UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação
ICMC

Distribuições combinadas

Tainá Santana Caldas

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Tainá Santana Caldas

Distribuições combinadas

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

Área de Concentração: Estatística

Orientador: Prof. Dr. Carlos Alberto Ribeiro Diniz

USP – São Carlos
Maio de 2021

Tainá Santana Caldas

Combined distributions

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

Concentration Area: Statistics

Advisor: Prof. Dr. Carlos Alberto Ribeiro Diniz

USP – São Carlos
May 2021

UNIVERSIDADE FEDERAL DE SÃO CARLOS

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

Folha de Aprovação

Defesa de Dissertação de Mestrado da candidata Tainá Santana Caldas, realizada em 21/04/2021.

Comissão Julgadora:

Prof. Dr. Carlos Alberto Ribeiro Diniz (UFSCar)

Profa. Dra. Carolina Costa Mota Paraíba (UFBA)

Prof. Dr. Marcio Luis Lanfredi Viola (UFSCar)

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

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

AGRADECIMENTOS

Aos meus pais, Silas e Helena, pela paciência, espera, apoio e orações;

Ao prof. Dr. Carlos Diniz, pela orientação e pelo tema da pesquisa;

A todos os professores e colegas do PIPGES;

À Primeira Igreja Batista, minha família em São Carlos-SP.

“Porque dele, por ele e para ele são todas as coisas.”

(Romanos 11:36, Bíblia Sagrada)

RESUMO

CALDAS, T. S. **Distribuições combinadas**. 2021. 103 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Propõe-se uma possível nova linha de pesquisa em que se combinam distribuições. Distribuições combinadas são uma alternativa para ajuste de dados contínuos que apresentam comportamentos probabilísticos distintos. Desenvolve-se uma distribuição combinada denominada Gucumatz, composta pelas distribuições Normal e Cauchy simétricas. Estuda-se algumas propriedades desta distribuição, realiza-se estimação por máxima verossimilhança de dados aumentados (Algoritmo EM), desenvolve-se modelo de regressão e estudos de simulações. Um conjunto de dados reais é ajustado com utilização da distribuição Gucumatz.

Palavras-chave: Distribuição Gucumatz, Algoritmo EM, Distribuição Normal, Distribuição Cauchy.

ABSTRACT

CALDAS, T. S. **Combined distributions**. 2021. 103 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

It is proposed a possible new line of research in which distributions are combined. Combined distributions are an alternative for adjusting continuous data that present different probabilistic behaviors. A combined distribution called Gucumatz is developed, composed of the symmetrical Normal and Cauchy distributions. Some properties of this distribution are studied, maximum likelihood estimation of augmented data is performed (EM Algorithm), regression model and simulation studies are developed. A set of real data is adjusted using the Gucumatz distribution.

Keywords: Gucumatz distribution, EM Algorithm, Normal distribution, Cauchy distribution.

LISTA DE ILUSTRAÇÕES

LISTA DE ALGORITMOS

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Em análise de dados cujo comportamento indica a presença de duas ou mais distribuições de probabilidade, é comum se utilizar mistura de distribuições. Uma mistura de distribuições representa a presença de subpopulações não identificadas diretamente nos dados observados. Uma variável aleatória $Y \in \mathbb{R}$ que segue uma distribuição de mistura composta pelas funções de distribuição $F_1(.)$ e $F_2(.)$, dada uma variável aleatória não-observável K tal que $K = 1$ se $Y \sim F_1(y)$ e $K = 2$ se $Y \sim F_2(y)$, e $K \sim Binomial(\tau)$, tem função de distribuição de probabilidade dada por $F(y) = \tau F_1(y) + (1-\tau)F_2(y), \quad 0 \leq \tau \leq 1$.

No entanto, dados com tais características podem indicar a presença de combinação de distribuições, ao invés de mistura. Considere um conjunto de dados disponibilizados por Erisen (2019), de geração de energia eólica coletados em uma região da Turquia durante o ano de 2018, com variável resposta "energia gerada pela turbina"e covariável "direção do vento".

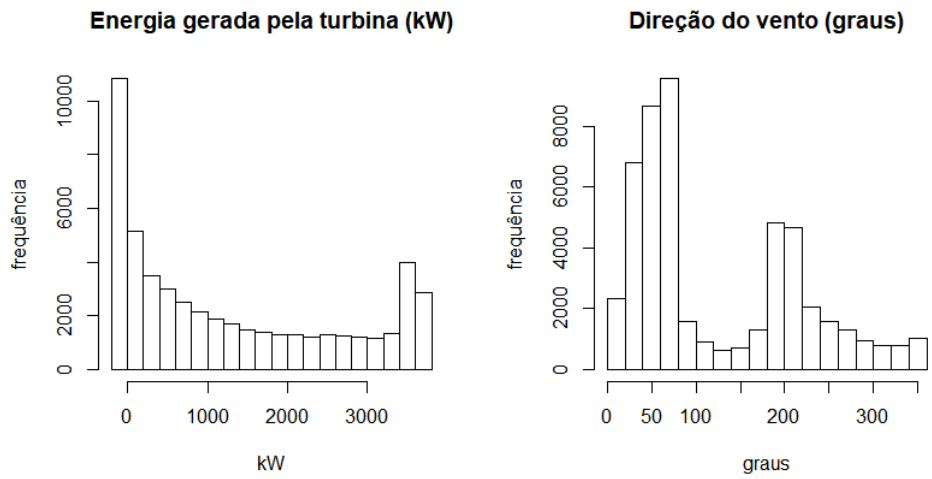


Figura 1 – Histogramas dos dados de energia gerada pela turbina e direção do vento, de um conjunto de dados de geração de energia eólica coletados em uma região da Turquia durante o ano de 2018

Nos histogramas na Figura 1, observa-se que há duas modas, entre as quais não há suavidade, e os dados apresentam diferentes comportamentos, de distribuições de probabilidade gama, uniforme e normal que não se misturam, mas se combinam de forma articulada, lado a lado, e não sobrepostas. Trata-se de uma possível combinação de distribuições, que dá origem a uma distribuição combinada.

Distribuições combinadas são definidas por Kubo *et al.* (1991) como a seguir, com adaptações, para possibilitar melhor compreensão. Seja uma variável aleatória $Y \in \mathbb{R}$. Se Y segue uma distribuição combinada, a sua função densidade de probabilidade é dada por:

$$f(y) = \begin{cases} q_1 f_1(y) & (-\infty < y < y_0) \\ q_2 f_2(y) & (y_0 \leq y < \infty) \end{cases}, \quad (0 < q_1, q_2).$$

O ponto y_0 pode ser chamado de ponto de fronteira entre as distribuições de probabilidades componentes $f_1(.)$ e $f_2(.)$.

Os escalares q_1 e q_2 não são pesos entre as funções densidades $f_1(.)$ e $f_2(.)$, isto é, $q_1 + q_2 \neq 1$; os escalares q_1 e q_2 tem como papel redimensionar, ou multiplicar as funções densidades $f_1(.)$ e $f_2(.)$ de modo que $f(.)$ seja uniformemente contínua e tenha probabilidade igual a 1 em $\mathbb{R}$.

A função densidade $f(.)$ é uniformemente contínua, ou seja:

$$q_1 f_1(y_0) = q_2 f_2(y_0). \tag{1.1}$$

A função de distribuição acumulada de Y é dada por:

$$F(y) = \int_{-\infty}^{y} f(t)dt = \begin{cases} q_1 F_1(y) & se \quad y \in (-\infty, y_0) \\ q_1 F_1(y_0) + q_2 F_2(y) & se \quad y \in [y_0, \infty) \end{cases}.$$

A variável Y tem probabilidade igual a 1 em $\mathbb{R}$, isto é, $\int_{-\infty}^{\infty} f(y)dy = 1$. De fato,

$$\int_{-\infty}^{\infty} f(y)dy = \int_{-\infty}^{y_0} q_1 f_1(y)dy + \int_{y_0}^{\infty} q_2 f_2(y)dy =$$
$$q_1[F_1(y_0) - \lim_{y \to -\infty} F_1(y)] + q_2[\lim_{y \to \infty} F_2(y) - F_2(y_0)] = q_1[F_1(y_0)] + q_2[1 - F_2(y_0)] = 1.$$

Em conformidade com a definição de Kubo *et al.* (1991), uma distribuição combinada denominada Gucumatz (ROGERS; TUKEY, 1972) é apresentada a seguir, com adaptações para possibilitar melhor compreensão. A distribuição Gucumatz é composta pela distribuição Normal padrão no intervalo $[-1, 1]$ e distribuição Cauchy padrão nas caudas, fora do intervalo $[-1, 1]$. As probabilidades Normal e Cauchy são multiplicadas pelos escalares 0,6930665 e 1,0537015, respectivamente, para que a probabilidade da distribuição Gucumatz em $\mathbb{R}$ seja igual a 1 e a

função densidade seja uniformemente contínua. Seja uma variável aleatória $Y \in \mathbb{R}$. A função densidade Gucumatz é dada por:

$$f(y) = \begin{cases} 0,6930665 \cdot \dfrac{e^{\frac{-y^2}{2}}}{\sqrt{2\pi}}, & se \quad -1 \leq y \leq 1 \\ \dfrac{1,0537015}{\pi(1+y^2)}, & c.c. \end{cases} . \tag{1.2}$$

Os pontos -1 e 1 são pontos de fronteira entre as funções densidades das distribuições componentes Normal e Cauchy.

A variável $Y \sim$ Gucumatz apresenta probabilidade igual a 1 em $\mathbb{R}$. De fato:

$$\begin{aligned} \int_{-\infty}^{\infty} f(y)dy &= \int_{-\infty}^{-1} \frac{1,0537015}{\pi(1+y^2)}dy + \int_{-1}^{1} 0,6930665 \cdot \frac{e^{\frac{-y^2}{2}}}{\sqrt{2\pi}}dy + \int_{1}^{\infty} \frac{1,0537015}{\pi(1+y^2)}dy \\ &= 1,0537015(0,25) + 0,6930665(0,6826895) + 1,0537015 \cdot 0,25 = 1. \end{aligned}$$

A distribuição Gucumatz (ROGERS; TUKEY, 1972) não tem parâmetros desconhecidos; todos os parâmetros são pré-determinados, dos quais, 0 e 1 são provenientes das distribuições componentes Normal padrão e Cauchy padrão. Portanto, pode-se denotá-la como Gucumatz (0,1), ou Gucumatz padrão.

A curva da função densidade Gucumatz (0,1) é visualizada juntamente com as curvas das funções densidades Normal padrão e Cauchy padrão, na Figura 2.

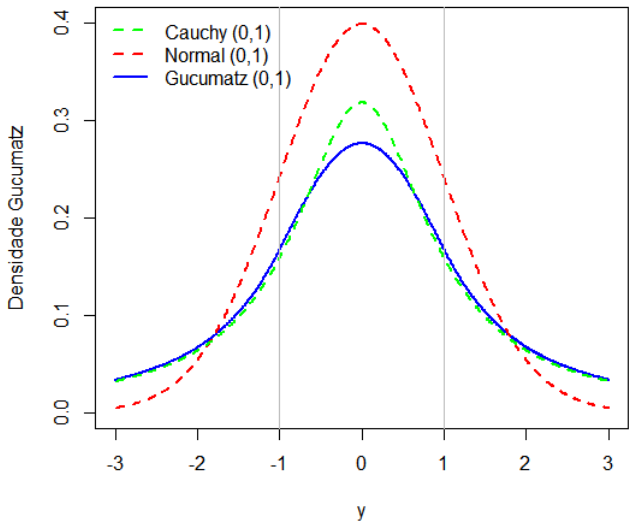


Figura 2 – Comparação entre os comportamentos das funções densidades Gucumatz, Normal padrão e Cauchy padrão.

Pela Figura 2, nota-se que a distribuição Gucumatz (0,1) está muito próxima à distribuição Cauchy padrão. Mas, visto que a distribuição Gucumatz (0,1) Rogers e Tukey (1972) apresenta todos os parâmetros pré-determinados, este trabalho pretende expandir a distribuição Gucumatz.

1.1 Objetivos

O objetivo geral deste trabalho é apontar para uma possível nova linha de pesquisa em distribuições combinadas. O objetivo específico é expandir e desenvolver a distribuição combinada Gucumatz, ou nova distribuição Gucumatz, como extensão daquela apresentada no trabalho de Rogers e Tukey (1972), que se torna caso particular desta. Os passos envolvem estabelecer parâmetros desconhecidos e estimá-los, apresentar algumas propriedades, construir modelo de regressão e aplicar o modelo proposto em dados reais e comparar modelos ajustados via distribuição Gucumatz com modelos ajustados às distribuições Normal e Cauchy. A motivação deste estudo está na utilidade da distribuição Gucumatz como alternativa para modelar dados contínuos que apresentam caudas pesadas e em incentivar pesquisas sobre novas distribuições combinadas.

Visto que a distribuição Gucumatz é composta por frações das distribuições Normal e Cauchy, são apresentados conceitos fundamentais sobre estas distribuições nas Seções 1.2 e 1.3.

1.2 Distribuição Normal

Seja $Y \in \mathbb{R}$ uma variável aleatória com distribuição Normal, cujos parâmetros são $\mu \in \mathbb{R}$ e $\sigma^2 > 0$. A função densidade de probabilidade de Y é dada por:

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-(y-\mu)^2/2\sigma^2\}.$$

A esperança e variância de Y são:

$$E(Y) = \mu, \quad Var(Y) = \sigma^2.$$

A função densidade da distribuição Normal é simétrica ao redor da média μ e apresenta formato de sino.

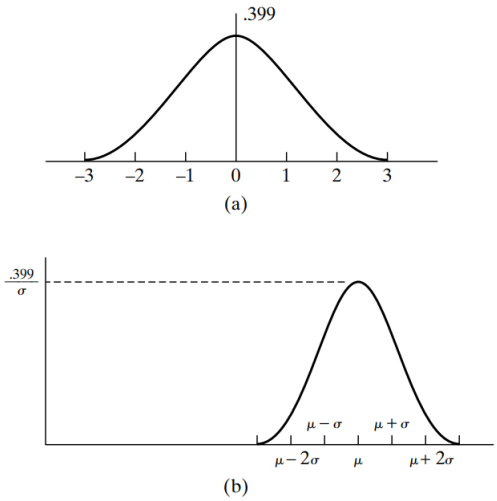


Figura 3 – Função densidade de probabilidade normal (Extraído de Ross (2010)).

Nota-se que quase toda a área sob a curva Normal está situada no intervalo $(\mu - 3\sigma, \mu + 3\sigma)$.

Uma variável aleatória $Z = (Y - \mu)/\sigma$ tem distribuição Normal (0,1), também denominada por distribuição Normal padrão.

A função de distribuição acumulada Normal é, usualmente, denotada por $\Phi(y)$. A sua expressão considera uma variável aleatória Normal padronizada $Z = (Y - \mu)/\sigma \sim N(0,1)$. Assim,

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{z} e^{-t^2/2} dt.$$

Sejam $Y_i \in \mathbb{R}$ $(i = 1, 2, ..., n)$ variáveis aleatórias independentes que seguem a distribuição Normal(μ_i, σ^2), com realizações $y_i \in \mathbb{R}$. Seja a matriz de planejamento $\boldsymbol{X}$, cujos vetores linhas $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, ..., x_{ip})^T$ correspondem aos valores observados das p covariáveis associadas ao i-ésimo indivíduo. Seja o vetor de parâmetros desconhecidos $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$, em que $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ tem dimensão $(p+1)$x1. O modelo de regressão linear múltipla, amplamente conhecido, é dado por:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad 1 \leq i \leq n.$$

$$\mu_i = E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}.$$

A função de verossimilhança do vetor de parâmetros $\boldsymbol{\theta}$, dadas as observações, é escrita como:

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n} \log \left\{ \frac{1}{\sqrt{2\pi\sigma}} \exp[-(y_i - \mu)^2/2\sigma^2] \right\},$$

a qual é maximizada nos parâmetros.

1.3 Distribuição de Cauchy

Seja $Y \in \mathbb{R}$ uma variável aleatória com distribuição de Cauchy, cujos parâmetros são $\mu \in \mathbb{R}$, a mediana de Y, e $\sigma > 0$. A função densidade de probabilidade de Y é dada por:

$$f(y) = \frac{1}{\pi\sigma\left\{1+\left(\frac{y-\mu}{\sigma}\right)^2\right\}}.$$

A função de distribuição acumulada é dada por:

$$F(y) = \frac{1}{2} + \frac{1}{\pi}\arctan(y), \quad y \in \mathbb{R}.$$

A esperança na distribuição de Cauchy não existe, isto é, $E|Y| = \infty$, e não existe nenhum momento, pois todos os momentos absolutos são iguais a ∞. (CASELLA; BERGER, 2016).

Pode-se definir um modelo de regressão Cauchy como a seguir. Sejam $Y_i \in \mathbb{R}$ $(i = 1,2,...,n)$ variáveis aleatórias independentes que seguem a distribuição Cauchy(μ_i, σ), com realizações $y_i \in \mathbb{R}$. Seja a matriz de planejamento $\boldsymbol{X}$, cujos vetores linhas $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, ..., x_{ip})^T$ correspondem aos valores observados das p covariáveis associadas ao i-ésimo indivíduo. Seja o vetor de parâmetros desconhecidos $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$, em que $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ tem dimensão $(p+1)$x1. O modelo de regressão Cauchy pode ser definido por meio da seguinte relação funcional:

$$\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}, \quad 1 \leq i \leq n.$$

A função de verossimilhança do vetor de parâmetros $\boldsymbol{\theta}$, dadas as observações, é escrita como:

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = -\sum_{i=1}^{n} \log\left\{\pi\sigma\left[1+\left(\frac{y-\mu}{\sigma}\right)^2\right]\right\},$$

a qual é maximizada nos parâmetros.

1.4 Organização do trabalho

No Capítulo 1 apresentou-se motivações e fundamentos teóricos para utilização de distribuições combinadas, bem como para o desenvolvimento da distribuição Gucumatz.

No Capítulo 2, apresenta-se a distribuição Gucumatz expandida, na qual são desenvolvidas funções densidade e de distribuição acumulada, apresentam-se algumas propriedades, função

quantil, algoritmo para simulação de dados e o parâmetro de dispersão quantílica.

No Capítulo 3, apresenta-se estimação por máxima verossimilhança, intervalos de confiança e estudo de simulação; em seguida apresenta-se um caminho para a estimação bayesiana nesta distribuição.

No Capítulo 4, apresenta-se o modelo de regressão Gucumatz, realiza-se estimação frequentista, apresentam-se fundamentos teóricos para realização de inferências, testes de hipóteses, seleção de modelos e diagnósticos de modelos, e são realizados estudos de simulação. Por fim, um conjunto de dados reais é ajustado ao modelo de regressão Gucumatz.

No Capítulo 5, apresenta-se uma alternativa de estimação, em que um parâmetro da distribuição Gucumatz é pré-determinado.

