\delta_{AB} \mid A)/P(\delta_{AB} \mid B) = P_A(x)/P_B(x)$$

.

De acordo com Wagenmakers et al. (2004), o critério para decisão da escolha entre os modelos denominado ótimo, definido como o critério que maximiza a probabilidade de uma classificação binária correta, é calculado por $P_A(x)/P_B(x) = 1$. O viés causado pelo mimetismo (β_m) é quantificado pela subtração entre o critério nominal, $\Delta GOF_{AB} = 0$ e o ótimo (WAGENMAKERS et al., 2004).

Considerando a importância de explicar o comportamento dos volumes de precipitação mensal; da possibilidade do uso do método PBC para comparação do ajuste de duas distribuições concorrentes para explicar a complexidade das distribuições, e do método PBC não ter sido utilizado em modelos de variáveis climatológicas de acordo com a revisão da literatura, objetivou-se neste trabalho realizar a discriminação e a

comparação do mimetismo das distribuições Gama e a Nakagami, por meio da aplicação da variação do PBC ajustadas aos dados referentes as precipitações mensais observadas na estação meteorológica convencional de Maringá - PR.

3.2 Materiais e métodos

3.2.1 Dados

Foram utilizados dados da estação meteorológica convencional da cidade de Maringá - PR (Figura 8), compilados a partir das séries históricas de precipitação mensais obtidas no Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) do Instituto Nacional de Meteorologia (INMET) órgão responsável pela coleta e disponibilização informações meteorológicas oficiais.

Figura 8 – Localização da estação meteorológica convencional de Maringá - PR. Fonte: Google Maps (2017).

Considerou-se como variável de estudo o total de precipitação acumulada no mês, a fim de realizar a comparação entre duas distribuições propostas (Gama e Nakagami) para a explicação de tal fenômeno. Nos dados obtido do INMET encontram-se disponíveis, de forma digital, apenas os registros mensais a partir do ano de 1961. Entretanto, foram selecionadas as séries históricas disponíveis no período entre janeiro

de 1964 e dezembro de 2016, já que entre os anos de 1961 e 1963 os registros da estação convencional de Maringá foram realizados apenas para alguns meses.

As informações referentes aos volumes de precipitação mensais não estavam disponíveis para todos os meses do período considerado, sendo que no total, foram utilizadas 527 observações, divididas em 12 séries mensais para o ajuste das distribuições.

3.2.2 Distribuições

Para análise da pluviosidade total mensal, utilizou-se a distribuição Gama, comum no estudo de variáveis de precipitação, e a Nakagami. A seguir descreve-se as distribuições objeto de estudo do presente trabalho.

3.2.2.1 Distribuição Nakagami

Uma variável aleatória não negativa X com distribuição Nakagami tem, respectivamente, função de densidade e de distribuição escritas nas formas:

$$f\left(x \mid \mathbf{\Theta}\right) = \frac{2}{\Gamma\left(\alpha\right)} \left(\frac{\alpha}{\theta}\right)^{\alpha} x^{2\,\alpha-1} \exp\left[-\frac{\alpha}{\theta} x^2\right] \tag{3.1}$$

e

$$F\left(x \mid \mathbf{\Theta}\right) = \Gamma\left(\frac{\alpha}{\theta} x^2, \alpha, 1\right), \tag{3.2}$$

em que $\mathbf{\Theta} = (\theta, \alpha)$ é o vetor de parâmetros, $\theta > 0$ o parâmetro de escala, $\alpha \geq 0.5$ o parâmetro de forma, $\Gamma\left(\cdot\right)$ a função gama e $\Gamma\left(\cdot, a, b\right)$ é a função de distribuição acumulada de uma variável aleatória com distribuição Gama com parâmetro de forma a e escala b. Para $\alpha = 1$, temos a distribuição Half-Normal e, para $\alpha = 0.5$, a distribuição Rayleigh. A partir do método de transformação de variáveis aleatórias mostra-se facilmente que se $X \sim Nakagami(\theta, \alpha)$ então $Y = X^2$ tem distribuição Gama com parâmetro escala $\frac{\theta}{\alpha}$ e forma α.

3.2.2.2 Distribuição Gama

Uma variável aleatória não negativa X com distribuição Gama tem, respectivamente, função de densidade e de distribuição escritas nas formas:

$$f\left(x \mid \mathbf{\Theta}\right) = \frac{1}{\theta^{\alpha}\Gamma\left(\alpha\right)} x^{\alpha-1} \exp\left(-\frac{x}{\theta}\right), \tag{3.3}$$

e

$$F\left(x \mid \Theta\right) = \int_0^x \frac{1}{\theta^{\alpha}\Gamma\left(\alpha\right)} x^{\alpha-1} \exp\left(-\frac{x}{\theta}\right), \tag{3.4}$$

em que $\Theta = (\theta, \alpha)$ é o vetor de parâmetros, $\theta > 0$ o parâmetro de escala, $\alpha > 0$ o parâmetro de forma, $\Gamma\left(\cdot\right)$ a função gama. Para $\alpha = 1$, temos a distribuição Exponencial e, para $\alpha = n/2$ e $\theta = 2$, a distribuição qui-quadrado com n graus de liberdade.

3.2.3 Aplicação do método PBC

Para realizar a aplicação do método PBC, utilizou-se as séries históricas de precipitação mensais da estação convencional de Maringá. Foram realizadas $M = 1000$ iterações, sendo assim, as densidades empíricas das distribuições de ΔGOF^*_{AB} foram estimadas a partir de 1000 valores. Utilizou-se a estatística KS como medida de GOF, dada por:

$$KS = \max_{1 \leq i \leq n} \left(\widehat{z}_i - \frac{i}{n}, \frac{i}{n} - \widehat{z}_i\right) \text{ para } i = 1, \ldots, n$$

.

Para descrever dados de precipitações, foram selecionadas as distribuições Gama e Nakagami, comparadas por meio do método PBC. É importante enfatizar que a Gama é a distribuição usada pelo INMET para a modelagem da precipitação. O algoritmo para aplicação do método de discriminação foi implementado no ambiente estatístico R (R Core Team, 2016), assim como todas as análises realizadas.

3.3 Resultados e discussão

Os resultados apresentados na Tabela 7 mostram que os maiores volumes são observados entre os meses de dezembro e fevereiro, chegando a uma precipitação média de 204 mm no mês de janeiro. Em contrapartida, os meses de inverno (junho a agosto) foram aqueles em que os menores volumes de precipitação média foram observados, e para tais meses a precipitação mínima foi de 0 mm, indicando que houveram anos em que não ocorreu precipitação nos meses em questão.

Tabela 7 – Medidas resumo das séries históricas de precipitação agrupadas por mês, da estação meteorológica convencional de Maringá – PR, de 1964 a 2016.

Mês	n	Média	DP	CV	Mediana	Mínimo	Máximo
Janeiro	45	204,01	91,83	45%	205,10	27,60	419,80
Fevereiro	45	191,30	99,66	52%	168,50	46,20	426,00
Março	45	152,12	74,05	49%	130,50	33,60	340,00
Abril	44	115,35	63,82	55%	104,75	1,00	346,30
Maio	43	115,99	93,56	81%	89,70	0,70	396,40
Junho	43	105,21	93,09	88%	92,80	0,00	396,70
Julho	43	70,53	69,83	99%	52,10	0,00	378,60
Agosto	42	53,91	52,14	97%	38,95	0,00	219,80
Setembro	45	119,01	79,41	67%	94,00	21,40	319,60
Outubro	43	164,32	82,83	50%	142,10	45,00	345,60
Novembro	44	138,17	83,31	60%	114,45	26,20	369,60
Dezembro	45	176,64	72,28	41%	176,80	45,00	360,40

* n: número de observações; DP: Desvio padrão; CV: Coeficiente de variação.

Observa-se uma grande variabilidade na distribuição dos volumes de precipitação, com coeficientes de variação chegando a 99% para o mês de julho. Também se nota que em geral, as distribuições dos volumes mensais de chuva (Figura 9) apresentam assimetria à direita, com medianas inferiores as respectivas médias de precipitação. Também nota-se a presença de alguns valores atípicos, sobretudo para os meses entre abril e julho.

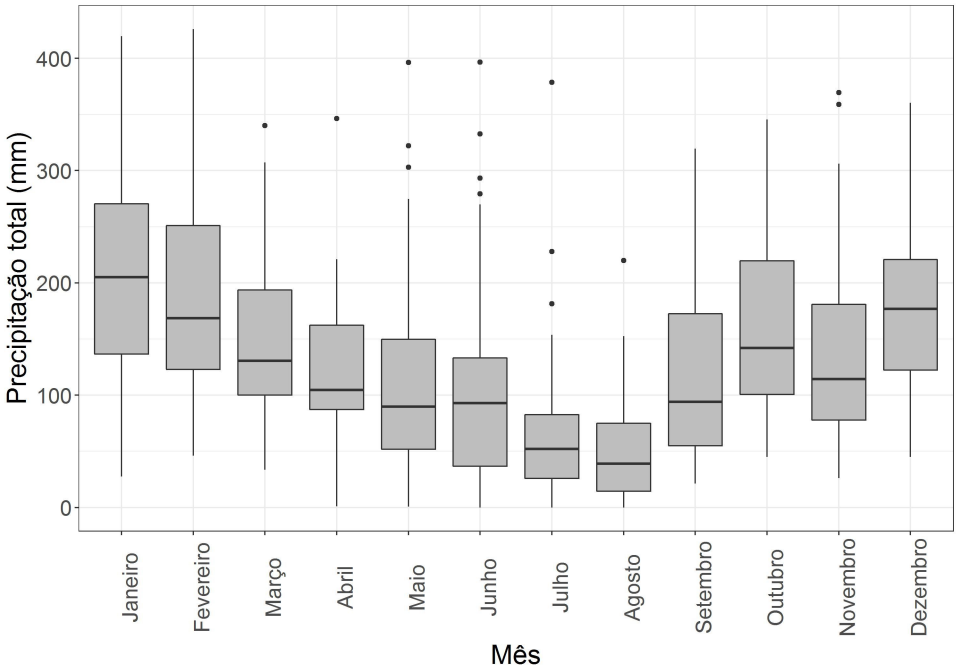


Figura 9 – Box-plot das séries históricas de precipitação agrupadas por mês, da estação meteorológica convencional de Maringá – PR, de 1964 a 2016.