CAPÍTULO

2

DISTRIBUIÇÃO GUCUMATZ

Neste capítulo, desenvolve-se a distribuição Gucumatz com parâmetros desconhecidos por meio de um exercício indutivo de generalização dos parâmetros, de modo que a distribuição Gucumatz apresentada em (1.2) torna-se caso particular desta. Em seguida, são apresentadas algumas propriedades desta distribuição e o parâmetro de dispersão.

Sejam as distribuições Normal (μ_1, σ_1^2) e Cauchy (μ_2, σ_2), componentes da distribuição Gucumatz. Por simplicidade, pode-se considerar $\mu = \mu_1 = \mu_2$ e $\sigma = \sigma_1 = \sigma_2$ como parâmetros da distribuição Gucumatz, de modo que as suas distribuições componentes sejam Normal (μ, σ^2) e Cauchy (μ, σ). Os pontos de fronteira, outrora apresentados como -1 e 1, podem ser reescritos como pontos desconhecidos simétricos ao redor do parâmetro μ, como $\mu - \sigma$ e $\mu + \sigma$. Assim, a distribuição componente Normal (μ, σ^2) é definida no intervalo $[\mu - \sigma, \mu + \sigma]$, enquanto que a componente Cauchy (μ, σ) é definida fora deste intervalo. Os parâmetros μ e σ estabelecidos até este momento são provenientes das distribuições Normal e Cauchy. Por sua vez, os escalares que multiplicam as funções densidades componentes Normal e Cauchy são reconhecidos como novos parâmetros α_1 e α_2, próprios da distribuição Gucumatz.

Devido a necessidade de conferir maior flexibilidade àquela distribuição Gucumatz apresentada no capítulo 1, adiciona-se mais um novo parâmetro λ na reta suporte, de modo que os pontos de fronteira passam a ser $\mu - \lambda\sigma$ e $\mu + \lambda\sigma$.

Portanto, a distribuição Gucumatz, com parâmetros $\mu, \sigma, \lambda, \alpha_1$ e α_2, é composta pelas distribuições Normal (μ, σ^2), no intervalo $[\mu - \lambda\sigma, \mu + \lambda\sigma]$ e Cauchy (μ, σ) em $\mathbb{R} - [\mu - \lambda\sigma, \mu + \lambda\sigma]$, fracionadas pelos escalares α_1 e α_2 respectivamente, para que a distribuição Gucumatz tenha probabilidade igual a 1 em $\mathbb{R}$ e seja uniformemente contínua.

Nas seções seguintes, são desenvolvidas a função densidade de probabilidade, a função de distribuição de probabilidade acumulada, e algumas propriedades.

2.1 Função densidade de probabilidade

Seja a variável aleatória $Y \in \mathbb{R}$. A função densidade de probabilidade Gucumatz com parâmetros $\mu \in \mathbb{R}, \sigma > 0$, $\lambda > 0$, α_1, α_2 é definida por:

$$f(y) = \begin{cases} \dfrac{\alpha_1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right], & se \quad \mu - \lambda\sigma \leq y \leq \mu + \lambda\sigma \\ \dfrac{\alpha_2}{\sigma\pi\left[1+\left(\frac{y-\mu}{\sigma}\right)^2\right]}, & c.c. \end{cases} . \tag{2.1}$$

A distribuição Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$ satisfaz as seguintes condições:

- A probabilidade desta distribuição no domínio $\mathbb{R}$ é igual a 1, isto é, $\int_{\infty}^{\infty} f(y)dy = 1$;
- A função densidade de probabilidade é uniformemente contínua, isto é, sem saltos nos pontos de fronteira $\mu - \lambda\sigma$ e $\mu + \lambda\sigma$,

condições verificadas por meio das Proposições 1 e 2, cujos resultados determinam os valores dos parâmetros α_1 e α_2.

Proposição 1. A função densidade $f(y)$ da distribuição Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$ tem probabilidade igual a 1 em $\mathbb{R}$ se, e somente se, $\alpha_2 = \dfrac{1 - \alpha_1[\Phi(\lambda) - \Phi(-\lambda)]}{1 + \frac{1}{\pi}[\arctan(-\lambda) - \arctan(\lambda)]}$.

Demonstração. $\displaystyle\int_{-\infty}^{\infty} f(y|\mu,\sigma)dy =$

$$= \int_{-\infty}^{\mu-\lambda\sigma} \frac{\alpha_2}{\sigma\pi\left[1+\left(\frac{y-\mu}{\sigma}\right)^2\right]}dy + \int_{\mu-\lambda\sigma}^{\mu+\lambda\sigma} \alpha_1 \cdot \frac{\exp\left[\frac{-1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]}{\sigma\sqrt{2\pi}}dy + \int_{\mu+\lambda\sigma}^{\infty} \frac{\alpha_2}{\sigma\pi\left[1+\left(\frac{y-\mu}{\sigma}\right)^2\right]}dy$$

$$= \alpha_2 \cdot \left[\frac{1}{\pi}\arctan(-\lambda) + \frac{1}{2}\right] + \alpha_1 \cdot [\Phi(\lambda) - \Phi(-\lambda)] + \alpha_2 \cdot \left[\frac{1}{2} - \frac{1}{\pi}\arctan(\lambda)\right] =$$

$$= \alpha_2 \cdot \left\{1 + \frac{1}{\pi}[\arctan(-\lambda) - \arctan(\lambda)]\right\} + \alpha_1[\Phi(\lambda) - \Phi(-\lambda)] = 1 \Leftrightarrow$$

$$\alpha_2 = \frac{1 - \alpha_1[\Phi(\lambda) - \Phi(-\lambda)]}{1 + \frac{1}{\pi}[\arctan(-\lambda) - \arctan(\lambda)]}. \tag{2.2}$$

□

Proposição 2. A função densidade $f(y)$ da distribuição Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$ é uniformemente contínua em $\mathbb{R}$ se, e somente se, $\alpha_2 = \dfrac{\alpha_1 e^{\frac{-\lambda^2}{2}} \pi(1+\lambda^2)}{\sqrt{2\pi}}$.

Demonstração. Para verificar que a função densidade Gucumatz é contínua nos pontos de fronteira $(y = \mu - \lambda\sigma)$ e $(y = \mu + \lambda\sigma)$, os limites à esquerda e direita do ponto $(y = \mu - \lambda\sigma)$ deverão ser iguais, e da mesma forma para o ponto $(y = \mu + \lambda\sigma)$, conforme (1.1).

Nos pontos $(y = \mu - \lambda\sigma)$ e $(y = \mu + \lambda\sigma)$:

$$\alpha_2 \cdot \frac{1}{\sigma\pi\left[1+\left(\frac{(\mu\pm\lambda\sigma)-\mu}{\sigma}\right)^2\right]} = \alpha_1 \cdot \frac{\exp\left[\frac{-1}{2}\left(\frac{(\mu\pm\lambda\sigma)-\mu}{\sigma}\right)^2\right]}{\sigma\sqrt{2\pi}} \Leftrightarrow$$

$$\alpha_2 \cdot \frac{1}{\sigma\pi\left[1+(\pm\lambda)^2\right]} = \alpha_1 \cdot \frac{\exp\left[\frac{-1}{2}\cdot(\pm\lambda)^2\right]}{\sigma\sqrt{2\pi}} \Leftrightarrow$$

$$\alpha_2 \cdot \frac{1}{\sigma\pi\left[1+\lambda^2\right]} = \alpha_1 \cdot \frac{\exp\left[\frac{-\lambda^2}{2}\right]}{\sigma\sqrt{2\pi}} \quad \Leftrightarrow$$

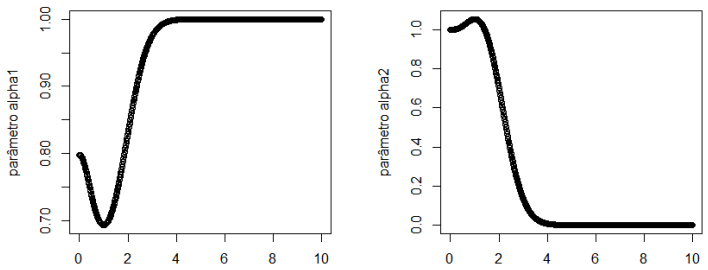
$$\alpha_2 = \frac{\alpha_1 \cdot e^{\frac{-\lambda^2}{2}}\pi(1+\lambda^2)}{\sqrt{2\pi}}. \tag{2.3}$$

□

De (2.2) e (2.3), os parâmetros α_1 e α_2 são escritos como

$$\alpha_1 = \left\{\frac{e^{\frac{-\lambda^2}{2}}(1+\lambda^2)}{\sqrt{2\pi}}\left[\pi - 2\arctan(\lambda)\right] + 2\Phi(\lambda) - 1\right\}^{-1}, \quad \alpha_2 = \frac{1-\alpha_1(2\Phi(\lambda)-1)}{1-\frac{2}{\pi}\arctan(\lambda)}. \tag{2.4}$$

Os comportamentos dos parâmetros α_1 e α_2 em função do parâmetro λ, podem ser vistos na Figura 4.



(a) α_1 em função de λ (b) α_2 em função de λ

Figura 4 – (a) comportamento do parâmetro α_1 em função do parâmetro λ, (b) comportamento do parâmetro α_2 em função do parâmetro λ.

Pela Figura 4a, α_1 apresenta ponto mínimo $\alpha_1 = 0,693$, quando $\lambda = 1$, e tende a 1 quando $\lambda > 4$. Pela Figura 4b, α_2 tende a 0, quando $\lambda > 4$, e apresenta ponto máximo $\alpha_2 = 1,054$, quando $\lambda = 1$. Nota-se que, quando $\lambda > 4$ a distribuição Gucumatz começa a assumir a característica de distribuição Normal, visto que α_2 tende a zero e α_1 tende a 1. Em virtude dos comportamento de α_1 e α_2 em função de λ e estimações realizados neste trabalho, entende-se que o modelo não é identificável quando $\lambda < 1$ e $\lambda > 4$. Portanto, $\alpha_1 \in (0,693;1)$ e $\alpha_2 \in (0;1,054)$.

Desta forma, a função densidade de probabilidade da distribuição Gucumatz com parâmetros $\mu \in \mathbb{R}, \sigma > 0$, $1 \leq \lambda \leq 4$, $\alpha_1 \in (0,693;1)$ e $\alpha_2 \in (0;1,054)$ é dada por

$$f(y) = \begin{cases} \dfrac{\alpha_1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right], & se \quad \mu - \lambda\sigma \leq y \leq \mu + \lambda\sigma \\ \dfrac{\alpha_2}{\sigma\pi\left[1+\left(\frac{y-\mu}{\sigma}\right)^2\right]}, & c.c. \end{cases}, \quad \text{em que}$$

$$\alpha_1 = \left\{ \frac{e^{\frac{-\lambda^2}{2}}(1+\lambda^2)}{\sqrt{2\pi}} [\pi - 2\arctan(\lambda)] + 2\Phi(\lambda) - 1 \right\}^{-1}, \quad \alpha_2 = \frac{1-\alpha_1(2\Phi(\lambda)-1)}{1-\frac{2}{\pi}\arctan(\lambda)}. \quad (2.5)$$

A probabilidade da distribuição Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$ no intervalo central $[\mu - \lambda\sigma, \mu + \lambda\sigma]$ é dada por

$$P(Y \in [\mu - \lambda\sigma, \mu + \lambda\sigma]) = \int_{\mu-\lambda\sigma}^{\mu+\lambda\sigma} \alpha_1 \frac{\exp\left[\frac{-1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]}{\sigma\sqrt{2\pi}} dy = \alpha_1 \cdot [\Phi(\lambda) - \Phi(-\lambda)]. \quad (2.6)$$

Seguem, como exemplo, algumas funções densidade de probabilidade da distribuição Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$ para alguns valores de parâmetro λ pré-determinados:

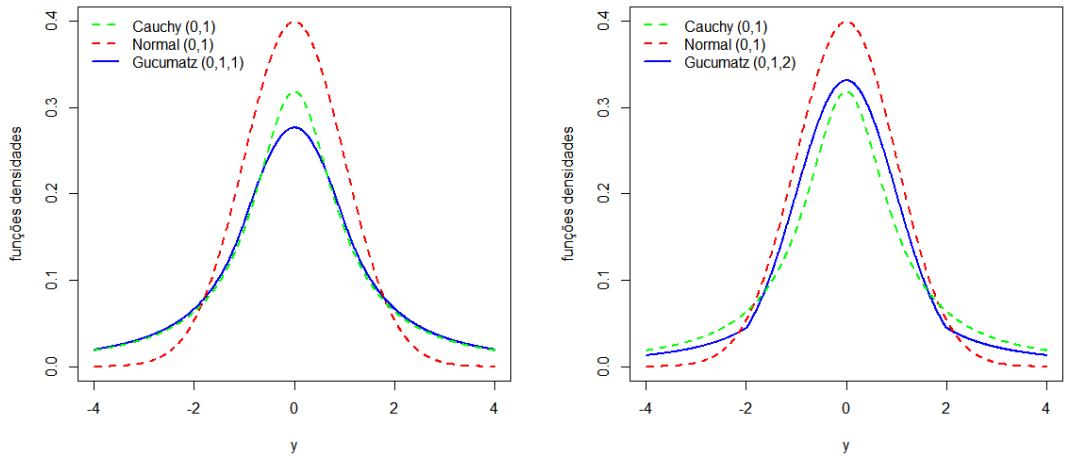
Função densidade de probabilidade da distribuição Gucumatz $(\mu, \sigma, \lambda = 1)$:

$$f(y) = \begin{cases} \dfrac{0,6930665}{\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right], & se \quad \mu - \sigma \leq y \leq \mu + \sigma \\ \dfrac{1,0537015}{\sigma\pi\left[1+\left(\frac{y-\mu}{\sigma}\right)^2\right]}, & c.c. \end{cases}.$$

Função densidade de probabilidade da distribuição Gucumatz $(\mu, \sigma, \lambda = 2)$:

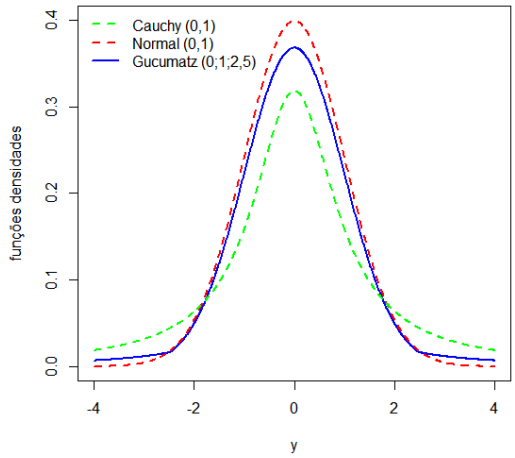
$$f(y) = \begin{cases} \dfrac{0,8299942}{\sigma\sqrt{2\pi}} \exp\left[\dfrac{-1}{2}\left(\dfrac{y-\mu}{\sigma}\right)^2\right], & se \quad \mu - 2\sigma \leq y \leq \mu + 2\sigma \\ \dfrac{0,7039082}{\sigma\pi\left[1+\left(\dfrac{y-\mu}{\sigma}\right)^2\right]}, & c.c. \end{cases} .$$

O comportamento da curva da função densidade de probabilidade da distribuição Gucumatz é comparado aos comportamentos das curvas das funções densidades das distribuições componentes. Para isto, são pré-determinados os parâmetros $\mu = 0$, $\sigma = 1$, enquanto o parâmetro λ varia, na Figura 5, em $\lambda = 1$ (Figura 5a), $\lambda = 2$ (Figura 5b), $\lambda = 2,5$ (Figura 5c) e $\lambda = 3$ (Figura 5d).

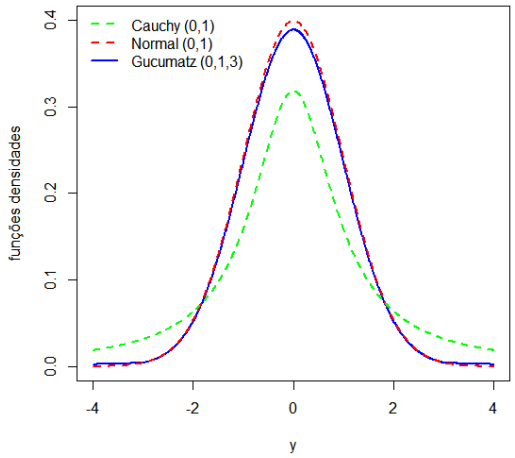


(a) $\lambda = 1$

(b) $\lambda = 2$



(c) $\lambda = 2,5$



(d) $\lambda = 3$

Figura 5 – Comparação entre as curvas das funções densidades de probabilidade da distribuição Gucumatz $(0,1,\lambda)$, Cauchy (0,1) e Normal (0,1) quando (a) $\lambda = 1$, (b) $\lambda = 2$, (c) $\lambda = 2,5$, (d) $\lambda = 3$.

Observa-se pelas Figuras 5b e 5c que quando $\lambda = 2$ ou $\lambda = 2,5$, a funções densidades das distribuições Gucumatz (0, 1, 2) e Gucumatz (0; 1; 2,5) se distinguem completamente das funções densidades das distribuições Normal (0,1) e Cauchy (0,1). A Figura 5a é igual à Figura 2 apresentada no Capítulo 1, que ilustra a função densidade da distribuição Gucumatz padrão, ou Gucumatz (0,1), que se distingue da densidade da distribuição Cauchy apenas no pico. Observa-se também, pela Figura 5d, que se $\lambda = 3$, a função densidade da distribuição Gucumatz (0, 1, 3) se aproxima quase completamente da densidade da distribuição Normal (0,1), apesar de apresentar caudas menos leves.

2.2 Função distribuição de probabilidade acumulada

Uma variável aleatória $Y \in \mathbb{R}$ com distribuição Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$ apresenta função distribuição de probabilidade acumulada $F : \mathbb{R} \to [0,1]$ dada por:

$$F(y) = P(Y \leq y) = \int_{-\infty}^{y} f(t)dt =$$

$$= \begin{cases} \alpha_2 \left[\frac{1}{\pi} \arctan\left(\frac{y-\mu}{\sigma} \right) + \frac{1}{2} \right], & \text{se} \quad y < \mu - \lambda\sigma \\ \alpha_2 \left[\frac{1}{\pi} \arctan(-\lambda) + \frac{1}{2} \right] + \alpha_1 \left[\Phi\left(\frac{y-\mu}{\sigma} \right) - \Phi(-\lambda) \right], & \text{se} \quad \mu - \lambda\sigma \leq y \leq \mu + \lambda\sigma \\ \alpha_2 \left[\frac{1}{\pi} \arctan(-\lambda) + \frac{1}{2} \right] + \alpha_1 \left[\Phi(\lambda) - \Phi(-\lambda) \right] + \frac{\alpha_2}{\pi} \left[\arctan\left(\frac{y-\mu}{\sigma} \right) - \arctan(\lambda) \right], \\ \text{se} \quad y > \mu + \lambda\sigma \end{cases}.$$

Usando $\arctan(-\lambda) = -\arctan(\lambda)$,

$$= \begin{cases} \alpha_2 \left[\frac{1}{\pi} \arctan\left(\frac{y-\mu}{\sigma} \right) + \frac{1}{2} \right], & \text{se} \quad y < \mu - \lambda\sigma \\ \alpha_2 \left[\frac{1}{\pi} \arctan(-\lambda) + \frac{1}{2} \right] + \alpha_1 \left[\Phi\left(\frac{y-\mu}{\sigma} \right) - \Phi(-\lambda) \right], & \text{se} \quad \mu - \lambda\sigma \leq y \leq \mu + \lambda\sigma \\ \alpha_2 \left[\frac{-2}{\pi} \arctan(\lambda) + \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{y-\mu}{\sigma} \right) \right] + \alpha_1 [\Phi(\lambda) - \Phi(-\lambda)], & \text{se} \quad y > \mu + \lambda\sigma \end{cases}. \tag{2.7}$$

Se $Y \sim$ Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$, a sua função distribuição de probabilidade acumulada apresenta três propriedades demonstradas nas Proposições 3, 4 e 5.

Proposição 3. A função de distribuição Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$ é não decrescente, isto é, $y_1 \leq y_2 \Rightarrow F(y_1) \leq F(y_2)$.

Demonstração. Sejam y_1 e y_2 valores quaisquer da variável aleatória $Y \in \mathbb{R}$, tais que $y_1 \leq y_2$. Da função de distribuição acumulada 2.7, tem-se:

Quando $y < \mu - \lambda\sigma$,

$$y_1 \leq y_2 \Rightarrow \alpha_2 \left[\frac{1}{\pi}\arctan\left(\frac{y_1-\mu}{\sigma}\right)+\frac{1}{2}\right] \leq \alpha_2 \left[\frac{1}{\pi}\arctan\left(\frac{y_2-\mu}{\sigma}\right)+\frac{1}{2}\right] \Leftrightarrow F(y_1) \leq F(y_2),$$

pois $\left[\frac{1}{\pi}\arctan\left(\frac{y-\mu}{\sigma}\right)+\frac{1}{2}\right]$ é uma função de distribuição Cauchy não decrescente.

Quando $\mu - \lambda\sigma \leq y \leq \mu + \lambda\sigma$,

$$y_1 \leq y_2 \Rightarrow \alpha_2 \left[\frac{1}{\pi}\arctan(-\lambda)+\frac{1}{2}\right] + \alpha_1 \left[\Phi\left(\frac{y_1-\mu}{\sigma}\right) - \Phi(-\lambda)\right] \leq$$

$$\alpha_2 \left[\frac{1}{\pi}\arctan(-\lambda)+\frac{1}{2}\right] + \alpha_1 \left[\Phi\left(\frac{y_2-\mu}{\sigma}\right) - \Phi(-\lambda)\right] \Rightarrow F(y_1) \leq F(y_2),$$

pois $\Phi(\frac{y-\mu}{\sigma})$ é função de distribuição acumulada normal padrão, não decrescente;

Quando $y > \mu + \lambda\sigma$,

$$y_1 \leq y_2 \Rightarrow \alpha_2 \left[\frac{-2}{\pi}\arctan(\lambda)+\frac{1}{2}+\frac{1}{\pi}\arctan\left(\frac{y_1-\mu}{\sigma}\right)\right] + \alpha_1[\Phi(\lambda)-\Phi(-\lambda)] \leq$$

$$\alpha_2 \left[\frac{-2}{\pi}\arctan(\lambda)+\frac{1}{2}+\frac{1}{\pi}\arctan\left(\frac{y_2-\mu}{\sigma}\right)\right] + \alpha_1[\Phi(\lambda)-\Phi(-\lambda)]$$

$\Rightarrow F(y_1) \leq F(y_2)$, pois $\arctan\left(\frac{y-\mu}{\sigma}\right)$ é uma função não decrescente e os demais termos se mantém e são não negativos. □

Proposição 4. A função de distribuição acumulada Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$ é contínua à direita, isto é, o limite de $F(y_n)$ quando y_n tende a y pela direita é igual a $F(y)$, ou seja, $\lim_{y_n \downarrow y^+} F(y_n) = F(y)$.

Demonstração. Da função de distribuição acumulada Gucumatz (2.7), mostra-se que $\lim_{y_n \downarrow y^+} F(y_n) = F(y)$.

Quando $y < \mu - \lambda\sigma$,

$F(y) = \alpha_2 \left[\frac{1}{\pi}\arctan\left(\frac{y-\mu}{\sigma}\right)+\frac{1}{2}\right]$. Então,

$$\lim_{y_n \downarrow y^+} F(y_n) = \lim_{y_n \downarrow y^+} \alpha_2 \left[\frac{1}{\pi}\arctan\left(\frac{y_n-\mu}{\sigma}\right)+\frac{1}{2}\right] =$$

$$= \alpha_2 \lim_{y_n \downarrow y^+} \left[\frac{1}{\pi}\arctan\left(\frac{y_n-\mu}{\sigma}\right)+\frac{1}{2}\right] = \alpha_2 \left[\frac{1}{\pi}\arctan\left(\frac{y-\mu}{\sigma}\right)+\frac{1}{2}\right] = F(y),$$

pois $\left[\frac{1}{\pi}\arctan\left(\frac{y_n-\mu}{\sigma}\right)+\frac{1}{2}\right]$ é função de distribuição Cauchy contínua à direita.

Quando $\mu - \lambda\sigma \leq y < \mu + \lambda\sigma$,

$F(y) = \alpha_2 \left[\frac{1}{\pi}\arctan(-\lambda)+\frac{1}{2}\right] + \alpha_1 \lim_{y_n \downarrow y^+} \left[\Phi\left(\frac{y_n-\mu}{\sigma}\right) - \Phi(-\lambda)\right]$. Então,

$$\lim_{y_n \downarrow y^+} F(y_n) = \lim_{y_n \downarrow y^+} \alpha_2 \left[\frac{1}{\pi}\arctan(-\lambda)+\frac{1}{2}\right] + \lim_{y_n \downarrow y^+} \alpha_1 \left[\Phi\left(\frac{y_n-\mu}{\sigma}\right) - \Phi(-\lambda)\right] =$$

$$\alpha_2 \left[\frac{1}{\pi}\arctan(-\lambda)+\frac{1}{2}\right] + \alpha_1 \lim_{y_n \downarrow y^+} \left[\Phi\left(\frac{y_n-\mu}{\sigma}\right) - \Phi(-\lambda)\right] = F(y),$$

pois $\Phi\left(\frac{y_n-\mu}{\sigma}\right)$ é função de distribuição acumulada Normal padrão, contínua à direita.

Quando $y = \mu + \lambda\sigma$,

$F(y) = \alpha_2 \left[\frac{1}{\pi}\arctan(-\lambda)+\frac{1}{2}\right] + \alpha_1 [\Phi(\lambda) - \Phi(-\lambda)]$. Então,

$$\lim_{y_n \downarrow y^+} F(y_n) = \lim_{y_n \downarrow y^+} \alpha_2 \left[\frac{-2}{\pi}\arctan(\lambda)+\frac{1}{2}+\frac{1}{\pi}\arctan\left(\frac{y_n-\mu}{\sigma}\right)\right] + \lim_{y_n \downarrow y^+} \alpha_1 [\Phi(\lambda) - \Phi(-\lambda)];$$

$$\lim_{y_n \downarrow y^+} F(y_n) = \alpha_2 \left[\frac{-2}{\pi}\arctan(\lambda)+\frac{1}{2}+\frac{1}{\pi}\arctan\left(\frac{y-\mu}{\sigma}\right)\right] + \alpha_1 [\Phi(\lambda) - \Phi(-\lambda)];$$

$$\lim_{y_n \downarrow y^+} F(y_n) = \alpha_2 \left[\frac{-2}{\pi}\arctan(\lambda)+\frac{1}{2}+\frac{1}{\pi}\arctan\left(\frac{(\mu+\lambda\sigma)-\mu}{\sigma}\right)\right] + \alpha_1 [\Phi(\lambda) - \Phi(-\lambda)];$$

$$\lim_{y_n \downarrow y^+} F(y_n) = \alpha_2 \left[\frac{-2}{\pi}\arctan(\lambda)+\frac{1}{2}+\frac{1}{\pi}\arctan(\lambda)\right] + \alpha_1 [\Phi(\lambda) - \Phi(-\lambda)];$$

$$\lim_{y_n \downarrow y^+} F(y_n) = \alpha_2 \left[\frac{1}{\pi}\arctan(-\lambda)+\frac{1}{2}\right] + \alpha_1 [\Phi(\lambda) - \Phi(-\lambda)] = F(y).$$

Quando $y > \mu + \lambda\sigma$,

$$F(y) = \alpha_2 \left[\frac{-2}{\pi} \arctan(\lambda) + \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{y-\mu}{\sigma} \right) \right] + \alpha_1 [\Phi(\lambda) - \Phi(-\lambda)];$$

$$\lim_{y_n \downarrow y^+} F(y_n) = \lim_{y_n \downarrow y^+} \alpha_2 \left[\frac{-2}{\pi} \arctan(\lambda) + \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{y_n-\mu}{\sigma} \right) \right] + \lim_{y_n \downarrow y^+} \alpha_1 [\Phi(\lambda) - \Phi(-\lambda)]$$

$$= \alpha_2 \left[\frac{-2}{\pi} \arctan(\lambda) + \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{y-\mu}{\sigma} \right) \right] + \alpha_1 [\Phi(\lambda) - \Phi(-\lambda)] = F(y)$$

Pois a função $\arctan\left(\frac{y_n-\mu}{\sigma} \right)$ é contínua em $\mathbb{R}$. □

Proposição 5. $\lim_{y_n \to -\infty} F(y_n) = 0$

Demonstração. Da função de distribuição acumulada (2.7),

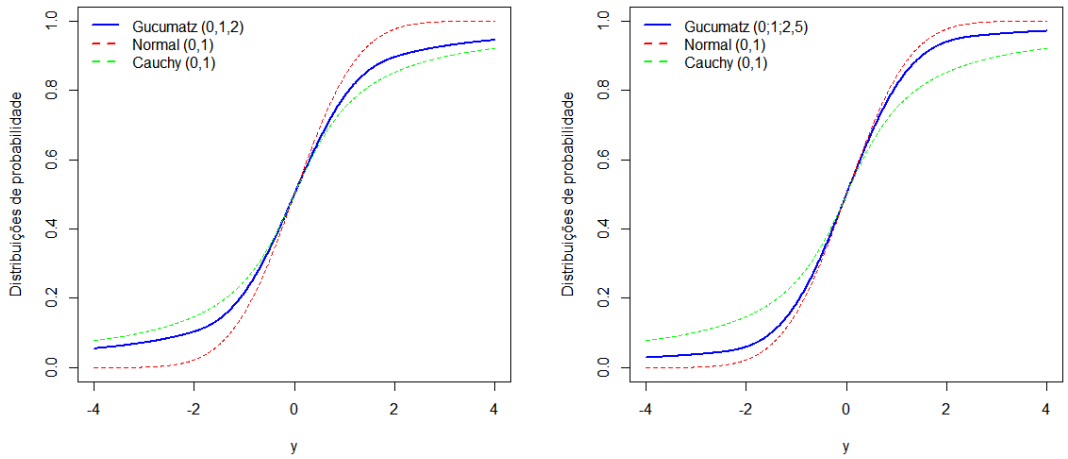
$$\lim_{y_n \to -\infty} F(y_n) =$$

$$\lim_{y_n \to -\infty} \alpha_2 \left[\frac{1}{\pi} \arctan\left(\frac{y_n-\mu}{\sigma} \right) + \frac{1}{2} \right] = \alpha_2 \lim_{y_n \to -\infty} \left[\frac{1}{\pi} \arctan\left(\frac{y_n-\mu}{\sigma} \right) + \frac{1}{2} \right] =$$

$$\alpha_2 \left[\frac{1}{\pi} \arctan(-\infty) + \frac{1}{2} \right] = \alpha_2 \left[\frac{1}{\pi} \left(\frac{-\pi}{2} \right) + \frac{1}{2} \right] = \alpha_2 \cdot 0 = 0.$$

□

Uma comparação entre as curvas das funções de distribuição acumulada dos modelos Gucumatz $(0,1,\lambda)$, Cauchy (0,1) e Normal (0,1), para $\lambda = 2$ e $\lambda = 2,5$ é apresentada na Figura 6.



(a) $\lambda = 2$ (b) $\lambda = 2,5$

Figura 6 – Comparação entre as curvas das funções de distribuição acumulada Gucumatz $(0,1,\lambda)$, Cauchy (0,1) e Normal (0,1) quando (a) $\lambda = 2$ e (b) $\lambda = 2,5$.

Pelas Figuras 6a e 6b, observa-se que as curvas das funções de distribuição acumulada

Gucumatz (0, 1, 2) e Gucumatz (0; 1; 2,5) se distinguem das funções de distribuição acumulada Normal (0,1) e Cauchy (0,1).

2.3 Propriedades

Se uma variável aleatória $Y \in \mathbb{R}$ segue distribuição Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$, algumas propriedades são verificadas por meio das Proposições 6, 7, 8 e 9.

Proposição 6. $E(Y)$ é indefinida, e portanto, a variância de Y em torno da média $E(Y)$ também é indefinida.

Demonstração. $E(Y) = \int_{-\infty}^{\infty} y f(y) dy =$

$$= \int_{-\infty}^{\mu-\lambda\sigma} \frac{y \cdot \alpha_2}{\sigma\pi\left[1+\left(\frac{y-\mu}{\sigma}\right)^2\right]} dy + \int_{\mu-\lambda\sigma}^{\mu+\lambda\sigma} y \cdot \alpha_1 \cdot \frac{\exp\left[\frac{-1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]}{\sigma\sqrt{2\pi}} dy + \int_{\mu+\lambda\sigma}^{\infty} \frac{y \cdot \alpha_2}{\sigma\pi\left[1+\left(\frac{y-\mu}{\sigma}\right)^2\right]} dy;$$

Resolvendo cada uma das três integrais separadamente, usando $z = \dfrac{y-\mu}{\sigma}$,

$$\int_{-\infty}^{\mu-\lambda\sigma} \frac{y\alpha_2}{\sigma\pi\left[1+\left(\frac{y-\mu}{\sigma}\right)^2\right]} dy = \frac{\alpha_2}{\pi}\int_{-\infty}^{-\lambda} \frac{\sigma z+\mu}{1+z^2} dz = \frac{\alpha_2\sigma}{\pi}\int_{-\infty}^{-\lambda} \frac{z}{1+z^2} dz + \frac{\alpha_2\mu}{\pi}\int_{-\infty}^{-\lambda} \frac{dz}{1+z^2}.$$

Usando $s = 1+z^2 \Rightarrow ds = 2zdz \Rightarrow zdz = ds/2$, obtém-se

$$= \frac{\alpha_2\sigma}{2\pi}\int_{-\infty}^{1+(-\lambda)^2} \frac{ds}{s} + \frac{\alpha_2\mu}{\pi}\left[\arctan(-\lambda) - \lim_{z\to-\infty}\arctan(z)\right]$$

$$= \frac{\alpha_2\sigma}{2\pi}\left\{\log[1+\lambda^2] - \lim_{z\to-\infty}\log(1+z^2)\right\} + \frac{\alpha_2\mu}{\pi}\left[\arctan(-\lambda)+\frac{\pi}{2}\right] =$$

$$= \frac{\alpha_2\sigma}{2\pi}[\log(1+\lambda^2) - \infty] + \frac{\alpha_2\mu}{\pi}\left[\arctan(-\lambda)+\frac{\pi}{2}\right] = -\infty.$$

Resolvendo a segunda integral,

$$\alpha_1\int_{\mu-\lambda\sigma}^{\mu+\lambda\sigma} y\frac{\exp\left[\frac{-1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]}{\sigma\sqrt{2\pi}} dy = \alpha_1\int_{-\lambda}^{\lambda} y\frac{e^{\frac{-z^2}{2}}}{\sqrt{2\pi}} dy = \alpha_1 \cdot [\Phi(\lambda) - \Phi(-\lambda)].$$

Resolvendo a terceira integral,

$$\alpha_2 \int_{\mu+\sigma}^{\infty} \frac{y}{\sigma\pi\left[1+\left(\frac{y-\mu}{\sigma}\right)^2\right]} dy = \frac{\alpha_2}{\pi}\int_1^{\infty} \frac{\sigma z+\mu}{1+z^2} dz = \frac{\alpha_2\sigma}{\pi}\int_1^{\infty} \frac{z}{1+z^2} dz + \frac{\alpha_2\mu}{\pi}\int_1^{\infty} \frac{dz}{1+z^2}.$$

Usando $s = 1+z^2 \Rightarrow ds = 2zdz \Rightarrow zdz = ds/2$, obtém-se

$$= \frac{\alpha_2\sigma}{2\pi}\int_{1+\lambda^2}^{\infty} \frac{ds}{s} + \frac{\alpha_2\mu}{\pi}\left[\lim_{z\to\infty} \arctan(z) - \arctan(\lambda)\right]$$

$$= \frac{\alpha_2\sigma}{2\pi}\left\{\lim_{z\to\infty} \log(1+z^2) - \log[1+\lambda^2]\right\} + \frac{\alpha_2\mu}{\pi}\left\{\frac{\pi}{2} - \arctan(\lambda)\right\} =$$

$$= \frac{\alpha_2\sigma}{2\pi}\left\{\infty - \log(1+\lambda^2)\right\} + \frac{\alpha_2\mu}{\pi}\left\{\frac{\pi}{2} - \arctan(\lambda)\right\} = \infty.$$

Assim, $E(Y) = \int_{-\infty}^{\infty} yf(y)dy = -\infty + \alpha_1 \cdot [\Phi(\lambda) - \Phi(-\lambda)] + \infty$.

□

Proposição 7. A variância de Y em torno do parâmetro μ é indefinida.

Demonstração. $E(Y-\mu)^2 = \int_{-\infty}^{\infty} (y-\mu)^2 f(y)dy =$

$$= \int_{-\infty}^{\mu-\lambda\sigma} \frac{(y-\mu)^2\alpha_2}{\sigma\pi\left[1+\left(\frac{y-\mu}{\sigma}\right)^2\right]} dy + \int_{\mu-\lambda\sigma}^{\mu+\lambda\sigma} \frac{(y-\mu)^2\alpha_1 \exp\left[\frac{-1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]}{\sigma\sqrt{2\pi}} dy + \int_{\mu+\lambda\sigma}^{\infty} \frac{(y-\mu)^2\alpha_2}{\sigma\pi\left[1+\left(\frac{y-\mu}{\sigma}\right)^2\right]} dy;$$

Fazendo $z = \dfrac{y-\mu}{\sigma}$ e resolvendo a primeira integral,

$$\alpha_2 \int_{-\infty}^{\mu-\lambda\sigma} \frac{(y-\mu)^2}{\sigma\pi\left[1+\left(\frac{y-\mu}{\sigma}\right)^2\right]} dy = \frac{\alpha_2}{\pi}\int_{-\infty}^{-\lambda} \frac{(\sigma z+\mu-\mu)^2}{1+z^2} dz = \frac{\alpha_2}{\pi}\int_{-\infty}^{-\lambda} \frac{(\sigma z)^2}{1+z^2} dz$$

$$= \frac{\alpha_2\sigma^2}{\pi}\int_{-\infty}^{-\lambda} \frac{z^2}{1+z^2} dz$$

$$= \frac{\alpha_2\sigma^2}{\pi}\left\{1 + (-\lambda)^2 - \log|1+(-\lambda)^2| - \lim_{z\to-\infty}[1+z^2 - \log|1+z^2|]\right\}.$$

Visto que z^2 tende a ∞ e $-\log|1+z^2|$ tende a $-\infty$, então $Var(Y)$ é indefinida.