O algoritmo do método PBC, foi aplicado as séries históricas de volumes de precipitação de cada mês da estação convencional de Maringá no período considerado, resultando em 12 conjuntos de dados. Observa-se na Figura 10 que na maior parte dos meses observados, a distribuição Gama foi mais flexível para o ajuste aos dados obtidos pelo modelo concorrente, apresentando maior mimetismo. Também observa-se que para três quartos dos meses observados, o PBC indicou que a distribuição Nakagami se ajusta melhor aos dados, e além disso, nos meses de Junho e Julho, a análise isolada do KS apontou uma conclusão contrária.

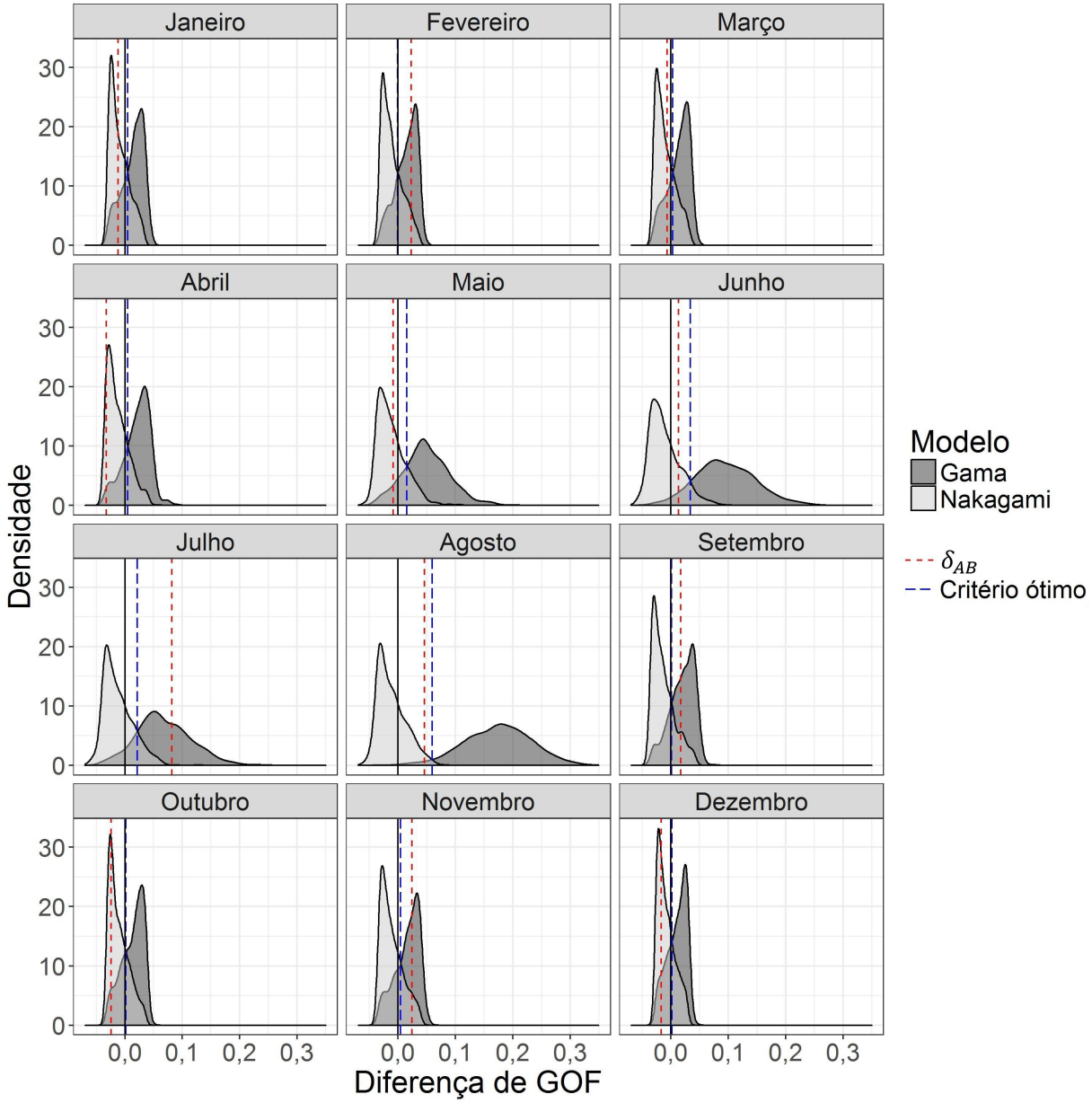


Figura 10 – Distribuições das diferenças de GOF obtidas pelo DIPBC aos dados mensais de precipitação da estação meteorológica convencional de Maringá – PR, de 1964 a 2016, para comparação das distribuições Gama e Nakagami.

Ao contrário da distribuição Gama, uma das mais utilizadas na modelagem de dados climáticos, ainda é pequeno o número de aplicações da Nakagami a este tipo de variáveis, sendo que em geral, a mesma é utilizada na área de engenharia de comunica-

ções (KARAGIANNIDIS; SAGIAS; MATHIOPOULOS, 2007). Ressalta-se que qualquer distribuição de probabilidade, com suporte nos números reais positivos, pode ser utilizada na descrição do comportamento de séries climatológicas, mas em se tratando de séries de pluviosidade total diária,decendial, mensal, entre outras, as distribuições Weibull e Gama são as mais utilizadas (MAZUCHELI; EMANUELLI, 2015).

No presente estudo, observou-se que em oito das 12 séries históricas consideradas, o critério de decisão do método PBC apontou que a Nakagami é a distribuição mais plausível para a descrição do comportamento da variável de interesse. Ramos e Moala (2014) também compararam distribuições utilizadas frequentemente em climatologia (Gama, Weibull e Lognormal), com uma distribuição não usual (Exponencial Geométrica Estendida), mostrando que distribuições menos usuais podem ser aplicadas com sucesso em dados referentes a precipitação.

Um estudo recente, na análise da pluviosidade total mensal, determinou que a Nakagami foi a distribuição mais apropriada em $34,43\%$ das séries seguida por $29,72\%$ e $22,88\%$ pelas distribuições Weibul e Gama, respectivamente (MAZUCHELI; EMANUELLI, 2015). Outra modelagem de dados de precipitação mensal apresentou resultados indicando que a distribuição Gama ajustou-se mais adequadamente que as distribuições log-normal e Weibull (RODRIGUES; FILHO; CHAVES, 2013). Em dados de chuva total anual, além da Gama, foram encontrados na literatura ajustes satisfatórios para as distribuições Gumbel, Normal e Weibull (SILVA et al., 2013). No entanto, nenhum deste trabalhos testou a característica de mimetismo visando mostrar os modelos mais simplistas. No presente estudo testou-se exatamente isso.

Com os dados observados foi possível notar, que o viés causado pelo mimetismo foi relativamente baixo para a maior parte dos meses, sendo que os maiores valores do viés foram observados para os dados referentes a junho e agosto, meses nos quais o PBC indicou a distribuição, contrário ao apontado pela estatística isoladamente, favorecendo a escolha pela Nakagami. Como comentado anteriormente, a estatística , uma medida de GOF, não considera a complexidade funcional das distribuições para selecioná-las, tendendo a escolher distribuições mais complexas.

De um modo geral, distribuições complexas serão capazes de imitar distribuições mais simples. No entanto, o aumento da complexidade não aumenta necessari-

amente o mimetismo. Assim, o DIPBC só pode discriminar entre as distribuições na medida em que estes são funcionalmente diferentes. Se uma distribuição é estruturalmente mais complexa, mas funcionalmente idêntica a uma distribuição mais simples, a seleção por esse método será potencialmente enganosa (WAGENMAKERS et al., 2004).

As áreas sobrepostas entre as distribuições de ΔGOF^*_{AB}, que são delimitadas pelo critério definido como ótimo, representam a habilidade de mimetismo de cada distribuição. Assim, observa-se que pelos resultados do PBC, a distribuição Gama exibiu maior flexibilidade na explicação dos dados provenientes da Nakagami do que o contrário. Ressalta-se que os resultados gerados pelo método são específicos aos conjuntos de observações analisadas.