□

Proposição 8. A função inversa da distribuição de probabilidade acumulada Gucumatz, também denominada função quantil $q_u = F^{-1}(u)$, $0 \le u \le 1$, é dada por:

$$q_u = \sigma \cdot \tan\left[\pi\left(\frac{u}{\alpha_2} - \frac{1}{2}\right)\right] + \mu$$

$$\left(\text{se} \quad u < \alpha_2 \left[\frac{1}{\pi}\arctan(-\lambda)+\frac{1}{2}\right]\right), \text{ referente ao intervalo } y < \mu - \lambda\sigma;$$

$$q_u = \sigma \cdot \Phi^{-1}\left\{\frac{u}{\alpha_1} - \frac{\alpha_2}{\alpha_1}\cdot\left[\frac{1}{\pi}\arctan(-\lambda)+\frac{1}{2}\right] + \Phi(-\lambda)\right\} + \mu$$

$$\left(\text{se} \quad \alpha_2 \left[\frac{1}{\pi}\arctan(-\lambda)+\frac{1}{2}\right] \leq u \leq \alpha_2 \left[\frac{1}{\pi}\arctan(-\lambda)+\frac{1}{2}\right] + \alpha_1[\Phi(\lambda)-\Phi(-\lambda)]\right),$$

referente ao intervalo central $\mu + \lambda\sigma \leq y \leq \mu + \lambda\sigma$;

$$q_u = \sigma \cdot \tan\left\{\frac{u\pi}{\alpha_2} - \pi\left[\frac{1}{\pi}\arctan(-\lambda)+\frac{1}{2}\right] - \frac{\pi}{\alpha_2}\alpha_1 \cdot [\Phi(\lambda)-\Phi(-\lambda)] + \arctan(\lambda)\right\} + \mu,$$

$$\left(\text{se} \quad u > \alpha_2 \cdot \left\{1 + \frac{1}{\pi}\left[\arctan(-\lambda) - \arctan(\lambda)\right]\right\} + \alpha_1[\Phi(\lambda)-\Phi(-\lambda)]\right),$$

referente ao intervalo $y \geq \mu + \lambda\sigma$.

Proposição 9. A mediana da distribuição Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$ é igual ao parâmetro μ.

Demonstração. Calcula-se a função quantil Gucumatz correspondente ao intervalo central $\mu + \lambda\sigma \leq y \leq \mu + \lambda\sigma$ na probabilidade $u = 0,5$.

$$q_{(0,5)} = F^{-1}(0,5) = \sigma \cdot \Phi^{-1}\left\{\frac{0,5}{\alpha_1} - \frac{\alpha_2}{\alpha_1}\cdot\left[\frac{1}{\pi}\arctan(-\lambda)+\frac{1}{2}\right] + \Phi(-\lambda)\right\} + \mu;$$

De α_1 e α_2 em (2.5), e das propriedades $\Phi(-\lambda) = 1 - \Phi(\lambda)$ e $\arctan(-\lambda) = -\arctan(\lambda)$, após manipulações algébricas,

$$q_{(0,5)} = \sigma \cdot \Phi^{-1}\left\{\frac{e^{\frac{-\lambda^2}{2}}(1+\lambda^2)}{\sqrt{2\pi}}\left[-0,5(-\arctan(\lambda)) - 0,5\arctan(\lambda)\right] + 0,5\right\} + \mu;$$

$$q_{(0,5)} = \sigma \cdot \Phi^{-1}(0,5) + \mu = \sigma \cdot 0 + \mu = \mu.$$

□

Apesar da distribuição Gucumatz apresentar variância indefinida, demonstra-se que o parâmetro σ é o parâmetro de dispersão quantílica, o qual possibilita uma ordenação em variabilidade entre distribuições Gucumatz.

2.4 Parâmetro de dispersão quantílica

A dispersão quantílica (*quantile spread*) é uma medida introduzida por Townsend e Colonius (2005) que possibilita uma ordenação dispersiva entre variáveis aleatórias. A dispersão

quantílica de uma variável aleatória Y, que segue uma função de distribuição uniformemente contínua F, é dada por

$$QS_Y(u) = F^{-1}(1-u) - F^{-1}(u), \quad 0 < u < 0,5, \tag{2.8}$$

de modo que, dadas duas variáveis aleatórias Y_1 e Y_2 com funções de distribuição F_{Y_1} e F_{Y_2} tais que $QS_{Y_1}(u) \leq QS_{Y_2}(u)$, então considera-se que Y_1 é menor do que Y_2 em ordem de dispersão quantílica, o que denota-se por $Y_1 \leq_{QS} Y_2$. Isto significa que a distância $QS_{Y_2}(u) = F_{Y_2}^{-1}(1-u) - F_{Y_2}^{-1}(u)$ entre quaisquer quantis simétricos em relação à mediana da variável Y_2 é igual ou maior do que a distância $QS_{Y_1}(u) = F_{Y_1}^{-1}(1-u) - F_{Y_1}^{-1}(u)$ entre os mesmos quantis simétricos em relação à mediana da variável Y_1.

Com base em Shaked e Shanthikumar (1994), Townsend e Colonius (2005) afirmam que, em distribuições simétricas, a ordenação em dispersão quantílica entre duas variáveis Y_1 e Y_2 corresponde a uma ordenação em variabilidade, isto é, $Y_1 \leq_{QS} Y_2$ implica $Var(Y_1) \leq Var(Y_2)$.

Na distribuição Gucumatz, que é simétrica, mostra-se que o parâmetro σ é o parâmetro de dispersão quantílica, com base em Mitnik e Baek (2012). Seja $Y_1 \sim$ Gucumatz (μ, σ_1, λ) e $Y_2 \sim$ Gucumatz (μ, σ_2, λ). Afirmar que σ é o parâmetro de dispersão quantílica significa dizer que $Y_1 \leq_{QS} Y_2$ se e só se, $\sigma_1 \leq \sigma_2$, que por sua vez, implica em $Var(Y_1) \leq Var(Y_2)$.

Proposição 10. Seja $Y_1 \sim$ Gucumatz (μ, σ_1, λ) e $Y_2 \sim$ Gucumatz (μ, σ_2, λ). Então Y_1 é menor do que Y_2 em ordem de dispersão quantílica, isto é, $Y_1 \leq_{QS} Y_2$ se e só se, $\sigma_1 \leq \sigma_2$.

Demonstração. Seja a função quantil da distribuição Gucumatz dada na Proposição 8. Se $Y_1 \sim$ Gucumatz (μ, σ_1, λ), $Y_2 \sim$ Gucumatz (μ, σ_2, λ) e $0 < u < 0,5$, então a dispersão quantílica de Y_1 e Y_2 são dadas, respectivamente, por

$$QS_{Y_1}(u) = \sigma_1 \left\{ \tan\left[\frac{(1-u)\pi}{\alpha_2} + 2\arctan(\lambda) - \frac{\pi}{2} - \frac{\pi\alpha_1}{\alpha_2}(2\Phi(\lambda) - 1) \right] - \tan\left(\frac{\pi u}{\alpha_2} - \frac{\pi}{2} \right) \right\} \quad \text{e}$$

$$QS_{Y_2}(u) = \sigma_2 \left\{ \tan\left[\frac{(1-u)\pi}{\alpha_2} + 2\arctan(\lambda) - \frac{\pi}{2} - \frac{\pi\alpha_1}{\alpha_2}(2\Phi(\lambda) - 1) \right] - \tan\left(\frac{\pi u}{\alpha_2} - \frac{\pi}{2} \right) \right\},$$

se $u = F_{Y_1}(y_1)$ e $u = F_{Y_2}(y_2)$ correspondem aos intervalos $y_1 < \mu - \lambda\sigma_1$ e $y_2 < \mu - \lambda\sigma_2$ respectivamente; e

$$QS_{Y_1}(u) = \sigma_1 \left\{ \Phi^{-1}\left[\frac{1-u}{\alpha_1} - \frac{\alpha_2}{\alpha_1}\left(\frac{\arctan(-\lambda)}{\pi} + \frac{1}{2} \right) + \Phi(-\lambda) \right] - \Phi^{-1}\left[\frac{u}{\alpha_1} - \frac{\alpha_2}{\alpha_1}\left(\frac{\arctan(-\lambda)}{\pi} + \frac{1}{2} \right) + \Phi(-\lambda) \right] \right\}$$

e

$$QS_{Y_2}(u) = \sigma_2 \left\{ \Phi^{-1}\left[\frac{1-u}{\alpha_1} - \frac{\alpha_2}{\alpha_1}\left(\frac{\arctan(-\lambda)}{\pi} + \frac{1}{2} \right) + \Phi(-\lambda) \right] - \Phi^{-1}\left[\frac{u}{\alpha_1} - \frac{\alpha_2}{\alpha_1}\left(\frac{\arctan(-\lambda)}{\pi} + \frac{1}{2} \right) + \Phi(-\lambda) \right] \right\},$$

se $u = F_{Y_1}(y_1)$ e $u = F_{Y_2}(y_2)$ correspondem aos intervalos $\mu + \lambda\sigma_1 < y_1 \leq \mu + \lambda\sigma_1$ e $\mu + \lambda\sigma_2 < y_2 \leq \mu + \lambda\sigma_2$, respectivamente.

Suponha que $QS_{Y_1}(u) \leq QS_{Y_2}(u)$. Visto que todos os termos na desigualdade $QS_{Y_1}(u) \leq QS_{Y_2}(u)$ são iguais nos dois membros, exceto $\sigma_1 > 0$ e $\sigma_2 > 0$, que são diretamente proporcionais a QS_{Y_1} e QS_{Y_2} respectivamente, então $QS_{Y_1}(u) \leq QS_{Y_2}(u)$ se e só se, $\sigma_1 \leq \sigma_2$. Portanto, σ é o parâmetro de dispersão quantílica, que possibilita comparações em variabilidade entre variáveis que seguem distribuição Gucumatz. □

2.5 Simulação

A simulação de um conjunto de dados cuja variável resposta segue distribuição Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$ é realizada por meio do seguinte algoritmo

Algoritmo 1 – Simulação de amostra Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$

1. Gere vetor $\boldsymbol{u}$ de tamanho n com distribuição Uniforme (0,1);
2. $\boldsymbol{y}$ recebe valores de $F^{-1}(\boldsymbol{u})$; (O código da função quantil F^{-1} está no Apêndice B)
3. Retorne $\boldsymbol{y}$.

Um histograma de uma amostra de tamanho $n = 100$ da distribuição Gucumatz $(0; 1; 2,5)$, simulada com auxílio do software RStudio Team (2018), é apresentado na Figura 7. Neste histograma estão sobrepostas uma curva da função densidade da distribuição Gucumatz $(0; 1; 2,5)$ e curvas de funções densidades estimadas por máxima verossimilhança das distribuições Normal, Cauchy e Logística, utilizando os dados gerados da distribuição Gucumatz.

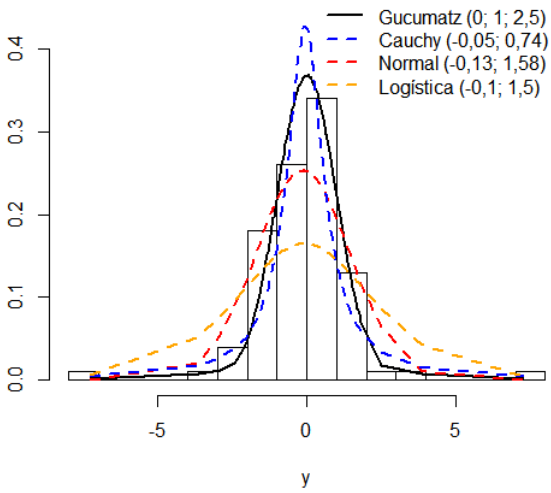


Figura 7 – Curvas de funções densidades estimadas das distribuições Cauchy, Normal e Logística para uma amostra simulada da distribuição Gucumatz $(0; 1; 2,5)$

Pela Figura 7, observa-se que a curva da função densidade da distribuição Gucumatz (0;1;2,5) ($\alpha_1 = 0,9222622, \alpha_2 = 0,3681986$) está visualmente bem ajustada ao histograma. Além disto, a função densidade Gucumatz distingue-se das funções densidades estimadas das distribuições Normal, Cauchy e Logística.

Neste capítulo, foram estabelecidos os parâmetros $\mu, \sigma, \lambda, \alpha_1, \alpha_2$ da distribuição Gucumatz, na qual α_1 e α_2 são definidos em função de λ. Foram apresentadas a função densidade de probabilidade, a função de distribuição acumulada, a função quantil, a mediana μ da distribuição, e o parâmetro de dispersão quantílica σ, que expressa uma ordenação em variabilidade na distribuição Gucumatz. Demonstrou-se que a distribuição Gucumatz apresenta média e variâncias indefinidas, assim como ocorre com a distribuição Cauchy. Constatou-se, visualmente, que a distribuição Gucumatz apresenta função densidade com forma distinta das distribuições Normal e Cauchy quando $2 \leq \lambda \leq 2,5$. Por fim, realizou-se uma simulação de dados, que seguem a distribuição Gucumatz (0,1, 2,5), por meio da inversa da função de distribuição acumulada.

CAPÍTULO

3

ESTIMAÇÃO GUCUMATZ

Neste capítulo é desenvolvida a estimação frequentista por máxima verossimilhança dos parâmetros da distribuição Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$, intervalos de confiança e estudo de simulação. Em seguida, apresenta-se um possível caminho para a estimação bayesiana, embora seja dada preferência à estimação frequentista neste trabalho. Primeiramente, são estabelecidos alguns resultados, e em seguida são apresentadas as estimações nas Seções 3.1 e 3.2.

A função densidade de uma variável aleatória $Y \sim$ Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$ dada em (2.5), em que $\mu \in \mathbb{R}, \sigma > 0$, $\lambda > 0$, $\alpha_1 \in (0{,}693; 1)$ e $\alpha_2 \in (0; 1{,}054)$. Estima-se apenas o vetor de parâmetros $\boldsymbol{\theta} = (\mu, \sigma, \lambda)$, pois os parâmetros α_1 e α_2 são estimados em função de $\hat{\lambda}$, isto é:

$$\hat{\alpha}_1 = \left\{ \frac{e^{\frac{-\hat{\lambda}^2}{2}} \pi (1+\hat{\lambda}^2)}{\sqrt{2\pi}} \left[1 + \frac{1}{\pi} [\arctan(-\hat{\lambda}) - \arctan(\hat{\lambda})] \right] + \Phi(\hat{\lambda}) - \Phi(-\hat{\lambda}) \right\}^{-1}; \tag{3.1}$$

$$\hat{\alpha}_2 = \frac{\hat{\alpha}_1 e^{\frac{-\hat{\lambda}^2}{2}} \pi (1+\hat{\lambda}^2)}{\sqrt{2\pi}}. \tag{3.2}$$

A função de verossimilhança do vetor de parâmetros $\boldsymbol{\theta} = (\mu, \sigma, \lambda)$, dado uma amostra $\boldsymbol{y} = (y_1, y_2, ..., y_n)$ de tamanho n, é escrita como

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{n} \begin{cases} \dfrac{\alpha_1}{\sigma\sqrt{2\pi}} \exp\left[\dfrac{-1}{2} \left(\dfrac{y_i - \mu}{\sigma} \right)^2 \right], & se \quad \mu - \lambda\sigma \leq y_i \leq \mu + \lambda\sigma \\ \dfrac{\alpha_2}{\sigma\pi \left[1 + \left(\dfrac{y_i - \mu}{\sigma} \right)^2 \right]}, & c.c. \end{cases} . \tag{3.3}$$

O logaritmo natural aplicado à função de verossimilhança dada em (3.3), denominada

por função de log-verossimilhança, é escrita como

$$\log L(\boldsymbol{\theta}|\boldsymbol{y}) = l(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n} \log \begin{cases} \dfrac{\alpha_1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2\right], & se \quad \mu-\lambda\sigma \leq y_i \leq \mu+\lambda\sigma \\ \dfrac{\alpha_2}{\sigma\pi\left[1+\left(\frac{y_i-\mu}{\sigma}\right)^2\right]}, & c.c. \end{cases} . \tag{3.4}$$

Seja uma variável aleatória não-observável K com realizações

$$K = \begin{cases} 1 & \text{se} \quad \mu-\lambda\sigma \leq Y \leq \mu+\lambda\sigma \\ 0 & c.c. \end{cases} ; \tag{3.5}$$

De (2.6), $P(K=1) = \alpha_1[\Phi(\lambda) - \Phi(-\lambda)]$. Para simplificação, denomina-se $\rho = \Phi(\lambda) - \Phi(-\lambda)$. Assim, $K \sim Bernoulli(\alpha_1\rho)$. Visto que $\alpha_1 \in (0,693;1)$ e $\rho \in [0,1]$, então $0 \leq \alpha_1\rho \leq 1$.

A função de verossimilhança do vetor de parâmetros $\boldsymbol{\theta}$ condicionada ao vetor de dados observáveis $\boldsymbol{y} = (y_1, y_2, ..., y_n)$ e ao vetor de dados não observáveis $\boldsymbol{k} = (k_1, k_2, ..., k_n)$, denominada por função de verossimilhança dos dados completos, é escrita como

$$L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{k}) = f(\boldsymbol{y}|\boldsymbol{k},\boldsymbol{\theta}) \cdot f(\boldsymbol{k}|\alpha_1\rho) = \prod_{i=1}^{n} \left\{ \left[\alpha_1 \frac{\exp(\frac{-1}{2}(\frac{y_i-\mu}{\sigma})^2)}{\sigma\sqrt{2\pi}}\right]^{k_i} \left[\frac{\alpha_2}{\sigma\pi[1+(\frac{y_i-\mu}{\sigma})^2]}\right]^{1-k_i} (\alpha_1\rho)^{k_i}(1-\alpha_1\rho)^{1-k_i} \right\}. \tag{3.6}$$

O logaritmo natural aplicado à função de verossimilhança dada em (3.6), denominada por função de log-verossimilhança dos dados completos, é escrita como

$$\begin{aligned} \log L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{k}) = l(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{k}) = \sum_{i=1}^{n} k_i \log(\alpha_1) + \sum_{i=1}^{n}\left[\frac{-k_i}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2\right] - \sum_{i=1}^{n} k_i \log(\sigma\sqrt{2\pi}) + \\ \sum_{i=1}^{n}(1-k_i)\log(\alpha_2) - \sum_{i=1}^{n}(1-k_i)\log(\sigma\pi) - \sum_{i=1}^{n}(1-k_i)\log\left[1+\left(\frac{y_i-\mu}{\sigma}\right)^2\right] + \\ \sum_{i=1}^{n} k_i \log(\alpha_1\rho) + (n - \sum_{i=1}^{n} k_i)\log(1-\alpha_1\rho). \end{aligned} \tag{3.7}$$

Apresentados os resultados iniciais, realizam-se as estimações.

3.1 Estimação frequentista

Realiza-se estimação por máxima verossimilhança de dados completos, com utilização do algoritmo Esperança-Maximização (Algoritmo EM), de Dempster, Laird e Rubin (1977), para maximizar a esperança da função de log-verossimilhança dos dados completos dada em 3.7 com relação à distribuição da variável $K|Y,\boldsymbol{\theta}$, como a seguir.

A distribuição de probabilidade da variável $K|Y,\boldsymbol{\theta}$ é dada por

$$\begin{aligned} P(K=1|y,\boldsymbol{\theta}) = \kappa = \begin{cases} 1, & \text{se} \quad \mu-\lambda\sigma \leq Y \leq \mu+\lambda\sigma \\ 0, & cc. \end{cases} \quad ; \\ P(K=0|y,\boldsymbol{\theta}) = 1-\kappa. \end{aligned} \tag{3.8}$$

A esperança da função de log-verossimilhança dos dados completos (3.7), na distribuição de $K|Y,\boldsymbol{\theta}$ mostrada em (3.8), é dada por

$$\begin{aligned} E_{K|Y,\boldsymbol{\theta}}[l(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{k})] = \sum_{i=1}^{n} \kappa_i [l(\boldsymbol{\theta}|y_i,k_i=1)] + \sum_{i=1}^{n}(1-\kappa_i)[l(\boldsymbol{\theta}|y_i,k_i=0)] = \\ \sum_{i=1}^{n} \kappa_i \log(\alpha_1) + \sum_{i=1}^{n}\left[\frac{-\kappa_i}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2\right] - \sum_{i=1}^{n}\kappa_i \log(\sigma\sqrt{2\pi}) + \sum_{i=1}^{n}(1-\kappa_i)\log(\alpha_2) \\ -\sum_{i=1}^{n}(1-\kappa_i)\log(\sigma\pi) - \sum_{i=1}^{n}(1-\kappa_i)\log\left[1+\left(\frac{y_i-\mu}{\sigma}\right)^2\right] + \sum_{i=1}^{n}\kappa_i\log(\alpha_1\rho) \\ + \sum_{i=1}^{n}(1-\kappa_i)\log(1-\alpha_1\rho) = l(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{\kappa}). \end{aligned} \tag{3.9}$$

O Algoritmo Esperança-Maximização é realizado como a seguir. Supondo que se esteja ao final da iteração t do Algoritmo EM, o vetor $\boldsymbol{\theta}^{(t)} = (\mu^{(t)},\sigma^{(t)},\lambda^{(t)})$ é atribuído ao vetor de parâmetros $\boldsymbol{\theta} = (\mu,\sigma,\lambda)$. Ao início da iteração $t+1$, é obtida a distribuição de probabilidade da variável $K_i|Y,\boldsymbol{\theta}^{(t)}$ $(1 \leq i \leq n)$ de (3.8), reescrita como

$$\begin{aligned} P(K_i=1|y_i,\boldsymbol{\theta}^{(t)}) = \kappa_i^{(t)} = \begin{cases} 1, & \text{se} \quad \mu^{(t)}-\lambda^{(t)}\sigma^{(t)} \leq Y_i \leq \mu^{(t)}+\lambda^{(t)}\sigma^{(t)} \\ 0, & cc. \end{cases} \quad ; \\ P(K_i=0|y_i,\boldsymbol{\theta}^{(t)}) = 1-\kappa_i^{(t)}. \end{aligned} \tag{3.10}$$

A esperança da log-verossimilhança dos dados completos (3.9) na distribuição da variável $K_i|Y,\boldsymbol{\theta}^{(t)}$ em (3.10) é dada por

$$E_{K|Y,\boldsymbol{\theta}^{(t)}}[l(\boldsymbol{\theta}^{(t)}|\boldsymbol{y},\boldsymbol{k})] = l(\boldsymbol{\theta}^{(t)}|\boldsymbol{y},\boldsymbol{\kappa}^{(t)}) = \sum_{i=1}^{n}\kappa_i^{(t)}\log(\alpha_1) + \sum_{i=1}^{n}\left[\frac{-\kappa_i}{2}\left(\frac{y_i-\mu^{(t)}}{\sigma^{(t)}}\right)^2\right]$$
$$-\sum_{i=1}^{n}\kappa_i^{(t)}\log(\sigma^{(t)}\sqrt{2\pi}) + \sum_{i=1}^{n}(1-\kappa_i^{(t)})\log(\alpha_2) - \sum_{i=1}^{n}(1-\kappa_i^{(t)})\log(\sigma^{(t)}\pi)$$
$$-\sum_{i=1}^{n}(1-\kappa_i^{(t)})\log\left[1+\left(\frac{y_i-\mu^{(t)}}{\sigma^{(t)}}\right)^2\right] + \sum_{i=1}^{n}\kappa_i^{(t)}\log(\alpha_1\rho) + (n-\sum_{i=1}^{n}\kappa_i^{(t)})\log(1-\alpha_1\rho). \tag{3.11}$$

Por fim, maximiza-se a esperança da log-verossimilhança dos dados completos (3.11) no vetor de parâmetros $\boldsymbol{\theta}^{(t+1)} = (\mu^{(t+1)}, \sigma^{(t+1)}, \lambda^{(t+1)})$. Para isto, é necessário decompor (3.11) em dois termos dados a seguir, para maximizá-los separadamente.

$$\sum_{i=1}^{n}\left[\frac{-\kappa_i}{2}\left(\frac{y_i-\mu^{(t)}}{\sigma^{(t)}}\right)^2\right] - \sum_{i=1}^{n}\kappa_i^{(t)}\log(\sigma^{(t)}\sqrt{2\pi})$$
$$-\sum_{i=1}^{n}(1-\kappa_i^{(t)})\log(\sigma^{(t)}\pi) - \sum_{i=1}^{n}(1-\kappa_i^{(t)})\log\left[1+\left(\frac{y_i-\mu^{(t)}}{\sigma^{(t)}}\right)^2\right], \tag{3.12}$$

e

$$\sum_{i=1}^{n}\kappa_i^{(t)}\log(\alpha_1) + \sum_{i=1}^{n}(1-\kappa_i^{(t)})\log(\alpha_2) + \sum_{i=1}^{n}\kappa_i^{(t)}\log(\alpha_1\rho) + (n-\sum_{i=1}^{n}\kappa_i^{(t)})\log(1-\alpha_1\rho). \tag{3.13}$$

Com auxílio do pacote *stats* do software RStudio Team (2018), maximiza-se (3.12) em $\mu^{(t+1)}$ e $\sigma^{(t+1)}$ com a utilização de um método de otimização numérica *Broyden-Flecther-Goldfarb-Shanno* (BFGS), de BROYDEN (1970), Fletcher (1970), Goldfarb (1970) e Shanno (1970), que minimiza uma função utilizando gradientes descendentes. Por sua vez, maximiza-se (3.13) em $\lambda^{(t+1)}$ com a utilização do método *Brent*, de Brent (1973), que aproxima raízes de uma função combinando métodos da bissecção, das secantes, e interpolação quadrática inversa. Nesta maximização, deve haver como restrição $0 \leq \lambda \leq 4$, pois, do contrário, quando $\kappa_i^{(.)} = 1$ para $1 \leq i \leq n$ durante as iterações, (3.13) apresenta como indeterminados valores máximos $\lambda \geq 4$.

Assim, o vetor $\boldsymbol{\theta}^{(t+1)} = (\mu^{(t+1)}, \sigma^{(t+1)}, \lambda^{(t+1)})$ é atribuído ao vetor de parâmetros $\boldsymbol{\theta} = (\mu, \sigma, \lambda)$. Este ciclo se repete até que a diferença $E_{K|Y,\boldsymbol{\theta}^{(t+1)}}[l(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{y},\boldsymbol{k})] - E_{K|Y,\boldsymbol{\theta}^{(t)}}[l(\boldsymbol{\theta}^{(t)}|\boldsymbol{y},\boldsymbol{k})]$ assume um valor arbitrário muito pequeno, quando $\hat{\mu}$, $\hat{\sigma}$ e $\hat{\lambda}$ alcançam convergência, como demonstrado por Dempster, Laird e Rubin (1977).

Os passos desta estimação estão detalhados no Algoritmo 4 (Apêndice A), o qual propõe as tentativas iniciais $\mu^{(0)} =$ mediana$(\boldsymbol{y})$, $\sigma^{(0)} = |84^{\text{o}}$ quantil amostral - $\mu^{(0)}|$ e $\lambda^{(0)} \sim U(2,3)$. Pode-se adotar outras tentativas iniciais não muito diferentes destas, como $\mu^{(0)} = \bar{\boldsymbol{y}}$, $\sigma^{(0)} =$ desvio padrão amostral e $\lambda^{(0)} \sim U(1,4)$ sem impactar o resultado da estimação. Esta afirmação é baseada nos estudos de simulação realizados, e não pode ser generalizada.

3.1.1 Vetor escore e matriz de informação observada

O vetor escore $\boldsymbol{U_\theta} = (U_\mu, U_\sigma, U_\lambda)$ é dado pelo gradiente da função de log-verossimilhança dos dados completos (3.7). Os elementos do vetor escore são

$$U_\mu = \frac{1}{\sigma}\sum_{i=1}^{n} k_i\left(\frac{y_i-\mu}{\sigma}\right) + \frac{2}{\sigma}\sum_{i=1}^{n}(1-k_i)\frac{\left(\frac{y_i-\mu}{\sigma}\right)}{1+\left(\frac{y_i-\mu}{\sigma}\right)^2}, \tag{3.14}$$

$$U_\sigma = \frac{1}{\sigma}\sum_{i=1}^{n} k_i\left(\frac{y_i-\mu}{\sigma}\right)^2 - \frac{n}{\sigma} + \frac{2}{\sigma}\sum_{i=1}^{n}(1-k_i)\frac{\left(\frac{y_i-\mu}{\sigma}\right)^2}{1+\left(\frac{y_i-\mu}{\sigma}\right)^2}, \tag{3.15}$$

e

$$\begin{aligned} U_\lambda = -2n\alpha_1\frac{e^{-\lambda^2/2}(\lambda-\lambda^3)}{\sqrt{2\pi}}[\pi - 2\arctan(\lambda)] + \sum_{i=1}^{n} k_i\frac{2e^{-\lambda^2/2}}{\sqrt{2\pi}[2\Phi(\lambda)-1]} \\ + 2\sum_{i=1}^{n}(1-k_i)\left[-\lambda + \frac{2\lambda}{1+\lambda^2}\right] + \sum_{i=1}^{n}(1-k_i)\frac{-2/(1+\lambda^2)}{\pi - 2\arctan(\lambda)}. \end{aligned} \tag{3.16}$$

Visto que a distribuição Gucumatz não apresenta média, a matriz de informação esperada $E\left[-\frac{\partial \boldsymbol{U_\theta}}{\partial \boldsymbol{\theta}}\right]$ é aproximada pela matriz de informação observada

$$\boldsymbol{J}(\boldsymbol{\theta}) = \left[-\frac{\partial \boldsymbol{U_\theta}}{\partial \boldsymbol{\theta}}\right] = -\begin{bmatrix} J_{\mu\mu} & J_{\mu\sigma} & J_{\mu\lambda} \\ J_{\sigma\mu} & J_{\sigma\sigma} & J_{\sigma\lambda} \\ J_{\lambda\mu} & J_{\lambda\sigma} & J_{\lambda\lambda} \end{bmatrix}, \quad \text{cujos elementos são:} \tag{3.17}$$

$$J_{\mu\mu} = \frac{-1}{\sigma^2}\sum_{i=1}^{n} k_i + \frac{2}{\sigma^2}\sum_{i=1}^{n}(1-k_i)\frac{-1+\left(\frac{y_i-\mu}{\sigma}\right)^2}{\left[1+\left(\frac{y_i-\mu}{\sigma}\right)^2\right]^2}, \tag{3.18}$$

$$J_{\mu\sigma} = \frac{-2}{\sigma^2}\sum_{i=1}^{n} k_i\left(\frac{y_i-\mu}{\sigma}\right) - \frac{4}{\sigma^2}\sum_{i=1}^{n}(1-k_i)\frac{\left(\frac{y_i-\mu}{\sigma}\right)}{\left[1+\left(\frac{y_i-\mu}{\sigma}\right)^2\right]^2}, \tag{3.19}$$

$$J_{\sigma\sigma} = \frac{-3}{\sigma^2}\sum_{i=1}^{n} k_i\left(\frac{y_i-\mu}{\sigma}\right)^2 + \frac{n}{\sigma^2} - \frac{2}{\sigma^2}\sum_{i=1}^{n}(1-k_i)\frac{3\left(\frac{y_i-\mu}{\sigma}\right)^2 + \left(\frac{y_i-\mu}{\sigma}\right)^4}{\left[1+\left(\frac{y_i-\mu}{\sigma}\right)^2\right]^2}, \tag{3.20}$$

$$J_{\lambda\lambda} = -2n\alpha_1 \frac{e^{-\lambda^2/2}}{\sqrt{2\pi}} \left[(\lambda^4 - 4\lambda^2 + 1)[\pi - 2\arctan(\lambda)] - \frac{2(\lambda - \lambda^3)}{1+\lambda^2} \right]$$
$$+ [n\alpha_1^2 e^{-\lambda^2}(\lambda - \lambda^3)^2/\pi][\pi - 2\arctan(\lambda)]^2 - \frac{2\lambda \sum_{i=1}^n k_i e^{-\lambda^2/2}}{\sqrt{2\pi}[2\Phi - 1]} - \frac{-2\sum_{i=1}^n k_i e^{-\lambda^2}}{\pi[2\Phi(\lambda) - 1]^2}$$
$$+ 2\sum_{i=1}^n (1 - k_i) \left\{ -1 + \frac{2 - 2\lambda^2}{(1+\lambda^2)^2} + \frac{2\lambda[\pi - 2\arctan(\lambda)] - 1}{(1+\lambda^2)^2[\pi - 2\arctan(\lambda)]^2} \right\}, \quad (3.21)$$

$$J_{\mu\lambda} = 0 \quad \text{e} \quad J_{\sigma\lambda} = 0. \quad (3.22)$$

Da matriz de informação observada (3.17) pode-se apresentar intervalos de confiança para os estimadores.

3.1.2 Intervalos de confiança

Dado um parâmetro θ, um estimador de máxima verossimilhança $\hat{\theta}$ apresenta a seguinte distribuição assintótica

$$\hat{\theta} \sim^a N(\theta, J(\hat{\theta})^{-1}), \quad (3.23)$$

isto é, o estimador $\hat{\theta}$ é consistente e assintoticamente eficiente (CASELLA; BERGER, 2016). Em outras palavras, $\hat{\theta}$ é assintoticamente não viciado e apresenta variância igual ao inverso da informação esperada de Fisher de θ avaliado em $\hat{\theta}$, quando a amostra é grande. Estas propriedades são verificadas mediante condições de regularidade que podem ser resumidas em duas, segundo Bolfarine e Sandoval (2001): o suporte da distribuição é independente dos parâmetros e é possível a troca das ordens das operações de derivação e integração, na distribuição em questão.

No entanto, na distribuição Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$, o conjunto suporte depende de todos os parâmetros, principalmente do parâmetro λ, o qual está presente apenas no suporte. Por sua vez, é possível a troca das ordens das operações de derivação e integração, de acordo com teorema fundamental do cálculo (este teorema pode ser visto em Leithold (1994)), visto que a distribuição Gucumatz é uniformemente contínua (Proposição 2). Portanto, tais condições de regularidade são parcialmente satisfeitas. Além disto, a média e a variância na distribuição Gucumatz são indefinidas, fazendo com que as esperanças e variâncias das escores sejam indefinidas.

Contudo, devido a distribuição Gucumatz apresentar componente distribuição Normal em torno da mediana μ, e com base em estudos de simulação no decorrer deste trabalho, para possibilitar utilização de procedimentos de inferência e diagnósticos mais conhecidos, admite-se que os estimadores $\hat{\mu}$ e $\hat{\sigma}$ apresentem distribuições assintoticamente normais,

$$\hat{\mu} \sim^a N(\mu, [\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}]_{1,1}) \quad e \quad \hat{\sigma} \sim^a N(\sigma, [\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}]_{2,2}), \quad (3.24)$$

em que $[\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}]_{1,1}$ é o elemento de posição $(1,1)$ e e $[\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}]_{2,2}$ é o elemento $(2,2)$ da inversa da matriz de informação observada (3.17), que aproxima a matriz de informação esperada de Fisher. O mesmo não se admite para o estimador $\hat{\lambda}$, pois o parâmetro λ depende somente do conjunto suporte da distribuição Gucumatz. Assim, pode-se também considerar que o vetor de estimadores $\hat{\boldsymbol{\theta}}$ apresenta normalidade assintótica,

$$\hat{\boldsymbol{\theta}} \sim^{a} N(\boldsymbol{\theta}, \boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}), \tag{3.25}$$

em que $\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}$ é a inversa da matriz de informação observada (equação 3.17). A normalidade assintótica do vetor $\hat{\boldsymbol{\theta}}$ estabelecida em (3.25) é deficiente, visto que a distribuição Gucumatz satisfaz parcialmente as condições de regularidade necessárias. Contudo, admite-se (3.25) para possibilitar o emprego de procedimentos de inferência e diagnóstico mais conhecidos.

Com base em (3.24) considera-se os seguintes intervalos de confiança assintóticos dos parâmetros μ e σ, com nível de confiança $1-\alpha$:

$$IC_{1-\alpha}(\mu) = \left(\hat{\mu} - z_{(1-\alpha/2)} \sqrt{[\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}]_{1,1}}, \quad \hat{\mu} + z_{(1-\alpha/2)} \sqrt{[\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}]_{1,1}} \right), \tag{3.26}$$

$$IC_{1-\alpha}(\sigma) = \left(\hat{\sigma} - z_{(1-\alpha/2)} \sqrt{[\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}]_{2,2}}, \quad \hat{\sigma} + z_{(1-\alpha/2)} \sqrt{[\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}]_{2,2}} \right), \tag{3.27}$$

em que $z_{(1-\alpha/2)}$ é o quantil $1-\alpha/2$ da função distribuição de probabilidade Normal padrão.

O intervalo de confiança bootstrap para o parâmetro λ está compreendido entre os percentis $100(\alpha/2)$ e $100(1-\alpha/2)$ de m estimativas $\hat{\lambda}$ simuladas ou reamostradas.