O PBC pode ser interpretado como uma implementação frequentista do método BPP (*Bayesian posterior predictive p-values*), usualmente utilizado para avaliar o ajuste de distribuições. Ambos os métodos geram distribuições de valores de GOF esperados para as distribuições sob consideração, que podem ser usadas para avaliar o mimetismo ou a adequação da distribuição. Entretanto, no procedimento BPP, a distribuição para gerar os dados simulados é explicitamente Bayesiano (WAGENMAKERS et al., 2004).

Se a distribuição dos parâmetros respectivos aos modelos obtida pelo *Bootstrap* não paramétrico é idêntica a distribuição à *posteriori* Bayesiana dos parâmetros, os dois métodos produzem os mesmos resultados, sendo que a distribuição não paramétrica tem vantagens práticas. Perturbando os dados, o *Bootstrap* aproxima do efeito Bayesiano de perturbação dos parâmetros, sendo normalmente muito mais simples de se realizar (HASTIE; TIBSHIRANI; FRIEDMAN, 2001).

3.4 Considerações finais

O desenvolvimento do presente trabalho permitiu a avaliação da utilização de um método recente de discriminação de distribuições, o PBC, para o ajuste de variáveis climatológicas, área na qual o método ainda não foi explorado, de acordo com pesquisa na literatura. Embora a Nakagami não seja uma distribuição usualmente utilizada para o ajuste de variáveis climatológicas, por meio da aplicação do método PBC a dados de precipitação reais, foi verificado que a mesma apresenta melhor ajuste se comparada

a distribuição Gama para a maior parte dos meses considerados e que em geral, a distribuição Gama apresenta maior complexidade funcional em relação a Nakagami, com maior viés de mimetismo para todas as séries históricas consideradas.

Ainda, os resultados da aplicação do método indicaram a importância da consideração do mimetismo na discriminação entre distribuições, uma vez que justamente para os casos em que foram observados os maiores valores de viés de mimetismo, o PBC apontou uma conclusão contrária a sugerida pela medida de qualidade de ajuste utilizada, a estatística . Tais constatações recomendam o método PBC para confrontar duas distribuições levando em consideração as respectivas complexidades funcionais, se comparado aos critérios de GOF usuais. Entretanto, é preferível que a discriminação entre distribuições seja baseada na aplicação de ambas as técnicas, já que estas avaliam diferentes questões.

Diante da utilidade do PBC apresentada neste trabalho, torna-se evidente a necessidade de sua exploração e aplicação em diversos contextos, sobretudo em climatologia, destacando-se ainda a proposta de Schultheis e Naidu (2014), que apresentam uma extensão do PBC, denominada método MMPBC (*Multi-Model Parametric Bootstrap Cross-Fitting*), que é aplicável, em princípio, a um número arbitrário de modelos concorrentes. Visto isso, a pesquisa pode ser continuada estendendo a comparação para outros modelos.

Capítulo 4

Aplicação da distribuição Gumbel baseada em valores de recordes

Resumo

Este capítulo apresenta a caracterização frequentista da distribuição Gumbel baseada apenas nos valores de recorde, que pode ser de grande utilidade não só para a estimação dos parâmetros da distribuição quando apenas os valores de recorde são observados, mas também para a previsão de recordes futuros. A metodologia foi aplicada a um conjunto de dados de precipitações mensais da estação meteorológica de Maringá - PR, observados entre 1964 e 2016. Foi verificado que, para alguns dos meses considerados, o ajuste aos dados apenas com os valores de recorde foi satisfatório quando comparado ao ajuste com os dados originais. Entretanto para alguns meses o ajuste pelos recordes se mostra bastante distante da distribuição empírica dos dados, assim como do ajuste por meio dos dados originais. Desta forma, a adoção de outras abordagens ou distribuições para o ajuste ao conjunto de dados deve ser considerada.

Palavras-chave: Recordes, Gumbel, Precipitação.

Abstract

This paper presents the frequentist characterization of the Gumbel distribution based only on record values, which can be very useful not only for the estimation of the parameters of the distribution, when only the record values are observed, but also in the prediction of future records. The methodology was applied to a set of monthly precipitation data from the meteorological station of Maringá - PR, observed between 1964 and 2016. It was verified that for some of the months considered, the fit to the data with only the values of record was satisfactory when compared the fit with the original data. However, for some months the fit for the records is quite different from the empirical distribution of the data, as well as the fit by the original data. Thus, the adoption of other approaches or distributions for fit the data set should be considered.

Keywords: Records, Gumbel, Precipitation.

4.1 Introdução

Em várias situações, apenas as observações que excedem o valor máximo atual - ou aquelas que ficam abaixo do valor mínimo atual - são registradas, sendo que essas observações são denominadas recordes. Os exemplos incluem meteorologia, hidrologia, esportes e mineração. O teste de esforço industrial também é um exemplo no qual apenas os itens mais fracos do que todos os itens observados são destruídos (AHMADI; ARGHAMI, 2003).

Um recorde é um registro em uma série temporal que é maior (ou menor) do que todos os registros anteriores. Nesse sentido, Wergen (2012) aponta que um recorde é um valor extremo que é definido em relação a todos os valores anteriores na série de tempo. O termo recorde deriva do latin *recordari* (lembrar), uma vez que um recorde provavelmente será lembrado pelos observadores, por geralmente se tratar de eventos raros.

O esquema de recordes é um método para reduzir o tempo total no teste de uma experiência (DOOSTPARAST; DEEPAK; ZANGOIE, 2013). Nesse esquema, os itens são observados sequencialmente e somente valores menores que todos os anteriores

são gravados. Em algumas situações, quando as experiências são demoradas ou os itens são perdidos durante sua execução, o esquema de recordes domina o esquema de amostra aleatória usual (DOOSTPARAST; BALAKRISHNAN, 2011). Entretanto o interesse nesta área de pesquisa não é apenas na estimativa, mas também na previsão de recordes futuros (MBAH, 2007).

A previsão estatística de valores de recorde tem potenciais aplicações ambientais que lidam, por exemplo, com saltos climáticos bruscos, como a previsão de extremos de precipitação, de níveis de água mais altos na superfície do mar ou de recordes de temperatura do ar (MADI; RAQAB, 2004).

Chandler (1952) introduziu o estudo de valores de recorde, documentando muitas de suas propriedades básicas. Os valores de recordes podem ser vistos como estatísticas de ordem de uma amostra cujo tamanho é determinado pelos valores e pela ordem de ocorrência das observações (NAGARAJA, 1988).

O processo padrão de valor de recorde corresponde a uma sequência infinita de observações independentes e identicamente distribuídas (i.i.d.). Uma observação X_j será chamada de recorde superior (ou simplesmente recorde) se seu valor exceder o de todas as observações anteriores, isto é, X_j é um recorde se $X_j > X_i$ para todo $i < j$. O mesmo raciocínio é utilizado para a definição de recordes inferiores (ARNOLD; BALAKRISHNAN; NAGARAJA, 2011). Neste momento, serão consideradas apenas variáveis contínuas, sendo que algumas adaptações devem ser consideradas ao se tratar com variáveis de natureza discreta. Considere a seguinte definição de (MBAH, 2007):

Definição: Seja $X_1, X_2, ..., X_n$ uma sequência de variáveis aleatórias i.i.d. com função de distribuição acumulada $F(x)$ e seja $X_n = max\{X_1, X_2, ..., X_n\}$ para $n \geq 1$. Diz-se que X_j é um recorde superior de $X_n, n \geq 1$ se $X_j > X_{j-1}$, $j > 1$. Existe uma definição análoga para os valores de recorde inferiores. Por definição, X_1 é um recorde superior, bem como recorde inferior.

A definição acima da sequência de recordes supõe implicitamente que a $F(x)$ não produzirá nenhum recorde "inquebrável". Por exemplo, ao lançar um dado, o recorde 6 é inquebrável, uma vez que não é possível observar uma face com valor maior que 6. Este não será o caso se existir algum valor x_0 tal que $F(x_0) - F(x_0-) > 0$ e $F(x_0) = 1$, ou seja, se existe um maior valor real possível que pode ser alcançado com

probabilidade positiva pelos X_j's (ARNOLD; BALAKRISHNAN; NAGARAJA, 2011).

Dessa forma, considerando os valores de recorde superior X_n, $n = 1, 2, ...$, que se baseiam na sequência de variáveis aleatórias i.i.d. $X_1, X_2, ...$, com uma função de distribuição contínua $F(x \mid \theta)$. Usando a definição $X_n = M(T_n)$ e sendo $M(n) = max\{X_1, ..., X_n\}$, de acordo com Ahsanullah e Nevzorov (2015), pode-se escrever que:

$$\begin{aligned} P(X_n < x) &= \sum_{m=n}^{\infty} P(M(T_n) < x \mid T_n = m) P(T_n = m) \\ &= \sum_{m=n}^{P(X(n)<x)} P(M(m) < x \mid T_n = m) P(T_n = m) \\ &= \sum_{m=n}^{\infty} P(M(m) < x) P(T_n = m) \\ &= \sum_{m=n}^{\infty} F^m(x) P(T_n = m) \\ &= E(F(x))^{T_n} \\ &= Q_n(F(x)) \end{aligned}$$

em que a expressão correspondente para a função geradora $Q_n(s)$ é dada por:

$$Q_n(s) = \frac{1}{(n-1)!} \int_0^{-\log(1-s)} \nu^{n-1} exp(-\nu) d\nu. \tag{4.1}$$

Assim, tem-se que:

$$P(X_n < x) = \frac{1}{(n-1)!} \int_0^{-\log(1-F(x))} \nu^{n-1} exp(-\nu) d\nu, -\infty < x < \infty, n = 1, 2, ... \tag{4.2}$$

Se $F(x \mid \theta)$ é absolutamente contínua, diferenciando-a, obtem-se a função de densidade de probabilidade correspondente $f(x \mid \theta)$:

$$f(X_n(x) \mid \theta) = f(x) \frac{[-\log(1 - F(x))]^{n-1}}{(n-1)!}. \tag{4.3}$$

Assim, de acordo com Arnold, Balakrishnan e Nagaraja (2011) a função de verossimilhança baseada nos recordes é dada por:

$$L(\boldsymbol{\Theta} \mid \boldsymbol{x}) = f(x_n \mid \boldsymbol{\Theta}) \prod_{i=1}^{m-1} \frac{f(x_i \mid \boldsymbol{\Theta})}{1 - F(x_i \mid \boldsymbol{\Theta})}. \tag{4.4}$$

4.1.1 Revisão da literatura

A seguir, apresenta-se as principais distribuições, as abordagens e as aplicações geralmente utilizadas nos trabalhos envolvendo a metodologia de recordes, levantadas por meio de uma revisão da literatura.