3.1.3 Estudo de simulação

Para avaliar a consistência e normalidade assintótica dos estimadores de máxima verossimilhança, são simuladas $m = 1.000$ amostras Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$ de tamanhos $n = 10$, $n = 50$, $n = 100$ e $n = 200$. Neste estudo, são avaliados erro quadrático médio (EQM), variância, vício e probabilidades de cobertura (PC) do vetor de estimadores $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma}, \hat{\lambda})$. Seja o vetor de parâmetros $\boldsymbol{\theta} = (\mu, \sigma, \lambda)$ cujo elemento é θ_j em que $1 \leq j \leq 3$, o qual assume os valores reais pré-determinados $\mu = 0, \sigma = 1$, $\lambda = 2$ e $\lambda = 3$. São avaliados:

$$EQM(\hat{\theta}_j) = \frac{\sum_{l=1}^{m} (\hat{\theta}_{jl} - \theta_j)^2}{m-1}, \quad \text{var}(\hat{\theta}_j) = \frac{\sum_{l=1}^{m} (\hat{\theta}_{jl} - \bar{\hat{\theta}}_j)^2}{m-1} \quad \text{e vício}(\hat{\theta}_j) = \frac{\sum_{l=1}^{m} (\hat{\theta}_{jl} - \theta_j)}{m-1}, \tag{3.28}$$

em que $\bar{\hat{\theta}}_j = \sum_{l=1}^{m} \hat{\theta}_{jl}/m$.

A probabilidade de cobertura é a proporção de vezes em que o valor real do parâmetro é coberto pelos intervalos de confiança assintóticos (3.26) e (3.27) e pelos intervalos de confiança bootstrap do parâmetro λ, $IC_{95\%}(\lambda = 2) = (1{,}667; 2{,}616)$ e $IC_{95\%}(\lambda = 3) = (2{,}496; 4)$ obtidos mediante uma simulação prévia de m amostras de tamanho $n = 100$.

Algumas amostras simuladas apresentam observações muito discrepantes, fazendo com que haja números negativos na diagonal principal das inversas das matrizes de informação observada (3.17). Além disso, em alguns casos, para $n = 10$, não ocorre convergência. Nestas situações as amostras são substituídas. Os valores de EQM, variância, vício e PC dos estimadores da regressão Gucumatz são apresentados nas Tabelas 1 e 2.

Tabela 1 – Comportamento do EQM, variância, vício e PC dos estimadores da regressão Gucumatz quando $\mu = 0, \sigma = 1$ e $\lambda = 2$.

n	θ_j	Média	$EQM(\hat{\theta}_j)$	variância $(\hat{\theta}_j)$	vício $(\hat{\theta}_j)$	PC $(\hat{\theta}_j)$
10	μ	0,131	0,586	0,568	0,131	0,870
50	μ	$9,6 \cdot 10^{-4}$	0,039	0,039	$9,6 \cdot 10^{-4}$	0,919
100	μ	-0,007	0,017	0,017	0,007	0,925
200	μ	-0,001	0,008	0,008	-0,001	0,936
10	σ	1,071	0,360	0,355	0,071	0,729
50	σ	0,948	0,043	0,041	-0,052	0,725
100	σ	0,960	0,022	0,020	-0,040	0,719
200	σ	0,981	0,011	0,011	-0,019	0,759
10	λ	2,549	0,998	0,696	0,550	0,732
50	λ	2,187	0,129	0,094	0,187	0,878
100	λ	2,124	0,070	0,055	0,125	0,963
200	λ	2,064	0,036	0,032	0,064	0,980

Tabela 2 – Comportamento do EQM, variância, vício e PC dos estimadores da regressão Gucumatz quando $\mu = 0, \sigma = 1$ e $\lambda = 3$.

n	θ_j	Média	$EQM(\hat{\theta}_j)$	variância $(\hat{\theta}_j)$	vício $(\hat{\theta}_j)$	PC $(\hat{\theta}_j)$
10	μ	0,028	0,135	0,134	0,028	0,800
50	μ	$9,7 \cdot 10^{-4}$	0,022	0,022	$9,7 \cdot 10^{-4}$	0,929
100	μ	-0,002	0,011	0,011	-0,002	0,932
200	μ	-0,004	0,005	0,005	-0,004	0,949
10	σ	1,443	0,815	0,618	0,443	0,878
50	σ	1,054	0,025	0,022	0,054	0,879
100	σ	1,027	0,009	0,008	0,027	0,906
200	σ	1,012	0,004	0,003	0,012	0,924
10	λ	3,091	0,890	0,882	0,091	0,497
50	λ	3,079	0,291	0,285	0,079	0,906
100	λ	3,076	0,131	0,126	0,076	0,986
200	λ	3,026	0,050	0,049	0,026	0,998

Pelas Tabelas 1 e 2, observa-se que as medidas EQM, variância e vício dos estimadores decrescem com o aumento da amostra. Por sua vez, as probabilidades de cobertura (PC) dos estimadores crescem com o aumento de n. Portanto, esta simulação indica haver consistência dos estimadores $\hat{\mu}$, $\hat{\sigma}$ e $\hat{\lambda}$ e reforça a normalidade assintótica dos estimadores $\hat{\mu}$ e $\hat{\sigma}$, admitida em (3.24).

3.2 Estimação bayesiana

Apresenta-se um caminho para a estimação bayesiana do vetor de parâmetros $\boldsymbol{\theta} = (\mu, \sigma, \lambda)$ com utilização de métodos MCMC (Monte Carlo Cadeia de Markov) por meio dos algoritmos Metropolis-Hastings (HASTINGS, 1953) e Amostrador de Gibbs (GEMAN; GEMAN, 1987). Nesta estimação, utiliza-se função de perda quadrática, cujo estimador de Bayes é a média da distribuição atualizada. Utilizou-se o software RStudio Team (2018), cujos códigos estão no Apêndice B.

As distribuições a priori dos parâmetros μ, σ e λ são dadas por:

$$\mu|\sigma \sim N(0, \sigma^2), \quad \sigma \sim Gama(1, 100) \quad \text{e} \quad \lambda \sim U(1, 4). \tag{3.29}$$

A distribuição priori conjunta dos parâmetros μ, σ e λ é dada por:

$$\pi(\boldsymbol{\theta}) = \pi(\mu, \sigma, \lambda) = \pi(\mu|\sigma)\pi(\sigma)\pi(\lambda) = \frac{\exp(-\mu^2/2\sigma^2)}{\sigma\sqrt{2\pi}} exp(-\sigma/100)\frac{1}{3}. \tag{3.30}$$

A distribuição posteriori conjunta dos parâmetros μ, σ e λ é dada por:

$$\pi(\mu, \sigma, \lambda|\boldsymbol{y}, \boldsymbol{k}) \propto L(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{k})\pi(\mu|\sigma)\pi(\sigma)\pi(\lambda), \tag{3.31}$$

em que $L(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{k})$ é a função de verossimilhança dos dados completos, mostrada em (3.6).

O algoritmo Amostrador de Gibbs possibilita simular dados da distribuição posteriori conjunta (3.31) simulando sucessivamente das distribuições condicionais completas de cada um dos parâmetros, bem como da variável não-observável K. As distribuições condicionais completas dos parâmetros são dadas a seguir, enquanto que a distribuição condicional completa da variável K é dada em (3.8).

$$P(\mu|\sigma, \lambda, \boldsymbol{y}, \boldsymbol{k}) \propto L(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{k})\pi(\mu|\sigma)\pi(\sigma), \tag{3.32}$$

$$P(\sigma|\mu, \lambda, \boldsymbol{y}, \boldsymbol{k}) \propto L(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{k})\pi(\mu|\sigma)\pi(\sigma), \tag{3.33}$$

$$P(\lambda|\mu, \sigma, \boldsymbol{y}, \boldsymbol{k}) \propto L(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{k})\pi(\lambda). \tag{3.34}$$

Os passos do algoritmo Amostrador de Gibbs podem ser descritos como a seguir:

Algoritmo 2 – Algoritmo Amostrador de Gibbs

1. Inicialize o contador de iterações da cadeia $t = 0$;

2. Especifique os valores iniciais $\boldsymbol{\theta}^{(0)} = (\mu^{(0)}, \sigma^{(0)}, \lambda^{(0)})$;

3. Obtenha um novo valor de $\boldsymbol{\theta}^{(t+1)}$ a partir de $\boldsymbol{\theta}^{(t)}$ através da geração sucessiva dos valores

$$k_i^{(t+1)} \sim P(K_i|Y, \mu^{(t)}, \sigma^{(t)}, \lambda^{(t)}), \quad 1 \leq i \leq n; \tag{3.35}$$
$$\mu^{(t+1)} \sim P(\mu|\sigma^{(t)}, \lambda^{(t)}, \boldsymbol{y}, \boldsymbol{k}^{(t)}); \tag{3.36}$$
$$\sigma^{(t+1)} \sim P(\sigma|\mu^{(t)}, \lambda^{(t)}, \boldsymbol{y}, \boldsymbol{k}^{(t)}); \tag{3.37}$$
$$\lambda^{(t+1)} \sim P(\lambda|\mu^{(t)}, \sigma^{(t)}, \boldsymbol{y}, \boldsymbol{k}^{(t)}). \tag{3.38}$$

4. Incremente o contador de t para $t+1$, e retorne ao passo 2 até obter convergência.

Cada iteração se completa após 4 movimentos (3.35), (3.36), (3.37) e (3.38) ao longo dos eixos coordenados das componentes de $\boldsymbol{\theta}$. Após a convergência, os valores resultantes formam uma amostra da distribuição a posteriori (3.31).

De (3.32), a distribuição condicional completa de $\mu|\sigma, \lambda, \boldsymbol{k}, \boldsymbol{y}$, a menos dos termos que não dependem do parâmetro μ, é dada por:

$$P(\mu|\sigma, \lambda, \boldsymbol{k}, \boldsymbol{y}) \propto \exp\left\{\sum_{i=1}^{n} -k_i \frac{(y_i - \mu)^2}{2\sigma^2}\right\} \prod_{i=1}^{n} \left[1 + \left(\frac{y_i - \mu}{\sigma}\right)^2\right]^{-(1-k_i)} e^{\frac{-\mu^2}{2\sigma^2}}. \tag{3.39}$$

De (3.33), a distribuição condicional completa de $\sigma|\mu, \lambda, \boldsymbol{k}, \boldsymbol{y}$, a menos dos termos que não dependem do parâmetro σ, é dada por:

$$P(\sigma|\mu, \lambda, \boldsymbol{k}, \boldsymbol{y}) \propto \exp\left\{\sum_{i=1}^{n} -k_i \frac{(y_i - \mu)^2}{2\sigma^2}\right\} \left(\frac{1}{\sigma}\right)^n \prod_{i=1}^{n} \left[1 + \left(\frac{y_i - \mu}{\sigma}\right)^2\right]^{-(1-k_i)} e^{\frac{-\mu^2}{2\sigma^2}} e^{\frac{-\sigma}{100}}. \tag{3.40}$$

De (3.34), a distribuição condicional completa de $\lambda|\mu, \sigma, \boldsymbol{k}, \boldsymbol{y}$, a menos dos termos que não dependem do parâmetro λ, é dada por:

$$P(\lambda|\mu, \sigma, \boldsymbol{k}, \boldsymbol{y}) \propto (\alpha_1 \rho)^{\sum_{i=1}^{n} k_i} (1 - \alpha_1 \rho)^{\sum_{i=1}^{n} (1-k_i)} e^{-\lambda/100}. \tag{3.41}$$

Para amostrar das distribuições condicionais completas (3.39), (3.40) e (3.41), visto que não apresentam forma conhecida, pode-se utilizar passos do algoritmo Metropolis-Hastings, como a seguir. Suponha que em uma iteração t do algoritmo Amostrador de Gibbs se deseja simular um valor $\mu^{(t+1)}$ da distribuição condicional completa de $\mu|\sigma^{(t)}, \lambda^{(t)}, \boldsymbol{y}, \boldsymbol{k}^{(t)}$ (3.39). É proposto o Algoritmo:

Algoritmo 3 – Algoritmo Metropolis-Hastings

1) Gere um novo valor proposto μ' de uma distribuição auxiliar $q(\mu'|\mu^{(t)}) = U(\mu^{(t)} - 1, \mu^{(t)} + 1)$;

2) Calcule a probabilidade de aceitação $R(\mu'|\mu^{(t)}) = \min\left(1, \dfrac{P(\mu'|\sigma,\lambda,\boldsymbol{k},\boldsymbol{y}) \cdot q(\mu^{(t)}|\mu')}{P(\mu^{(t)}|\sigma,\lambda,\boldsymbol{k},\boldsymbol{y}) \cdot q(\mu'|\mu^{(t)})}\right)$ para o valor candidato μ';

3) Gere $u \sim U(0,1)$;

4) Se $u \leq R(\mu'|\mu^{(t)})$, faça $\mu^{(t+1)} = \mu'$. Caso contrário, faça $\mu^{(t+1)} = \mu^{(t)}$.

Similarmente, para simular um valor $\sigma^{(t+1)}$ da distribuição condicional completa de $\sigma|\mu^{(t)},\lambda^{(t)},\boldsymbol{y},\boldsymbol{k}^{(t)}$ dada em (3.40), um valor proposto σ' é gerado de uma distribuição auxiliar $q(\sigma'|\sigma^{(t)}) = U(10^{-2}, 10^{-2} + 2\sigma^{(t)})$, o qual é aceito com uma probabilidade $R(\sigma'|\sigma^{(t)})$, de modo que se $u \sim U(0,1)$ for tal que $u < R(\sigma'|\sigma^{(t)})$, então $\sigma^{(t+1)} = \sigma'$.

$$R(\sigma'|\sigma^{(t)}) = \min\left(1, \frac{P(\sigma'|\mu,\lambda,\boldsymbol{k},\boldsymbol{y})q(\sigma^{(t)}|\sigma')}{P(\sigma^{(t)}|\mu,\lambda,\boldsymbol{k},\boldsymbol{y})q(\sigma'|\sigma^{(t)})}\right). \tag{3.42}$$

Similarmente, para simular um valor $\lambda^{(t+1)}$ da distribuição condicional completa de $\lambda|\mu^{(t)},\sigma^{(t)},\boldsymbol{y},\boldsymbol{k}^{(t)}$ dada em (3.41), propõe-se um valor gerado de uma distribuição auxiliar $q(\lambda'|\lambda^{(t)}) = \lambda' \sim U(1,4)$ que é aceito com probabilidade $R(\lambda'|\lambda^{(t)})$, de modo que se $u \sim U(0,1)$ for tal que $u < R(\lambda'|\lambda^{(t)})$, então $\lambda^{(t+1)} = \lambda'$.

$$R(\lambda'|\lambda^{(t)}) = \min\left(1, \frac{P(\lambda'|\mu,\sigma,\boldsymbol{k},\boldsymbol{y})q(\lambda^{(t)}|\lambda')}{P(\lambda^{(t)}|\mu,\sigma,\boldsymbol{k},\boldsymbol{y})q(\lambda'|\lambda^{(t)})}\right). \tag{3.43}$$

Ao final de T iterações, sejam as cadeias de valores simulados das distribuições condicionais completas (3.39), (3.40) e (3.41) ao longo dos eixos coordenados das componentes de $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{(t)} = (\mu^{(t)}, \sigma^{(t)}, \lambda^{(t)}), \quad 0 \leq t \leq T. \tag{3.44}$$

Após descarte de B valores iniciais das cadeias para anular o efeito das tentativas iniciais $\mu^{(0)}$, $\sigma^{(0)}$ e $\lambda^{(0)}$, e imposição de saltos de tamanho S para anular correlações nas cadeias, os estimadores são dados por

$$\hat{\mu} = \frac{\sum_{t=B+1}^{T} \mu^{(t)}}{(T-B)/S}, \quad \hat{\sigma} = \frac{\sum_{t=B+1}^{T} \sigma^{(t)}}{(T-B)/S}, \quad \hat{\lambda} = \frac{\sum_{t=B+1}^{T} \lambda^{(t)}}{(T-B)/S}. \tag{3.45}$$

A seguir, é realizado um estudo de simulação para avaliar o comportamento assintótico dos estimadores $\hat{\mu}, \hat{\sigma}$ e $\hat{\lambda}$.

3.2.1 Estudo de simulação

Para avaliar o comportamento assintótico dos estimadores bayesianos, são simuladas $m = 500$ amostras Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$ de tamanhos $n = 10$, $n = 50$ e $n = 100$. Neste estudo, são avaliados raiz do erro quadrático médio (REQM), desvio padrão (DP), vício e probabilidades de cobertura (PC) do vetor de estimadores $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma}, \hat{\lambda})$ dados em (3.45). Seja o vetor de parâmetros $\boldsymbol{\theta} = (\mu, \sigma, \lambda)$ cujo elemento é θ_j em que $1 \leq j \leq 3$, o qual assume os valores reais pré-determinados $\mu = 0, \sigma = 1$, e $\lambda = 2$. São avaliados:

$$REQM(\hat{\theta}_j) = \sqrt{\frac{\sum_{l=1}^{m}(\hat{\theta}_{jl} - \theta_j)^2}{m-1}}, \quad DP(\hat{\theta}_j) = \sqrt{\frac{\sum_{l=1}^{m}(\hat{\theta}_{jl} - \bar{\hat{\theta}}_j)^2}{m-1}} \quad \text{e vício}(\hat{\theta}_j) = \frac{\sum_{l=1}^{m}(\hat{\theta}_{jl} - \theta_j)}{m-1}, \tag{3.46}$$

em que $\bar{\hat{\theta}}_j = \sum_{l=1}^{m} \hat{\theta}_{jl}/m$.

As medidas REQM, DP e vício dos estimadores, quando $n = 10, n = 50$ e $n = 100$, são apresentadas na Tabela 3.

Tabela 3 – Comportamento dos estimadores da distribuição Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$ em 500 simulações, quando $\mu = 0$, $\sigma = 1$, $\lambda = 2$.

n	θ	$REQM(\hat{\theta})$	DP$(\hat{\theta})$	Vício$(\hat{\theta})$
10	μ	0,409	0,408	0,016
10	σ	0,433	0,368	0,229
10	λ	0,549	0,406	0,370
50	μ	0,177	0,177	-0,006
50	σ	0,172	0,163	0,055
50	λ	0,306	0,298	0,070
100	μ	0,124	0,124	0,003
100	σ	0,126	0,126	0,013
100	λ	0,225	0,225	0,006

Pela Tabela 3, observa-se que o REQM, a variância e o vício dos estimadores diminuem satisfatoriamente com o aumento de n. No entanto, devido a problemas ocorridos nesta estimação bayesiana, é dada preferência à estimação frequentista na regressão Gucumatz, que apresenta bom desempenho e simplicidade, a qual é utilizada no capítulo seguinte.

Neste capítulo, apresentou-se estimação por máxima verossimilhança de dados aumentados, matriz de informação observada, intervalos de confiança assintóticos para os parâmetros μ e σ e intervalo de confiança bootstrap para o parâmetro λ. Em seguida, apresentou-se um caminho para realizar estimação bayesiana, embora, neste trabalho seja dada preferência pela estimação frequentista. Estudos de simulação indicam a consistência de todos os estimadores e a normalidade assintótica dos estimadores $\hat{\mu}$ e $\hat{\sigma}$.