4.1.1.1 Distribuições

Na literatura, vários autores concentraram-se na obtenção de estimadores de várias distribuições com base em valores e tempos de recorde. Dentre elas cita-se a distribuição de Gumbel (MOUSA; JAHEEN; AHMAD, 2002), distribuição de Pareto (MADI; RAQAB, 2004) e (DOOSTPARAST; AKBARI; BALAKRISHNA, 2011), distribuição Weibull (SOLIMAN; ELLAH; SULTAN, 2006), distribuição exponencial generalizada (MADI; RAQAB, 2007) e (DEY et al., 2013), entre outras.

A distribuição Weibull é relatada no trabalho de Soliman, Ellah e Sultan (2006), no qual estimativas de máxima verossimilhança (EMV) e Bayesianas baseadas em valores de recorde são derivadas para os dois parâmetros desconhecidos da distribuição, assim como para alguns parâmetros de tempo de sobrevivência e funções de confiabilidade e risco, apontando que os resultados podem ser de interesse em uma situação em que apenas os valores de recorde são armazenados.

Outras contribuições, considerando a distribuição Weibull, foram propostas por Teimouri e Nadarajah (2013) e Zakerzadeh e Jafari (2015). Os primeiros autores apresentam EMV corrigidos por viés para os parâmetros de forma e de escala da distribuição de Weibull, demonstrando seu desempenho superior. Já os últimos autores propõem métodos exatos e simples para testar e construir um intervalo de confiança para os parâmetros da distribuição Weibull, e para inferência sobre o parâmetro de forma, além de apresentar uma abordagem generalizada para a inferência sobre o parâmetro de escala.

Em Mousa, Jaheen e Ahmad (2002), os autores apresentam a estimativa Bayesiana para os dois parâmetros da distribuição de Gumbel, obtida com base em valores de recorde, além da previsão, pontual ou intervalar, para futuros valores de recorde inferiores, também a partir de um ponto de vista Bayesiano. A distribuição Gumbel, também conhecida como distribuição valor extremo do Tipo I, também é caracterizada no trabalho de Alzaid e Ahsanullah (2003).

A previsão Bayesiana de recordes de temperatura usando o modelo de Pareto é abordada por Madi e Raqab (2004), que desenvolvem a distribuição preditiva Bayesiana para recordes futuros e estabelece os correspondentes intervalos de maior densidade a posteriori (HPD). A distribuição de Pareto também é o foco em Doostparast, Akbari e Balakrishna (2011), no qual são desenvolvidos os EMV e estimadores Bayesianos dos dois parâmetros da distribuição, com base em valores e tempos de recorde.

Uma previsão Bayesiana de recordes de precipitação de Los Angeles usando a distribuição exponencial generalizada (GED) é apresentada em Madi e Raqab (2007), que utilizam a amostragem de importância para estimar os parâmetros do modelo, e os amostradores de Gibbs e Metropolis–Hastings para implementar o procedimento de predição. Já o trabalho de Dey et al. (2013) concentra-se na inferência Bayesiana da GED com base em valores de recordes inferiores, obtendo as EMV e as estimativas Bayesianas dos dois parâmetros da distribuição.

Em seu trabalho, Ahmadi e Doostparast (2006) apresentam a estimação e a previsão Bayesiana para algumas distribuições de vida com base em valores recorde superiores, incluindo Exponencial, Weibull, Pareto e Burr tipo XII.

Recentemente, Wang, Wang e Yu (2017) abordaram a inferência para a distribuição de Kumaraswamy, com base nos valores de recorde, obtendo as EMV para os parâmetros do modelo e construindo séries de intervalos de confiança exatos e regiões de confiança exatas.

Ainda, estimadores Bayesianos e não-Bayesianos usando valores de recorde são apresentados por Panaitescu et al. (2010) para a distribuição inversa modificada de Weibull, por Nadar e Papadopoulos (2011) para a distribuição Burr tipo XII, por Mubarak (2011) para a distribuição de Frèchet, por Doostparast, Deepak e Zangoie (2013) para a distribuição Lognormal e por Asgharzadeh et al. (2016) para a distribuição

Lindley.

Com enfoque na caracterização de classes gerais de distribuições baseadas na propriedade de independência dos valores de recorde transformados, Juhás e Skřivánková (2014) discutem exemplos de casos especiais de classes gerais como Gumbel, Fréchet, Weibull, distribuições exponenciais e lognormal.

4.1.1.2 Abordagens

Como visto acima, uma grande quantidade de trabalhos foram desenvolvidos considerando um enfoque Bayesiano para os problemas de estimação e inferência baseados em recordes, como em Mousa, Jaheen e Ahmad (2002), Ahmadi e Doostparast (2006), Madi e Raqab (2007), Nadar e Papadopoulos (2011), Doostparast, Akbari e Balakrishna (2011), entre outros. Nesses trabalhos são discutidos aspectos como as distribuições a priori utilizadas, os métodos de amostragem e as funções de perda utilizadas.

Também foram observados diversos trabalhos que comparam os resultados obtidos por meio das abordagens Bayesiana e frequentista, que se utilizam em geral da função de verossimilhança para derivação dos resultados. Entre estes trabalhos cita-se: Soliman, Ellah e Sultan (2006), Panaitescu et al. (2010), Dey et al. (2013) e Asgharzadeh et al. (2016).

Alguns trabalhos com enfoque não paramétrico para estimação e inferência baseada em valores de recorde são encontradas na literatura. Em Ahmadi e Arghami (2003), mostra-se como os valores de recorde podem ser usados para proporcionar intervalos de confiança livres de distribuição para quantis e intervalos de tolerância, sendo que os resultados podem ser de interesse na situação em que apenas os valores de recorde são armazenados. Por outro lado, intervalos de predição não paramétricos para futuros valores de recorde são construídos em Raqab (2009).

Já no recente trabalho dos autores Ahmadi, Basiri e Kundu (2017), os intervalos de confiança para os quantis são construídos com base em recordes superiores e inferiores interpolados e os intervalos de previsão são obtidos para futuros recordes superiores baseados em recordes superiores interpolados. Além disso, os limites superiores para o erro de cobertura desses intervalos de confiança e de predição são derivados.

4.1.1.3 Aplicações

Em vários desses estudos, que buscam caracterizar distribuições, obter estimativas, predições e desenvolver procedimentos de inferência baseados em valores de recorde, são relatadas aplicações da metodologia considerando conjuntos de dados reais de diversas áreas.

Nos trabalhos de Soliman, Ellah e Sultan (2006) e Zakerzadeh e Jafari (2015), que discutem aspectos da distribuição Weibull, os autores se utilizam do mesmo banco de dados, referente aos valores de recorde reais de um teste acelerado sobre o fluido isolante. Já Teimouri e Nadarajah (2013) realizam aplicações a três bancos de dados de: tempos de itens testados até a falha; vidas em anos e tempos entre falhas, em milhares de horas de bombas de reator secundário.

Já em Doostparast, Akbari e Balakrishna (2011), é analisado um conjunto de dados sobre as embarcações anuais de uma amostra de trabalhadores da linha de produção em uma grande empresa industrial, utilizando os procedimentos propostos. Enquanto que a aplicação apresentada em Wang, Wang e Yu (2017) utiliza os dados mensais de capacidade de água do reservatório de Shasta na Califórnia – EUA, referentes ao mês de fevereiro, durante o período de 1991 a 2010.

Também foram observadas algumas aplicações a dados que envolvem variáveis climatológicas. Um conjunto de dados que representa os valores recorde das temperaturas médias de julho em Neuenburg - Suíça, é usado em Madi e Raqab (2004) para ilustrar a aplicação ambiental do procedimento de previsão proposto.

Ainda, no trabalho de Ahmadi, Basiri e Kundu (2017), são realizadas aplicações a dados reais: da quantidade de precipitação anual (em polegadas) registrada no Centro Cívico de Los Angeles durante o período de 1890 até 1989, das temperaturas diárias (em graus Fahrenheit) registrados no National Center of Atmospheric Research (NCAR) durante o ano de 2005.

Por fim, em Asgharzadeh et al. (2016) os autores analisam a precipitação anual total (em polegadas) durante março, registrada no Centro Cívico de Los Angeles de 1973 a 2006.

4.2 Objetivos

Diante do exposto, o presente trabalho tem por objetivo caracterizar a distribuição Gumbel, com base nos valores de recorde superiores, aplicando a metodologia proposta para ajustar os recordes de precipitações mensais observadas na estação meteorológica convencional de Maringá - PR.

4.3 Materiais e métodos

4.3.1 Dados

Foram utilizados dados da estação meteorológica convencional da cidade de Maringá - PR (Figura 11), compilados a partir das séries históricas de precipitação mensais obtidas no Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) do Instituto Nacional de Meteorologia (INMET), órgão responsável pela coleta e disponibilização informações meteorológicas oficiais.

Figura 11 – Localização da estação meteorológica convencional de Maringá - PR. Fonte: Google Maps (2017).

Considerou-se como variável de estudo o total de precipitação acumulada no mês, a fim de ajustar a distribuição Gumbel para a explicação de tal fenômeno, com

base nos valores de recorde. Nos dados obtidos do INMET encontram-se disponíveis, de forma digital, apenas os registros mensais a partir do ano de 1961. Entretanto foram selecionadas as séries históricas disponíveis no período entre janeiro de 1964 e dezembro de 2016, já que entre os anos de 1961 e 1963 os registros da estação convencional de Maringá foram realizados apenas para alguns meses.

As informações referentes aos volumes de precipitação mensais não estavam disponíveis para todos os meses do período considerado, sendo que no total, foram utilizadas 527 observações, divididas em 12 séries mensais para o ajuste da distribuição.

4.3.2 Distribuição Gumbel

Segundo Nadarajah (2006), a distribuição Gumbel é possivelmente a distribuição estatística mais amplamente aplicada para modelagem climática. Diversos autores utilizaram a distribuição Gumbel nesse contexto, citanto a análise da temperatura mínima mensal (BARBOSA et al., 2014), a análise das precipitações máximas mensais (SANTOS et al., 2014) e a análise de máximos diários de velocidade do vento e (LISKA et al., 2013).

Ainda, em estudo recente, Cremoneze (2015) aponta que, de acordo com a avaliação das performances das generalizações da distribuição Gumbel, ajustas as mesmas séries climatológicas formadas pelos máximos mensais de precipitação nas estações da região sul do Brasil, verificou-se que a distribuição Gumbel obteve melhor ajuste segundo os critérios de discriminação adotados: Informação de Akaike (AIC), Informação de Akaike corrigido (AICc) e Informação Bayesiana (BIC). Assim, distribuição Gumbel padrão, com 2 parâmetros, obteve melhor performance comparada às suas generalizações.

Desta forma, a distribição Gumbel será utilizada para o ajuste aos dados de precipitação mensal e de previsões de recordes.

A distribuição Gumbel foi originalmente proposta por Fisher e Tippett (1928), em que os autores definiram três distribuições assintóticas de valores extremos, conhecidas como Gumbel, Fréchet e Weibull ou Valor Extremo tipo I, II e III, respectivamente.

A função densidade de probabilidade da distribuição Gumbel é definida por:

$$f(x \mid \boldsymbol{\Theta}) = \frac{1}{\sigma} \exp\left[-\frac{x-\mu}{\sigma} - \exp\left(-\frac{x-\mu}{\sigma}\right)\right], \tag{4.5}$$

em que $\boldsymbol{\Theta} = (\mu, \sigma)$, sendo $\mu \in \mathbb{R}$ o parâmetro de locação e $\sigma > 0$ o parâmetro de escala. Desta forma, a média e a variância de uma variável aleatória contínua que segue uma distribuição Gumbel são dadas, respectivamente, por:

$$E(X) = \mu + \gamma\sigma, \tag{4.6}$$

$$V(X) = \frac{\pi^2}{6}\sigma^2, \tag{4.7}$$

em que γ é a constante de Euler, $\gamma \approx 0,5772$ (PINHEIRO, 2013). A função de distribuição acumulada é escrita na forma:

$$F(x \mid \boldsymbol{\Theta}) = \exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right]. \tag{4.8}$$

Substituindo a função densidade e a função de distribuição acumulada dadas em 4.5 e 4.8, respectivamente, na equação 4.4, a função de verossimilhança baseada nos recordes é dada por:

$$\begin{aligned} L(\boldsymbol{\Theta} \mid X) &= \frac{1}{\sigma} \exp\left[-z_m - \exp\left(-z_m\right)\right] \prod_{i=1}^{m-1} \frac{\frac{1}{\sigma}\exp\left[-z_i - \exp\left(-z_i\right)\right]}{1 - \exp\left[-\exp\left(-z_i\right)\right]} \\ &= \frac{1}{\sigma^m} \prod_{i=1}^{m} \exp\left[-z_i - \exp\left(-z_i\right)\right] \prod_{i=1}^{m-1} \left(1 - \exp\left[-\exp\left(-z_i\right)\right]\right)^{-1} \end{aligned} \tag{4.9}$$

em que $z_i = \frac{x_i - \mu}{\sigma}$ e m é o número de recordes observados. Assim, o logaritmo da função de verossimilhança é dado por:

$$\ell(\boldsymbol{\Theta} \mid X) = m \log\sigma + \sum_{i=1}^{m}\left[-z_i - \exp\left(-z_i\right)\right] - \sum_{i=1}^{m-1} \log\left(1 - \exp\left[-\exp\left(-z_i\right)\right]\right) \tag{4.10}$$

Portanto o estimador de máxima verossimilhança (EMV) do vetor de parâmetros $\boldsymbol{\Theta}$, denotado por $\widehat{\Theta}$, pode ser obtitido maximizando a equação 4.10, com respeito a Θ.

A avaliação dos ajustes da distribuição aos conjuntos de dados será realizada por meio das estatísticas Kolmogorov Smirnov (KS), Anderson-Darling (AD) e Cramér-von Mises (CvM).

4.4 Resultados

Na Tabela 8, apresenta-se o número de observações n, o número de recordes m e os valores dos recordes observados nas 12 séries históricas de precipitação agrupadas por mês, da estação meteorológica convencional de Maringá – PR, de 1964 a 2016.

Tabela 8 – Recordes das séries históricas de precipitação agrupadas por mês, da estação meteorológica convencional de Maringá – PR, de 1964 a 2016.

Mês	**n**	**m**	**Recordes**							
Janeiro	45	5	71,9	271,3	287,4	354,6	419,8			
Fevereiro	45	4	339,9	359,0	362,6	426,0				
Março	45	7	130,5	142,2	203,4	256,0	268,8	277,3	340,0	
Abril	44	3	100,2	221,1	346,3					
Maio	43	7	74,4	131,0	140,4	159,0	237,8	322,0	396,4	
Junho	43	4	119,0	279,2	332,8	396,7				
Julho	43	7	60,4	78,0	108,0	126,7	148,7	227,9	378,6	
Agosto	42	5	87,1	89,3	139,3	152,6	219,8			
Setembro	45	8	21,4	33,4	84,4	94,0	162,4	212,7	308,9	319,6
Outubro	43	5	130,4	254,6	264,6	277,6	345,6			
Novembro	44	8	52,2	84,6	101,0	169,0	249,7	306,0	358,8	369,6
Dezembro	45	5	155,4	253,4	281,0	284,2	360,4			

* n: número de observações; m: número de recordes.

Observa-se que o número de observações caracterizadas como recordes varia entre os meses, sendo que em Abril foram observados apenas 3 recordes, enquanto que nos meses de Setembro e Novembro foram observados 8 valores de recorde. Apenas

nos meses de Janeiro e Fevereiro os recordes observados ultrapassam os 400 mm, ao passo que para Agosto, o máximo observado foi de 219,80 mm.

Utilizando apenas os valores de recorde apresentados na Tabela 8, ajustou-se a distribuição Gumbel, obtendo-se os EMV por meio da função descrita na equação 4.10. Os resultados deste ajuste e do ajuste aos dados originais, assim como a distribuição empírica da precipitação mensal, são apresentados na Figura 12.