CAPÍTULO

4

REGRESSÃO GUCUMATZ

Neste capítulo, apresenta-se o modelo de regressão Gucumatz $(\mu_i, \sigma, \lambda, \alpha_1, \alpha_2)$, no qual o parâmetro μ, mediana da distribuição, é modelado por covariáveis por meio de uma estrutura similar à utilizada em modelos lineares generalizados. É realizada estimação por máxima verossimilhança e são apresentadas matriz de informação observada e intervalos de confiança. Em seguida, são apresentados testes de hipótese e Critérios de Informação de Akaike (AIC) e Bayesiano (BIC) para seleção de modelos não hierárquicos. Em seguida, são apresentados resíduos quantílicos aleatorizados e medida de afastamento pela verossimilhança, para realizar diagnóstico do modelo. Por meio de estudos de simulação, avaliam-se as propriedades assintóticas dos estimadores, comparam-se modelos ajustados via distribuição Gucumatz com as distribuições Normal e Cauchy, e analisa-se a aplicabilidade das técnicas de diagnóstico. Por fim, um conjunto de dados reais é analisado utilizando-se o modelo de regressão Gucumatz.

4.1 Modelo de regressão Gucumatz

Modelos de regressão, em geral, ajustam a média ou uma função da média da variável resposta mediante covariáveis. Mas, devido à média da distribuição Gucumatz ser indefinida (Proposição 6), o parâmetro μ_i, mediana da distribuição, é modelado por covariáveis por meio de uma estrutura similar à que ocorre em modelos lineares generalizados. Neste modelo, utiliza-se função de ligação identidade, visto que $\mu_i \in \mathbb{R}$.

Sejam $Y_1, Y_2, ..., Y_n$ v.a.'s independentes, $Y_i \sim$ Gucumatz$(\mu_i, \sigma, \lambda, \alpha_1, \alpha_2)$, com realizações $y_i \in \mathbb{R}$, $1 \leq i \leq n$. Seja a matriz de planejamento $\boldsymbol{X}$, de posto completo $p+1$, cujo vetor linha $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, ..., x_{ip})$ corresponde aos valores observados das covariáveis associadas ao i-ésimo indivíduo. Seja o vetor de parâmetros desconhecidos $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma, \lambda)$, em que $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ tem dimensão $p+1$. O modelo de regressão Gucumatz $(\mu_i, \sigma, \lambda, \alpha_1, \alpha_2)$ relaciona o vetor de medianas $\boldsymbol{\mu}_{(nx1)}$, de elemento μ_i, com a parte sistemática $\boldsymbol{X\beta}_{(nx1)}$, de

elemento $\boldsymbol{x}_i^T\boldsymbol{\beta}$, por meio da seguinte relação funcional

$$\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}, \quad 1 \leq i \leq n. \tag{4.1}$$

A função densidade de Y_i é reescrita, de (2.5), considerando $\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, como segue:

$$f(Y_i = y_i) = \begin{cases} \dfrac{\alpha_1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2}\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)^2\right], & se \quad \boldsymbol{x}_i^T\boldsymbol{\beta} - \lambda\sigma \leq y_i \leq \boldsymbol{x}_i^T\boldsymbol{\beta} + \lambda\sigma \\ \dfrac{\alpha_2}{\sigma\pi\left[1 + \left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)^2\right]}, & c.c. \end{cases}, \quad \text{em que} \tag{4.2}$$

$$\alpha_1 = \left\{ \frac{e^{\frac{-\lambda^2}{2}}(1+\lambda^2)}{\sqrt{2\pi}} \left[\pi - 2\arctan(\lambda)\right] + 2\Phi(\lambda) - 1 \right\}^{-1} \quad \text{e} \quad \alpha_2 = \frac{1 - \alpha_1(2\Phi(\lambda) - 1)}{1 - \frac{2}{\pi}\arctan(\lambda)}.$$

Realiza-se estimação frequentista, de maneira igual à desenvolvida na Seção 3.1, considerando $\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$.

4.2 Estimação

O vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma, \lambda)$, de dimensão $p+3$, é estimado por máxima verossimilhança de dados completos, como na Seção 3.1. Seja a variável aleatória não-observável $K_i \sim Bernoulli(\alpha_1\rho)$ $(1 \leq i \leq n)$. Considerando $\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, (3.5) é reescrito como

$$K_i = \begin{cases} 1 & \text{se} \quad Y_i \in [\boldsymbol{x}_i^T\boldsymbol{\beta} - \lambda\sigma, \boldsymbol{x}_i^T\boldsymbol{\beta} + \lambda\sigma] \\ 0 & c.c. \end{cases}. \tag{4.3}$$

O Algoritmo Esperança-Maximização é realizado como a seguir. Supondo que se esteja na iteração $t+1$ do Algoritmo EM, a distribuição de probabilidade da variável $K_i|Y_i, \boldsymbol{\theta}^{(t)}$ $(1 \leq i \leq n)$ é

$$P(k_i = 1|y_i, \boldsymbol{\theta}^{(t)}) = \kappa_i^{(t)} = \begin{cases} 1, & se \quad \boldsymbol{x}_i^T\boldsymbol{\beta}^{(t)} - \lambda^{(t)}\sigma^{(t)} \leq y_i \leq \boldsymbol{x}_i^T\boldsymbol{\beta}^{(t)} + \lambda^{(t)}\sigma^{(t)} \\ 0, & cc. \end{cases}, \tag{4.4}$$

$$P(k_i = 0|y_i, \boldsymbol{\theta}^{(t)}) = 1 - \kappa_i^{(t)}.$$

A esperança da log-verossimilhança dos dados completos na distribuição de $K|Y, \boldsymbol{\theta}^{(t)}$ é reescrita a partir de (3.11), considerando $\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, como

$$E_{K|Y,\boldsymbol{\theta}^{(t)}}[l(\boldsymbol{\theta}^{(t)}|\boldsymbol{y},\boldsymbol{k})] = l(\boldsymbol{\theta}^{(t)}|\boldsymbol{y},\boldsymbol{\kappa}^{(t)}) = \sum_{i=1}^{n}\kappa_i^{(t)}\log(\alpha_1) + \sum_{i=1}^{n}\left[\frac{-\kappa_i}{2}\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{(t)}}{\sigma^{(t)}}\right)^2\right]$$
$$- \sum_{i=1}^{n}\kappa_i^{(t)}\log(\sigma^{(t)}\sqrt{2\pi}) + \sum_{i=1}^{n}(1-\kappa_i^{(t)})\log(\alpha_2) - \sum_{i=1}^{n}(1-\kappa_i^{(t)})\log(\sigma^{(t)}\pi)$$
$$- \sum_{i=1}^{n}(1-\kappa_i^{(t)})\log\left[1 + \left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{(t)}}{\sigma^{(t)}}\right)^2\right] + \sum_{i=1}^{n}\kappa_i^{(t)}\log(\alpha_1\rho) + (n - \sum_{i=1}^{n}\kappa_i^{(t)})\log(1-\alpha_1\rho). \tag{4.5}$$

A maximização de (4.5) é realizada dividindo-a em duas partes, e maximizando-as separadamente em $\boldsymbol{\beta}^{(t+1)}$, $\sigma^{(t+1)}$ e $\lambda^{(t+1)}$, assim como em (3.12) e (3.13).

Os passos desta estimação estão descritos no Algoritmo 5, no Apêndice A. A estimação foi realizada com auxílio do software RStudio Team (2018), cujos códigos se encontram no Apêndice B. Com base nas simulações estudadas neste trabalho, entende-se que é possível utilizar outras tentativas iniciais não muito diferentes daquelas propostas sem impactar o resultado da estimação, mas esta informação não pode ser generalizada.

4.2.1 Vetor escore e matriz de informação observada

O vetor escore $\boldsymbol{U}_\theta = (U_{\boldsymbol{\beta}}^T, U_\sigma, U_\lambda)$ é o gradiente da função de log-verossimilhança dos dados completos (3.7), considerando $\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, em relação ao vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma, \lambda)$, em que $\boldsymbol{\beta}^T = (\beta_0, \beta_1, ..., \beta_p)^T$. Os elementos U_{β_r}, $r = 0, 1, ..., p$, de $U_{\boldsymbol{\beta}}^T$, e U_σ do vetor escore são dados a seguir, enquanto que U_λ é dado em (3.16). A variável não-observável K_i, por sua vez, é definida em (4.3).

$$U_{\beta_r} = \frac{1}{\sigma}\sum_{i=1}^{n}k_i\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)x_{ir} + \frac{2}{\sigma}\sum_{i=1}^{n}(1-k_i)\frac{\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)x_{ir}}{1 + \left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)^2}, \tag{4.6}$$

e

$$U_\sigma = \frac{1}{\sigma}\sum_{i=1}^{n}k_i\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)^2 - \frac{n}{\sigma} + \frac{2}{\sigma}\sum_{i=1}^{n}(1-k_i)\frac{\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)^2}{1 + \left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)^2}. \tag{4.7}$$

A matriz de informação observada é dada pelo negativo do gradiente do vetor escore,

$$\boldsymbol{J}(\boldsymbol{\theta}) = \left[-\frac{\partial \boldsymbol{U}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right] = -\begin{bmatrix} \boldsymbol{J}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \boldsymbol{J}_{\boldsymbol{\beta}\sigma} & \boldsymbol{J}_{\boldsymbol{\beta}\lambda} \\ & J_{\sigma\sigma} & J_{\sigma\lambda} \\ & & J_{\lambda\lambda} \end{bmatrix}, \tag{4.8}$$

cujos elementos são dados como segue. O elemento $J_{\lambda\lambda}$ é dado em (3.21), $\boldsymbol{J}_{\boldsymbol{\beta}\lambda} = 0$, $J_{\sigma\lambda} = 0$. Por sua vez, $J_{\sigma\sigma}$ é reescrito a partir de (3.20) considerando $\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, o elemento $J_{\beta_r\beta_s}$, $r,s = 0,1,...,p$, da matriz bloco diagonal $\boldsymbol{J}_{\boldsymbol{\beta}\boldsymbol{\beta}}$, e o elemento $J_{\beta_r\sigma}$, do vetor $\boldsymbol{J}_{\boldsymbol{\beta}\sigma}$, são dados a seguir.

$$J_{\beta_r\beta_s} = \frac{-1}{\sigma^2}\sum_{i=1}^{n} k_i x_{ir}x_{is} + \frac{2}{\sigma^2}\sum_{i=1}^{n}(1-k_i)x_{ir}x_{is}\frac{-1+\left(\frac{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)^2}{\left[1+\left(\frac{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)^2\right]^2}, \tag{4.9}$$

$$J_{\beta_r\sigma} = \frac{-2}{\sigma^2}\sum_{i=1}^{n} k_i x_{ir}\left(\frac{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right) - \frac{4}{\sigma^2}\sum_{i=1}^{n}(1-k_i)\frac{\left(\frac{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)x_{ir}}{\left[1+\left(\frac{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)^2\right]^2}, \tag{4.10}$$

e

$$J_{\sigma\sigma} = \frac{-3}{\sigma^2}\sum_{i=1}^{n} k_i\left(\frac{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)^2 + \frac{n}{\sigma^2} - \frac{2}{\sigma^2}\sum_{i=1}^{n}(1-k_i)\frac{3\left(\frac{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)^2+\left(\frac{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)^4}{\left[1+\left(\frac{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)^2\right]^2}. \tag{4.11}$$

Nas seções seguintes, são apresentados intervalos de confiança, testes de hipótese, seleção de modelos, técnicas para diagnóstico e são realizados estudos de simulações.

4.3 Intervalos de confiança assintóticos

Nos termos da Subseção 3.1.2, considerando $\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ e $\hat{\beta}_r$, r-ésimo elemento do vetor $\hat{\boldsymbol{\beta}}$, pode-se considerar como distribuições assintóticas dos estimadores $\hat{\beta}_r$ e $\hat{\sigma}$

$$\hat{\beta}_r \sim^a N(\beta_r, [\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}]_{r,r}) \quad \text{e} \quad \hat{\sigma} \sim^a N(\sigma, [\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}]_{p+2,p+2}), \tag{4.12}$$

em que $[\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}]_{r,r}$ e $[\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}]_{p+2,p+2}$ são os elementos de posição (r,r) e $(p+2,p+2)$, respectivamente, da inversa da matriz de informação observada (4.8). Assim, os intervalos de confiança assintóticos dos parâmetros β_r e σ, com nível de confiança $100(1-\alpha)\%$, são dados por:

$$IC_{1-\alpha}(\beta_r) = \left(\hat{\beta}_r - z_{(1-\alpha/2)}\sqrt{[\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}]_{r,r}}, \quad \hat{\beta}_r + z_{(1-\alpha/2)}\sqrt{[\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}]_{r,r}}\right), \tag{4.13}$$

$$IC_{1-\alpha}(\sigma) = \left(\hat{\sigma} - z_{(1-\alpha/2)}\sqrt{[\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}]_{p+2,p+2}}, \quad \hat{\sigma} + z_{(1-\alpha/2)}\sqrt{[\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}]_{p+2,p+2}}\right), \tag{4.14}$$

em que $z_{(1-\alpha/2)}$ é o quantil $1-\alpha/2$ da função distribuição de probabilidade Normal padrão.

4.4 Testes de hipótese assintóticos

Testes de hipótese assintóticos podem ser realizados para o vetor de parâmetros $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$, de dimensão $p+1$, e para o parâmetro σ com base na normalidade assintótica dos estimadores $\hat{\boldsymbol{\beta}}$ e $\hat{\sigma}$, conforme (4.12).

Seja a hipótese nula $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$. A estatística da razão de verossimilhanças é escrita como a seguir, e de forma análoga para uma hipótese nula $H_0 : \sigma = \sigma_0$.

$$ERV = 2\{l(\hat{\boldsymbol{\beta}}|\boldsymbol{y}) - l(\boldsymbol{\beta}_0|\boldsymbol{y})\},$$

em que $l(\hat{\boldsymbol{\beta}}|\boldsymbol{y})$ é a função de log-verossimilhança (3.4) avaliada em $\hat{\boldsymbol{\beta}}$ e $l(\boldsymbol{\beta}_0|\boldsymbol{y})$ é a função de log-verossimilhança (3.4) avaliada em $\boldsymbol{\beta}_0$, considerando $\mu = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$.

$$l(\hat{\boldsymbol{\beta}}|\boldsymbol{y}) = \sum_{i=1}^{n} log \begin{cases} \dfrac{\hat{\alpha}_1}{\hat{\sigma}\sqrt{2\pi}} \exp\left[\dfrac{-1}{2}\left(\dfrac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)^2\right], & \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} - \hat{\lambda}\hat{\sigma} \leq y_i \leq \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} + \hat{\lambda}\hat{\sigma} \\ \dfrac{\hat{\alpha}_2}{\hat{\sigma}\pi\left[1 + \left(\dfrac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)^2\right]}, & c.c. \end{cases},$$

$$l(\boldsymbol{\beta}_0|\boldsymbol{y}) = \sum_{i=1}^{n} log \begin{cases} \dfrac{\hat{\alpha}_1}{\hat{\sigma}\sqrt{2\pi}} \exp\left[\dfrac{-1}{2}\left(\dfrac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_0}{\hat{\sigma}}\right)^2\right], & \boldsymbol{x}_i^T\boldsymbol{\beta}_0 - \hat{\lambda}\hat{\sigma} \leq y_i \leq \boldsymbol{x}_i^T\boldsymbol{\beta}_0 + \hat{\lambda}\hat{\sigma} \\ \dfrac{\hat{\alpha}_2}{\hat{\sigma}\pi\left[1 + \left(\dfrac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_0}{\hat{\sigma}}\right)^2\right]}, & c.c. \end{cases}.$$

A estatística de Wald para a hipótese nula $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ pode ser escrita como a seguir, e de forma análoga para $H_0 : \sigma = \sigma_0$.

$$EW = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \boldsymbol{J}(\hat{\boldsymbol{\theta}})_{\boldsymbol{\beta}\boldsymbol{\beta}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

em que $\boldsymbol{J}(\hat{\boldsymbol{\theta}})_{\boldsymbol{\beta}\boldsymbol{\beta}}$ é a matriz bloco diagonal principal de dimensão $p+1$ da matriz de informação observada de Fisher, isto é, o termo $\boldsymbol{J}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ da matriz em (4.8) avaliada em $\hat{\boldsymbol{\theta}}$. Estas estatísticas, sob hipótese nula, tem distribuição assintótica χ^2 com graus de liberdade igual à quantidade de restrições de parâmetros impostas por H_0, conforme Wasserman (2004) e Dobson (2002). Assim, para a hipótese nula $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, pode-se rejeitar H_0 se o valor observado da estatística é superior a $\chi^2_{p+1,1-\alpha}$.

4.5 Seleção de modelos

Seleções entre modelos não hierárquicos podem ser realizadas com utilização do Critério de Informação de Akaike (AIC), de Akaike (1974), e do Critério de Informação Bayesiano (BIC), de Schwarz (1978). Estes critérios indicam o modelo que apresenta ajuste com menor perda de informação, que são os modelos com menores valores de AIC e BIC. Visto que o total de parâmetros do modelo é $p+3$, as medidas AIC e BIC podem ser escritas como a seguir.

$$AIC = -2l(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + 2(p+3),$$

e

$$BIC = -2l(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + (p+3)\log(n),$$

em que $l(\hat{\boldsymbol{\theta}})$ é a função de log-verossimilhança, dada em (3.4), avaliada nas estimativas $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}, \hat{\lambda})$, considerando $\mu_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$.

Neste trabalho, os critérios AIC e BIC são utilizados para comparar modelos ajustados por distribuição Gucumatz com modelos ajustados nas distribuições Normal e Cauchy.

4.6 Técnicas para diagnósticos

Ajustando-se dados ao modelo de regressão Gucumatz $(\mu_i, \sigma, \lambda, \alpha_1, \alpha_2)$, é necessário utilizar técnicas que permitam diagnosticar possíveis anomalias no modelo, como falsa distribuição para a variável resposta, existência de discrepâncias que sejam aberrantes, presença de autocorrelação ou heterocedasticidade, ou presença de observações que exercem influência significativa nas estimativas dos parâmetros.

4.6.1 Resíduo quantílico aleatorizado

Os resíduos são importantes para detectar a presença de observações aberrantes que devem ser estudadas detalhadamente. O resíduo para a i-ésima observação pode ser definido como uma função do tipo $res_i = res(y_i, \hat{\mu}_i)$ que procura medir a discrepância entre o valor observado y_i e o valor ajustado $\hat{\mu}_i$ da i-ésima observação (CORDEIRO; DEMETRIO, 2008). Neste trabalho, utiliza-se o resíduo quantílico aleatorizado, proposto por Dunn e Smyth (1996). Seja $F(y_i|\hat{\boldsymbol{\theta}})$ a probabilidade de função de distribuição acumulada contínua para a observação y_i. O resíduo quantílico aleatorizado res_i para a i-ésima observação y_i pode ser escrito como a seguir:

$$res_i = \Phi^{-1}[F(y_i|\hat{\boldsymbol{\theta}})], \tag{4.15}$$

em que $\Phi^{-1}(.)$ é a inversa da função de distribuição acumulada normal padrão e $F(y_i|\hat{\boldsymbol{\theta}})$ é a função de distribuição acumulada Gucumatz dada em (2.7), considerando $\hat{\mu}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$. A distribuição de res_i converge para a normalidade se o vetor $\boldsymbol{\theta}$ é consistentemente estimado (DUNN; SMYTH, 1996). Demonstra-se, em Sadeghpour (2016), que (4.15) tem distribuição Normal padrão se $F(y_i|\hat{\boldsymbol{\theta}})$ é a verdadeira função de distribuição da variável resposta, qualquer que seja a distribuição $F(.)$.