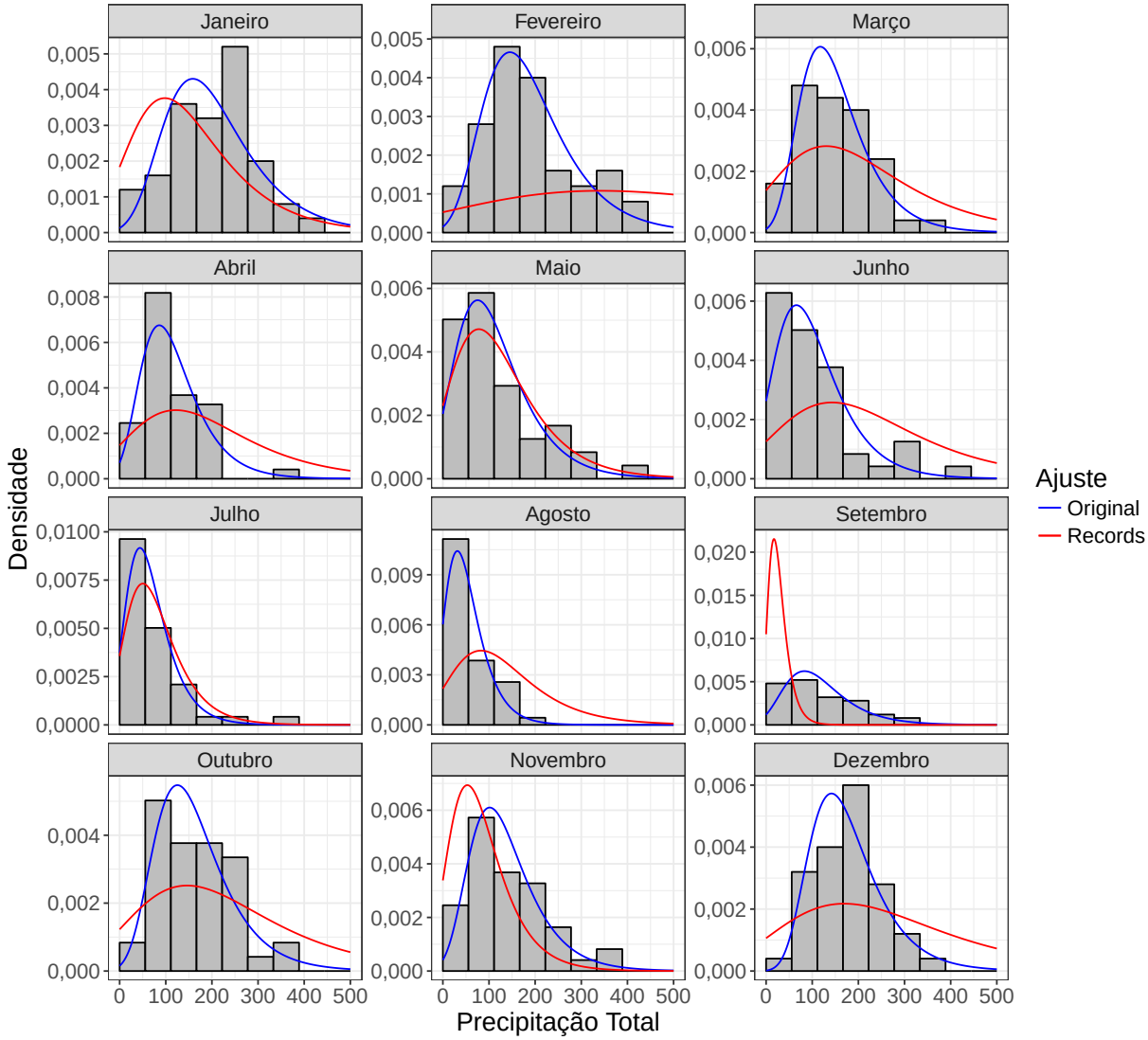


Figura 12 – Histograma e ajuste da distribuição Gumbel por meio dos dados originais e dos valores de recorde, às séries históricas de precipitação agrupadas por mês, da estação meteorológica convencional de Maringá – PR, de 1964 a 2016.

Pela Figura 12, nota-se que os ajustes obtidos por meio dos valores de recorde são bastante próximos dos ajustes obtidos por meio dos dados originais para os meses de Maio e Julho. Já para os meses de Fevereiro e Setembro, o ajuste pelos recordes se mostra bastante diferente da distribuição empírica dos volumes mensais de precipitação, assim como do ajuste por meio dos dados originais.

A Tabela 9 apresenta as estimativas e erros padrões (E.P.) dos parâmetros e das medidas de GOF do ajuste da distribuição Gumbel, por meio dos valores de recorde das séries históricas de precipitação agrupadas por mês, da estação meteorológica convencional de Maringá – PR, de 1964 a 2016.

Tabela 9 – Estimativas dos parâmetros e medidas de GOF do ajuste da distribuição Gumbel por meio dos valores de recorde, às séries históricas de precipitação agrupadas por mês, da estação meteorológica convencional de Maringá – PR, de 1964 a 2016.

Mês	$\widehat{\mu}$ **(E.P.)**	$\widehat{\sigma}$ **(E.P.)**	**KS**	**AD**	**CvM**
Janeiro	97,81 (61,04)	65,76 (26,89)	0,382	15,304	2,720
Fevereiro	340,74 (18,69)	19,31 (9,11)	0,852	—	10,107
Março	130,38 (27,31)	28,63 (10,42)	0,242	18,615	0,930
Abril	121,84 (68,19)	72,87 (37,49)	0,290	7,077	1,406
Maio	78,01 (40,93)	42,39 (15,17)	0,155	4,034	0,311
Junho	142,81 (59,72)	64,24 (28,88)	0,457	37,859	4,160
Julho	50,28 (42,24)	42,23 (15,40)	0,150	1,570	0,311
Agosto	82,66 (24,42)	25,26 (10,76)	0,576	134,900	5,145
Setembro	17,09 (34,36)	35,42 (12,09)	0,494	40,516	5,656
Outubro	146,12 (37,65)	40,56 (16,54)	0,272	14,054	1,030
Novembro	53,04 (36,06)	37,44 (12,80)	0,408	22,067	3,238
Dezembro	169,26 (35,61)	38,29 (15,58)	0,234	15,534	0,495

* E.P.: Erro Padrão.

Nota-se que mesmo para os meses que apresentaram ajustes próximos à distribuição empírica, os erros padrões dos estimadores são relativamente altos. Em relação às medidas de qualidade de ajuste, os resultados corroboram o exposto na Figura 12, sendo que o ajuste para os meses de Maio e Junho foram os que apresentaram as menores medidas, indicando um melhor ajuste.

Não foi observada uma relação clara entre o número de valores de recorde observados e a qualidade do ajuste da distribuição Gumbel com base nesses, visto que, tanto entre os piores quanto entre os melhores ajustes, haviam meses com poucos e muitos recordes observados.

4.5 Considerações finais

Com o presente trabalho foi avaliada a utilização do esquema de recordes para o ajuste de variáveis climatológicas, especificamente para o volume mensal de precipitação, considerando a distribuição Gumbel. Foi verificado que para alguns dos meses considerados, o ajuste aos dados apenas com os valores de recorde foi satisfatório quando comparado ao ajuste com os dados originais. Entretanto para alguns meses o ajuste pelos recordes se mostra bastante diferente da distribuição empírica dos volumes mensais de precipitação, assim como do ajuste por meio dos dados originais.

Embora seja levantada a hipótese de que o número de recordes ou mesmo de observações possa ser um dos motivos para o ajuste precário da distribuição Gumbel aos dados, considerando apenas os valores de recorde, tal relação não foi verificada na aplicação realizada, visto que, tanto entre os piores quanto entre os melhores ajustes, haviam meses com poucos e muitos recordes observados.

Desta forma, levanta-se a abordagem Bayesiana para a estimação dos parâmetros da distribuição Gumbel ou até mesmo a consideração de outra distribuição, como a Nakagami, possa apresentar melhores resultados, sendo que tais propostas podem ser consideradas no desenvolvimento de trabalhos futuros.

Assim, considerando não só a aplicação do esquema de recordes para o ajuste a uma variável aleatória para a qual apenas os recordes são registrados, o esquema de recordes apresenta grande utilidade na previsão de recordes futuros, aspecto que ainda pode ser explorado em trabalhos futuros.

Referências

AHMADI, J.; ARGHAMI, N. R. Nonparametric confidence and tolerance intervals from record values data. *Statistical Papers*, Springer, v. 44, n. 4, p. 455–468, 2003. Citado 2 vezes nas páginas 58 e 63.

AHMADI, J.; BASIRI, E.; KUNDU, D. Confidence and prediction intervals based on interpolated records. *Journal of Nonparametric Statistics*, Taylor & Francis, v. 29, n. 1, p. 1–21, 2017. Citado 2 vezes nas páginas 63 e 64.

AHMADI, J.; DOOSTPARAST, M. Bayesian estimation and prediction for some life distributions based on record values. *Statistical Papers*, Springer, v. 47, n. 3, p. 373–392, 2006. Citado 2 vezes nas páginas 62 e 63.

AHSANULLAH, M.; NEVZOROV, V. B. *Records via Probability Theory*. [S.l.]: Springer, 2015. Citado na página 60.

AKAIKE, H. A new look at the statistical model identification. *IEEE transactions on automatic control*, Ieee, v. 19, n. 6, p. 716–723, 1974. Citado na página 26.

ALZAID, A.; AHSANULLAH, M. A characterization of the gumbel distribution based on record values. Taylor & Francis, 2003. Citado na página 62.

ARNOLD, B. C.; BALAKRISHNAN, N.; NAGARAJA, H. N. *Records*. [S.l.]: John Wiley & Sons, 2011. v. 768. Citado 3 vezes nas páginas 59, 60 e 61.

ASGHARZADEH, A.; FALLAH, A.; RAQAB, M.; VALIOLLAHI, R. Statistical inference based on lindley record data. *Statistical Papers*, Springer, p. 1–21, 2016. Citado 3 vezes nas páginas 62, 63 e 64.

ATKINSON, A. C. *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. New York: Oxford University Press, 1985. Citado na página 25.

BARBOSA, E. C.; SILVA, C. H. O.; MANULI, R. C.; TAVARES, R. G.; NAZARé, T. B. Distribuição generalizada de valores extremos (gve): Um estudo aplicado a valores de temperatura mínima da cidade de viçosa-mg. *Revista da Estatística UFOP*, III, n. 3, p. 387–391, 2014. Citado na página 66.

BARNDORFF-NIELSEN, O.; JØRGENSEN, B. Some parametric models on the Simplex. *Journal of Multivariate Analysis*, v. 39, n. 1, p. 106–116, 1991. ISSN 0047-259X. Citado na página 18.

BAYES, C. L.; BAZÁN, J. L.; GARCÍA, C. et al. A new robust regression model for proportions. *Bayesian Analysis*, International Society for Bayesian Analysis, v. 7, n. 4, p. 841–866, 2012. Citado na página 18.

BEYRUTH, Z. Água, agricultura e as alterações climáticas globais. *Revista tecnologia & inovação agropecuária*, v. 1, n. 1, p. 74–89, 2008. Citado na página 45.

BONAT, W. H.; JR, P. J. R.; ZEVIANI, W. M. Regression models with response on the unity interval: Specification, estimation and comparison. *Biometric Brazilian Journal*, v. 30, n. 4, p. 415–431, 2012. Citado na página 17.

BUSSAB, W.; MORETTIN, P. *Estatística Básica*. 7ª edição. ed. [S.l.]: Saraiva, 2011. Citado na página 12.

CEPEDA-CUERVO, E. *Variability modeling in generalized linear models*. Tese (Doutorado) — Mathematics Institute, Universidade Federal do Rio de Janeiro, 2001. Citado na página 25.

CHANDLER, K. The distribution and frequency of record values. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 220–228, 1952. Citado na página 59.

CHOW, V.; MAIDMENT, D.; MAYS, L. *Applied Hydrology*. 2nd ed. ed. [S.l.]: Tata McGraw-Hill Education, 2013. (McGraw-Hill series in water resources and environmental engineering). Citado na página 45.

COE, R.; STERN, R. D. Fitting models to daily rainfall data. *Journal of Applied Meteorology*, v. 21, n. 7, p. 1024?1031, 1982. Citado na página 45.

COLLET, D. *Modelling binary data*. New York: Chapman & Hall/CRC, 2003. Second ed. Citado na página 25.

COX, D. R.; SNELL, E. J. A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 248–275, 1968. Citado na página 25.

CREMONEZE, I. Z. *Aplicação de algumas generalizações da Distribuição Gumbel na Análise de Dados Climatológicos*. Dissertação (Mestrado) — Universidade Estatual de Maringá, Maringá, 2015. Citado na página 66.

CRIBARI-NETO, F.; ZEILEIS, A. Beta regression in r. Department of Statistics and Mathematics x, WU Vienna University of Economics and Business, 2009. Citado na página 17.

DEY, S.; DEY, T.; SALEHI, M.; AHMADI, J. Bayesian inference of generalized exponential distribution based on lower record values. *American Journal of Mathematical and Management Sciences*, Taylor & Francis, v. 32, n. 1, p. 1–18, 2013. Citado 3 vezes nas páginas 61, 62 e 63.

DOOSTPARAST, M.; AKBARI, M. G.; BALAKRISHNA, N. Bayesian analysis for the two-parameter pareto distribution based on record values and times. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 81, n. 11, p. 1393–1403, 2011. Citado 4 vezes nas páginas 61, 62, 63 e 64.

DOOSTPARAST, M.; BALAKRISHNAN, N. Optimal record-based statistical procedures for the two-parameter exponential distribution. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 81, n. 12, p. 2003–2019, 2011. Citado na página 59.

DOOSTPARAST, M.; DEEPAK, S.; ZANGOIE, A. Estimation with the lognormal distribution on the basis of records. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 83, n. 12, p. 2339–2351, 2013. Citado 2 vezes nas páginas 58 e 62.

EFRON, B. Bootstrap methods: another look at the jackknife. In: *Breakthroughs in statistics*. [S.l.]: Springer, 1992. p. 569–593. Citado na página 46.

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 3 vezes nas páginas 17, 18 e 26.

FISHER, R. A.; TIPPETT, L. H. C. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In: CAMBRIDGE UNIV PRESS. *Mathematical Proceedings of the Cambridge Philosophical Society*. [S.l.], 1928. v. 24, n. 02, p. 180–190. Citado na página 66.

GALTON, F. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, JSTOR, v. 15, p. 246–263, 1886. Citado na página 16.

GRASSIA, A. On a family of distributions with argument between 0 and 1 obtained by transformation of the Gamma distribution and derived compound distributions. *Australian Journal of Statistics*, Blackwell Publishing Ltd, v. 19, n. 2, p. 108–114, 1977. ISSN 1467-842X. Citado na página 18.

GUPTA, R.; KUNDU, D. Generalized exponential distributions. *Australian and New Zealand Journal of Statistics*, v. 58, p. 173–188, 1999. Citado na página 19.

HAHN, E. D. Mixture densities for project management activity times: A robust approach to pert. *European Journal of Operational Research*, Elsevier, v. 188, n. 2, p. 450–459, 2008. Citado na página 18.

HANNAN, E. J.; QUINN, B. G. The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 190–195, 1979. Citado na página 26.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. [S.l.]: Springer New York, 2001. (Springer Series in Statistics). Citado na página 55.

HELD, L.; BOVÉ, D. S. Applied statistical inference. *Springer, Berlin Heidelberg, doi*, Springer, v. 10, p. 978–3, 2014. Citado na página 26.

JOHNSON, N. L. Systems of frequency curves generated by methods of translation. *Biometrika*, v. 36, n. 1/2, p. 149–176, 1949. Citado na página 18.

JUHÁS, M.; SKŘIVÁNKOVÁ, V. Characterization of general classes of distributions based on independent property of transformed record values. *Applied Mathematics and Computation*, Elsevier, v. 226, p. 44–50, 2014. Citado na página 63.

KARAGIANNIDIS, G. K.; SAGIAS, N. C.; MATHIOPOULOS, P. T. N $*$ nakagami: A novel stochastic model for cascaded fading channels. *IEEE Transactions on Communications*, IEEE, v. 55, n. 8, p. 1453–1458, 2007. Citado na página 54.

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 2 vezes nas páginas 17 e 18.

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 27.

KUMARASWAMY, P. A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, v. 46, n. 1, p. 79 – 88, 1980. Citado 2 vezes nas páginas 18 e 19.

KUTNER, M. H.; NACHTSHEIM, C. J.; NETER, J.; LI, W. *Applied Linear Statistical Models*. New York: McGraw-HillfIrwin, 2005. Fifth Edition. Citado na página 25.

LAWLESS, J. F. *Statistical Models and Methods for Lifetime Data*. Second. [S.l.]: John Wiley and Sons, 2003. Citado na página 25.

LEE, E. T.; WANG, J. W. *Statistical methods for survival data analysis*. Third. Hoboken, NJ: [s.n.], 2003. (Wiley Series in Probability and Statistics). Nenhuma citação no texto.

LEHMANN, E. J.; CASELLA, G. *Theory of Point Estimation*. [S.l.]: Springer Verlag, 1998. Citado na página 24.

LINDSAY, B. G.; LI, B. On second-order optimality of the observed Fisher information. *The Annals of Statistics*, The Institute of Mathematical Statistics, v. 25, n. 5, p. 2172–2199, 10 1997. Citado na página 24.

LISKA, G. R.; BORTOLINI, J.; SÁFADI, T.; BEIJO, L. A. Estimativas de velocidade máxima de vento em piracicaba–sp via séries temporais e teoria de valores extremos. *Rev. Bras. Biom., São Paulo*, v. 31, n. 2, p. 295–309, 2013. Citado na página 66.

MADI, M. T.; RAQAB, M. Z. Bayesian prediction of temperature records using the pareto model. *Environmetrics*, Wiley Online Library, v. 15, n. 7, p. 701–710, 2004. Citado 4 vezes nas páginas 59, 61, 62 e 64.

MADI, M. T.; RAQAB, M. Z. Bayesian prediction of rainfall records using the generalized exponential distribution. *Environmetrics*, Wiley Online Library, v. 18, n. 5, p. 541–549, 2007. Citado 3 vezes nas páginas 61, 62 e 63.

MARTIN, T. N.; NETO, D. D.; JUNIOR, P. A. V.; MANFRON, P. A. Homogeneidade espaçotemporal e modelos de distribuição para a precipitação pluvial no estado de são paulo. *Ceres*, v. 55, n. 5, 2015. Citado na página 45.

MAZUCHELI, J.; EMANUELLI, I. E. Aplicação da distribuição nakagami na análise de dados de precipitação. In: *60ª Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria (RBras) e 16º Simpósio de Estatística Aplicada a Experimentação Agronômica (SEAGRO)*. Presidente Prudente: [s.n.], 2015. Citado 2 vezes nas páginas 45 e 54.

MAZUCHELI, J.; MENEZES, A. F. B.; GHITANY, M. E. The unit-weibull distribution and associated inference. *Journal of Applied Probability and Statistics*, 2018. Citado 6 vezes nas páginas 13, 15, 16, 18, 19 e 42.

MBAH, A. K. *On the theory of records and applications*. [S.l.]: University of South Florida, 2007. Citado na página 59.

MCCULLAGH, P.; NELDER, J. *Generalized Linear Models*. Second. [S.l.]: Chapman and Hall, 1989. Citado na página 23.

MEYLAN, P.; FAVRE, A.; MUSY, A. *Predictive Hydrology: A Frequency Analysis Approach*. [S.l.]: CRC Press, 2011. Citado na página 45.

MITNIK, P. A.; BAEK, S. The kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation. *Statistical Papers*, Springer, p. 1–16, 2013. Citado 5 vezes nas páginas 15, 16, 18, 22 e 26.

MOUSA, A. M.; EL-SHEIKH, A. A.; ABDEL-FATTAH, M. A. A gamma regression for bounded continuous variables. *Advances and Applications in Statistics*, Pushpa Publishing House, v. 49, n. 4, p. 305, 2016. Citado na página 18.

MOUSA, M. A.; JAHEEN, Z.; AHMAD, A. Bayesian estimation, prediction and characterization for the gumbel model based on records. *Statistics: A Journal of Theoretical and Applied Statistics*, Taylor & Francis, v. 36, n. 1, p. 65–74, 2002. Citado 3 vezes nas páginas 61, 62 e 63.

MUBARAK, M. Estimation of the frèchet distribution parameters based on record values. *Arabian Journal for Science and Engineering*, Springer, v. 36, n. 8, p. 1597–1606, 2011. Citado na página 62.

NADAR, M.; PAPADOPOULOS, A. Bayesian analysis for the burr type xii distribution based on record values. *Statistica*, Università degli Studi di Bologna, Department of Statistical Sciences, Alma Mater Studiorum, v. 71, n. 4, p. 421, 2011. Citado 2 vezes nas páginas 62 e 63.

NADARAJAH, S. The exponentiated gumbel distribution with climate application. *Environmetrics*, Wiley Online Library, v. 17, n. 1, p. 13–23, 2006. Citado na página 66.

NAGARAJA, H. Record values and related statistics-a review. *Communications in Statistics-Theory and Methods*, Taylor & Francis, v. 17, n. 7, p. 2223–2238, 1988. Citado na página 59.

NAKAGAMI, M. The m-distribution-a general formula of intensity distribution of rapid fading. *Statistical Method of Radio Propagation*, Pergamon Press, p. 3–34, 1960. Citado na página 45.

NELDER, J. A.; BAKER, R. J. *Generalized linear models*. [S.l.]: Wiley Online Library, 1972. Citado na página 17.

PANAITESCU, E.; POPESCU, P. G.; COZMA, P.; POPA, M. Bayesian and non-bayesian estimators using record statistics of the modified-inverse weibull distribution. *Proceedings of the Romanian Academy, Series A*, v. 11, n. 3, p. 224–231, 2010. Citado 2 vezes nas páginas 62 e 63.

PAULA, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2004. Citado 2 vezes nas páginas 17 e 25.

PINHEIRO, E. C. *Contribuições em inferência e modelagem de valores extremos*. Tese (Doutorado) — Universidade de São Paulo, 2013. Citado na página 67.

PITT, M. A.; MYUNG, J. When a good fit can be bad. *Trends in Cognitive Sciences*, v. 6, p. 421–425, 2002. Citado na página 46.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2016. Disponível em: <https://www.R-project.org/>. Citado na página 50.

RAMOS, P. L.; MOALA, F. A. The extended geometric exponential distribution applied for modeling rainfall data. *Revista Brasileira de Meteorologia*, SciELO Brasil, v. 29, n. 4, p. 613–620, 2014. Citado na página 54.

RAQAB, M. Z. Distribution-free prediction intervals for the future current record statistics. *Statistical Papers*, Springer, v. 50, n. 2, p. 429–439, 2009. Citado na página 63.

RODRIGUES, J. A.; FILHO, J. dos S.; CHAVES, L. M. Funções densidade de probabilidade para a estimativa de precipitação mensal. *Semina: Ciências Exatas e Tecnológicas*, v. 34, n. 1, p. 3–8, 2013. Citado na página 54.

SALSBURG, D. S. *Uma senhora toma chá... como a estatística revolucionou a ciência no século XX*. [S.l.]: Zahar, 2009. Citado na página 12.

SANTOS, B.; BOLFARINE, H. Bayesian analysis for zero-or-one inflated proportion data using quantile regression. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 85, n. 17, p. 3579–3593, 2015. Citado na página 22.

SANTOS, W. d. O.; MESQUITA, F. d. O.; BATISTA, B. D. d. O.; BATISTA, R. O.; ALVES, A. d. S. Precipitações máximas para o município de mossoró de 1964 a 2011 pela distribuição de gumbel. *Irriga*, v. 19, n. 2, p. 207, 2014. Citado na página 66.

SAS. *User's Guide: The NLMIXED Procedure*. Cary, NC, 2010. v. 9.22, 4967–5062 p. Citado na página 27.

SCHMIT, J. T.; ROTH, K. Cost effectiveness of risk management practices. *Journal of Risk and Insurance*, JSTOR, p. 455–470, 1990. Citado 2 vezes nas páginas 27 e 28.

SCHULTHEIS, H.; NAIDU, P. Multi-model comparison using the cross-fitting method. *COGSCI*, p. 1389–1394, 2014. Citado 3 vezes nas páginas 46, 47 e 56.

SCHULTHEIS, H.; SINGHANIYA, A. Decision criteria for model comparison using cross-fitting. *Cognitive Systems Research*, v. 33, p. 100–121, 2013. Citado na página 47.

SCHWARTZ, J.; GODWIN, R. T.; GILES, D. E. Improved maximum-likelihood estimation of the shape parameter in the nakagami distribution. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 83, n. 3, p. 434–445, 2013. Citado na página 45.

SCHWARZ, G. et al. Estimating the dimension of a model. *The annals of statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. Citado na página 26.

SILVA, Í. N.; OLIVEIRA, J. B. de; FONTES, L. de O.; ARRAES, F. D. Distribuic? o de frequencia da chuva para regi? o centro-sul do ceara, brasil. *Revista Ciência Agronômica*, Universidade Federal do Ceará, Centro de Ciências Agrárias, v. 44, n. 3, p. 481, 2013. Citado na página 54.

SINGH, K.; SARKAR, S. Development of giuh for the catchment contributing to loktak lake, north east india. *Journal of the Indian Society of Remote Sensing*, Springer, v. 41, n. 2, p. 447–459, 2013. Citado na página 45.

SOLIMAN, A. A.; ELLAH, A. A.; SULTAN, K. Comparison of estimates using record statistics from weibull model: Bayesian and non-bayesian approaches. *Computational Statistics & Data Analysis*, Elsevier, v. 51, n. 3, p. 2065–2077, 2006. Citado 3 vezes nas páginas 61, 63 e 64.

SONG, P. X.-K.; TAN, M. Marginal models for longitudinal continuous proportional data. *Biometrics*, Wiley Online Library, v. 56, n. 2, p. 496–502, 2000. Citado na página 18.

TADIKAMALLA, P. R. On a family of distributions obtained by the transformation of the Gamma distribution. *Journal of Statistical Computation and Simulation*, v. 13, n. 3–4, p. 209–214, 1981. Citado na página 18.

TADIKAMALLA, P. R.; JOHNSON, N. L. Systems of frequency curves generated by transformations of logistic variables. *Biometrika*, [Oxford University Press, Biometrika Trust], v. 69, n. 2, p. 461–465, 1982. ISSN 00063444. Citado na página 18.

TEIMOURI, M.; NADARAJAH, S. Bias corrected MLEs for the Weibull distribution based on records. *Statistical Methodology*, v. 13, p. 12–24, 2013. Citado 2 vezes nas páginas 61 e 64.

TOPP, C. W.; LEONE, F. C. A family of J-Shaped frequency functions. *Journal of the American Statistical Association*, v. 50, n. 269, p. 209–219, 1955. Citado na página 18.

VUONG, Q. H. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, JSTOR, p. 307–333, 1989. Citado 2 vezes nas páginas 26 e 27.

WAGENMAKERS, E.; RATCLIFF, R.; GOMEZ, P.; IVERSONC, G. Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, v. 48, p. 28–50, 2004. Citado 4 vezes nas páginas 14, 46, 47 e 55.

WANG, B. X.; WANG, X. K.; YU, K. Inference on the Kumaraswamy distribution. *Communications in Statistics - Theory and Methods*, v. 46, n. 5, p. 2079–2090, 2017. Citado 2 vezes nas páginas 62 e 64.

WEIBULL, W. A statistical distribution of wide applicability. *Journal of Applied Mechanics*, v. 18, n. 3, p. 293–297, 1951. Citado na página 19.

WEISBERG, S. *Applied linear regression*. [S.l.]: John Wiley & Sons, 2005. v. 528. Citado na página 17.

WERGEN, G. *Records statistics beyond the standard model-Theory and applications*. Tese (Doutorado) — Universität zu Köln, 2012. Citado na página 58.

WOOLDRIDGE, J. M. Quasi-likelihood methods for count data. *Handbook of applied econometrics*, Blackwell Oxford, v. 2, p. 352–406, 1997. Citado na página 18.

ZAKERZADEH, H.; JAFARI, A. Inference on the parameters of two weibull distributions based on record values. *Statistical Methods & Applications*, Springer, v. 24, n. 1, p. 25–40, 2015. Citado 2 vezes nas páginas 61 e 64.

ZHANG, P.; QIU, Z.; SHI, C. simplexreg: An R package for regression analysis of proportional data using the simplex distribution. *Journal of Statistical Software, Articles*, v. 71, n. 11, p. 1–21, 2016. ISSN 1548-7660. Citado na página 33.

ZHAO, Y.; LEE, A. H.; YAU, K. K. W.; MCLACHLAN, G. J. Assessing the adequacy of Weibull survival models: A simulated envelope approach. *Journal of Applied Statistics*, v. 38, n. 10, p. 2089–2097, 2011. Citado na página 25.