Dunn e Smyth (1996) afirmam que o resíduo quantílico aleatorizado pode ser empregado em qualquer método de diagnóstico que use resíduos. Pode-se, portanto, utilizá-lo em técnicas gráficas, como em Chatterjee e Hadi (1988) e Demetrio e Zocchi (2011), da seguinte forma:

a) Gráfico de resíduos versus índices das observações: é utilizado para localizar observações com resíduo grande. Se os pontos formam padrão, o gráfico indica autocorrelação.

b) Gráfico de resíduos versus valores ajustados: é desejável que os pontos se distribuam aleatoriamente com média 0 e amplitude constante. Se a dispersão dos pontos cresce à medida que aumenta o valor ajustado, há indicação de heterocedasticidade.

4.6.2 Influência global

Uma etapa importante na análise de um ajuste de regressão é a verificação da existência de observações discrepantes com alguma interferência desproporcional ou inferencial nos resultados do ajuste. (PAULA, 2013). Um recurso bastante utilizado em medidas de avaliação de pontos influentes é a técnica da deleção de pontos. Pode-se utilizar uma medida de influência global denominada afastamento pela verossimilhança, denotada por LD_i, Esta medida reúne, em uma estrutura mais geral, duas medidas de influência global tradicionais, que são distância de Cook e $DFFITS_i$ (COOK; PENA; WEISBERG, 1988).

Seja $\hat{\boldsymbol{\theta}}$ a estimativa de $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma, \lambda)$ por máxima verossimilhança e $\hat{\boldsymbol{\theta}}_{(-i)}$ a estimativa mediante a retirada da observação i. O afastamento pela verossimilhança mede o afastamento entre $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_{(-i)}$ por meio da diferença entre as funções de log-verossimilhança avaliadas em $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_{(-i)}$, respectivamente. Pode ser escrita como

$$LD_i = 2\{l(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_{(-i)}|\boldsymbol{y})\}, \quad 1 \leq i \leq n, \tag{4.16}$$

em que $l(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ é a função de log-verossimilhança dada em (3.4), avaliada em $\hat{\boldsymbol{\theta}}$, e $l(\hat{\boldsymbol{\theta}}_{(-i)}|\boldsymbol{y})$ é a função de log-verossimilhança avaliada em $\hat{\boldsymbol{\theta}}$ mediante a retirada da observação i, dada por

$$l(\hat{\boldsymbol{\theta}}_{(-i)}|\boldsymbol{y}) = \sum_{i=1}^{n} log \begin{cases} \dfrac{\hat{\alpha}_{1(-i)}}{\hat{\sigma}_{(-i)}\sqrt{2\pi}} \exp\left[\dfrac{-1}{2}\left(\dfrac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(-i)}}{\hat{\sigma}_{(-i)}}\right)^2\right], & \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\lambda}_{(-i)}\hat{\sigma}_{(-i)} \leq y_i \leq \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(-i)} + \hat{\lambda}_{(-i)}\hat{\sigma}_{(-i)} \\ \dfrac{\hat{\alpha}_{2(-i)}}{\hat{\sigma}_{(-i)}\pi\left[1 + \left(\dfrac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(-i)}}{\hat{\sigma}_{(-i)}}\right)^2\right]}, & c.c. \end{cases}.$$

Visto que a distribuição Gucumatz satisfaz parcialmente as condições de regularidade, admite-se que o vetor $\hat{\boldsymbol{\theta}}$ apresenta normalidade assintótica (3.25) com alguma deficiência, como

visto na Subseção 3.1.2. Assim, *LDi*, mostrada em (4.16), pode ser interpretada em termos de uma região de confiança assintótica $\{\hat{\boldsymbol{\theta}} : 2[l(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - l(\boldsymbol{\theta}|\boldsymbol{y})] \leq \chi^2_{p+3,\alpha}\}$, segundo Cook e Weisberg (1982), em que $\chi^2_{p+3,\alpha}$ é o ponto de probabilidade α da cauda superior da distribuição qui-quadrado com $p+3$ graus de liberdade. No entanto, visto que pode ocorrer $LDi < 0$ para alguma observação i, adota-se, como alternativa para (4.16), uma aproximação por meio da expansão de Taylor de $l(\hat{\boldsymbol{\theta}}_{(-i)}|\boldsymbol{y})$ ao redor de $\hat{\boldsymbol{\theta}}$, apresentada em Cook e Weisberg (1982)

$$LDi \approx (\hat{\boldsymbol{\theta}}_{(-i)} - \hat{\boldsymbol{\theta}})^T \boldsymbol{J}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{(-i)} - \hat{\boldsymbol{\theta}}), \tag{4.17}$$

em que $\boldsymbol{J}(\hat{\boldsymbol{\theta}})$ é a matriz de informação observada (4.8) avaliada em $\hat{\boldsymbol{\theta}}$. Neste trabalho, a *LDi*, dada em (4.17), é comparada com quantis da distribuição χ^2_{p+3}, de modo que uma observação y_i é considerada influente quando $LD_i > \chi^2_{p+3,1-\alpha}$.

4.7 Estudos de Simulação

Nesta seção, valores da regressão Gucumatz $(\mu_i, \sigma, \lambda, \alpha_1, \alpha_2)$ são simulados, conforme Algoritmo 1, para verificação das propriedades assintóticas dos estimadores, para comparar modelos Gucumatz, Normal e Cauchy ajustados a partir de dados simulados da distribuição Gucumatz, e para verificar a aplicabilidade de técnicas de diagnóstico apresentadas na Seção anterior.

4.7.1 Propriedades assintóticas dos estimadores

Para avaliar o comportamento assintótico dos estimadores de máxima verossimilhança, são simuladas $m = 1.000$ amostras Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$ de tamanhos $n = 10$, $n = 50$, $n = 100$ e $n = 200$. Duas covariáveis, $X_1 \sim U(1,100)$ e $X_2 \sim N(50,10)$ são consideradas. Neste estudo, são avaliados erro quadrático médio (EQM), variância, vício e probabilidades de cobertura (PC) do vetor de estimadores $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}, \hat{\lambda})$. Seja o vetor de parâmetros $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \sigma, \lambda)$ cujo elemento é θ_j em que $1 \leq j \leq 5$, o qual assume os valores reais pré-determinados $\beta_0 = 5, \beta_1 = -3, \beta_2 = 3, \sigma = 1$, e $\lambda = 2,5$. São avaliados:

$$EQM(\hat{\theta}_j) = \frac{\sum_{l=1}^{m}(\hat{\theta}_{jl} - \theta_j)^2}{m-1}, \quad var(\hat{\theta}_j) = \frac{\sum_{l=1}^{m}(\hat{\theta}_{jl} - \bar{\hat{\theta}}_j)^2}{m-1} \quad \text{e vício}(\hat{\theta}_j) = \frac{\sum_{l=1}^{m}(\hat{\theta}_{jl} - \theta_j)}{m-1},$$

em que $\bar{\hat{\theta}}_j = \dfrac{\sum_{l=1}^{m}\hat{\theta}_{jl}}{m}$, $l = 1, 2, ..., m$.

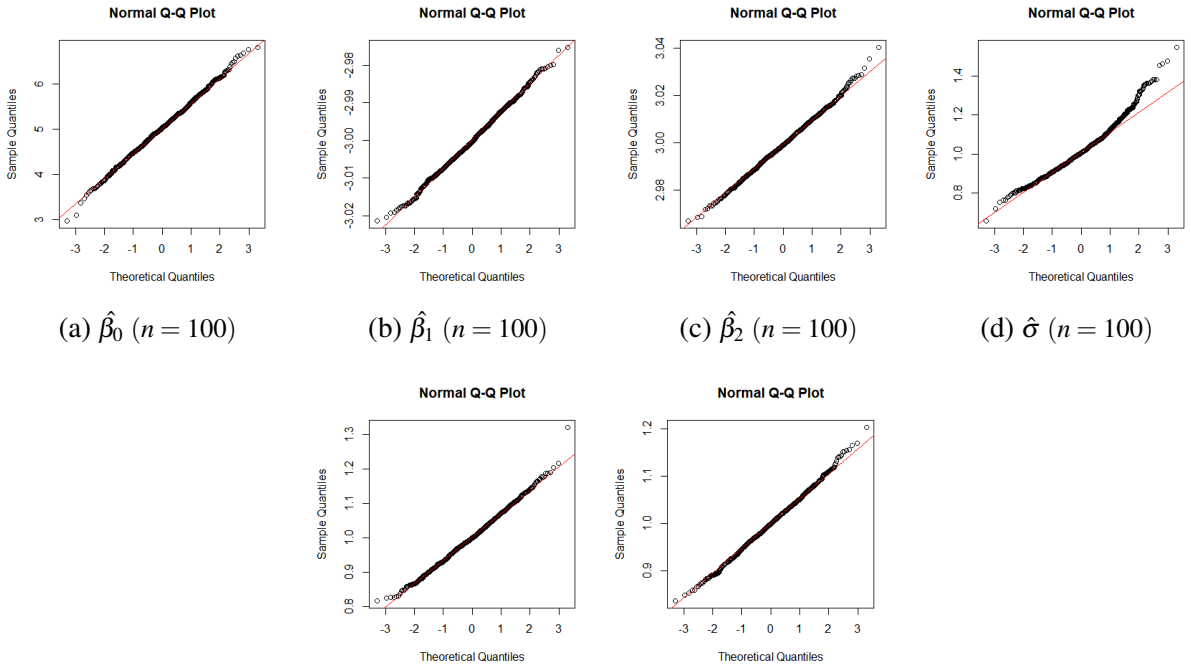
A probabilidade de cobertura é a proporção de vezes em que o valor real do parâmetro é coberto pelos intervalos de confiança assintóticos apresentados na Seção 4.3 e intervalo de confiança bootstrap $IC_{95\%}(\lambda = 2,5) = (2,176; 3,121)$, obtido por meio de simulação prévia de m estimativas $\hat{\lambda}$, de m amostras de tamanho $n = 100$.

Algumas amostras simuladas apresentam observações muito discrepantes, fazendo com que haja números negativos na diagonal principal das inversas das matrizes de informação observada (4.8). Além disso, em alguns casos, para $n = 10$, não ocorre convergência. Nestas situações as amostras são substituídas. Os valores de EQM, variância, vício e PC dos estimadores da regressão Gucumatz são apresentados na Tabelas 4.

Tabela 4 – Comportamento do EQM, variância, vício e PC dos estimadores da regressão Gucumatz quando $\beta_0 = 5; \beta_1 = -3; \beta_2 = 3; \sigma = 1; \lambda = 2,5$.

n	θ_j	Média	$EQM(\hat{\theta}_j)$	variância$(\hat{\theta}_j)$	vício$(\hat{\theta}_j)$	PC$(\hat{\theta}_j)$
10	β_0	5,368	161,8	161,6	0,368	0,824
50	β_0	5,058	1,142	1,138	0,058	0,913
100	β_0	4,973	0,389	0,388	-0,027	0,930
200	β_0	4,973	0,210	0,209	-0,027	0,924
10	β_1	-3,016	0,160	0,160	-0,016	0,837
50	β_1	-3,001	$1,8 \cdot 10^{-4}$	$1,8 \cdot 10^{-4}$	$-5,9 \cdot 10^{-4}$	0,914
100	β_1	-2,999	$7,2 \cdot 10^{-5}$	$7,2 \cdot 10^{-5}$	$5,4 \cdot 10^{-4}$	0,922
200	β_1	-3,000	$3,2 \cdot 10^{-5}$	$3,2 \cdot 10^{-5}$	$2,5 \cdot 10^{-4}$	0,918
10	β_2	3,005	0,121	0,121	0,005	0,824
50	β_2	2,999	$3,2 \cdot 10^{-4}$	$3,2 \cdot 10^{-4}$	$-6,6 \cdot 10^{-4}$	0,918
100	β_2	3,000	$1,3 \cdot 10^{-4}$	$1,3 \cdot 10^{-4}$	$1,6 \cdot 10^{-4}$	0,928
200	β_2	3,000	$6,5 \cdot 10^{-5}$	$6,5 \cdot 10^{-5}$	$3,9 \cdot 10^{-4}$	0,939
10	σ	0,934	0,363	0,359	-0,066	0,536
50	σ	0,991	0,028	0,028	-0,009	0,794
100	σ	1,004	0,014	0,014	0,004	0,838
200	σ	1,003	0,006	0,006	0,007	0,846
10	λ	3,899	2,109	0,148	1,409	0,062
50	λ	2,686	0,149	0,114	0,186	0,953
100	λ	2,590	0,068	0,060	0,090	0,954
200	λ	2,543	0,030	0,029	0,043	0,992

Na Tabela 4, observa-se que as medidas EQM, variâncias e vícios dos estimadores decrescem com o aumento da amostra e as probabilidades de cobertura (PC) dos estimadores crescem com o aumento do tamanho n da amostra. Portanto, os estimadores são consistentes, e verifica-se a normalidade assintótica (4.12) dos estimadores $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ e $\hat{\sigma}$, a qual pode ser também observada via gráficos Quantil-Quantil normal, na Figura 8.



(a) $\hat{\beta}_0$ $(n = 100)$ (b) $\hat{\beta}_1$ $(n = 100)$ (c) $\hat{\beta}_2$ $(n = 100)$ (d) $\hat{\sigma}$ $(n = 100)$

(e) $\hat{\sigma}$ $(n = 200)$ (f) $\hat{\sigma}$ $(n = 300)$

Figura 8 – Gráficos Q-Q normal das estimativas simuladas referentes aos estimadores (a) $\hat{\beta}_0$, (b) $\hat{\beta}_1$, (c) $\hat{\beta}_2$ e (d) $\hat{\sigma}$, quando $n = 100$, e referente ao estimador $\hat{\sigma}$, quando (e) $n = 200$ e (f) $n = 300$.

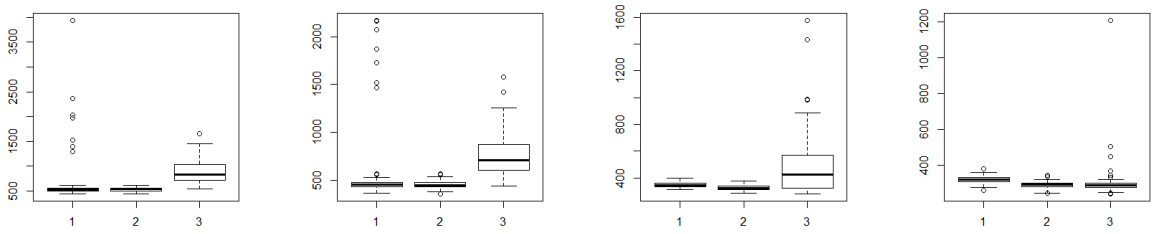
Pelas Figuras 8a, 8b e 8c, pode-se notar que, aparentemente, os estimadores de $\hat{\beta}_0, \hat{\beta}_1$ e $\hat{\beta}_2$, convergem para a normalidade em amostras de tamanho a partir de $n = 100$. Por sua vez, pelas Figuras 8d, 8e e 8f, pode-se dizer que, aparentemente, o estimador $\hat{\sigma}$ converge para a normalidade em amostras de tamanho a partir de $n = 300$.

4.7.2 Comparação entre modelos Gucumatz, Normal e Cauchy ajustados a dados Gucumatz

São comparados modelos de regressão Gucumatz $(\mu_i, \sigma, \lambda, \alpha_1, \alpha_2)$, Normal (μ_i, σ^2), apresentado na Seção 1.2, e Cauchy (μ_i, σ), apresentado na Seção 1.3, ajustados a dados simulados da distribuição Gucumatz, por meio do critério AIC em seleção de modelos.

Para isto, foram simulados 4 grupos de 100 amostras Gucumatz $(\mu_i, \sigma, \lambda, \alpha_1, \alpha_2)$ de tamanho $n = 100$, com valores de parâmetros reais pré-determinados como na simulação anterior $\beta_0 = 5, \beta_1 = -3, \beta_2 = 3, \sigma = 1$ e duas covariáveis $X_1 \sim U(1, 100)$ e $X_2 \sim N(50, 10)$. Cada grupo de 100 amostras foi simulado mediante um valor de λ pré-determinado $\lambda = 1, \lambda = 2, \lambda = 3, \lambda = 4$.

Cada amostra simulada foi ajustada aos modelos de regressão Cauchy, Gucumatz e Normal, e os AIC's de cada ajuste foram plotados em 3 diagramas de caixa (1, 2 e 3) referentes aos modelos Cauchy (1), Gucumatz (2) e Normal (3), nesta ordem, sobre dados simulados da distribuição Gucumatz com $\lambda = 1$ (9a), $\lambda = 2$ (9b), $\lambda = 3$ (9c) e $\lambda = 4$ (9d).



(a) Simulação $\lambda = 1$ (b) Simulação $\lambda = 2$ (c) Simulação $\lambda = 3$ (d) Simulação $\lambda = 4$

Figura 9 – Diagramas de caixa formados por AIC's de ajustes de dados simulados da distribuição Gucumatz mediante (a) $\lambda = 1$, (b) $\lambda = 2$, (c) $\lambda = 3$ e (d) $\lambda = 4$ aos modelos Cauchy (1), Gucumatz (2) e Normal (3).

Pelos diagramas de caixa 1, 2 e 3 formados por AIC's de ajustes de dados simulados da distribuição Gucumatz aos modelos Cauchy (1), Gucumatz (2) e Normal (3) sobre amostras Gucumatz simuladas com $\lambda = 1$ (Figura 9a), $\lambda = 2$ (Figura 9b), $\lambda = 3$ (Figura 9c), $\lambda = 4$ (Figura 9d), nota-se que, aparentemente, os AIC's referentes aos ajustes Gucumatz apresentam valores menores ou iguais aos AIC's referentes aos modelos Cauchy e Normal. Isto significa que, visualmente, o critério AIC identifica corretamente o modelo verdadeiro Gucumatz, para dados que seguem a distribuição Gucumatz. Além disso, quando $\lambda = 1$ (Figura 9a), os AIC's dos ajustes Gucumatz se igualam aos AIC's dos ajustes Cauchy exceto por 4 amostras. Similarmente, quando $\lambda = 4$, os AIC's dos ajustes Gucumatz se igualam aos AIC's dos ajustes Normal exceto por 4 amostras. Isto indica que, quando λ decresce, a distribuição Gucumatz tende à Cauchy, e se $\lambda > 4$, a distribuição Gucumatz tende à Normal.

4.7.3 Aplicabilidade de técnicas de diagnóstico

Simula-se um conjunto de dados com variável resposta Y_i que segue distribuição Gucumatz $(\mu_i, \sigma, \lambda, \alpha_1, \alpha_2)$ mediante valores de parâmetros reais pré-determinados, $\beta_0 = 5, \beta_1 = -3, \beta_2 = 3, \sigma = 1$ e $\lambda = 2,5$ e duas covariáveis $X_1 \sim U(1,100)$ e $X_2 \sim N(50,10)$, como na Subseção 4.7.1. O conjunto de dados simulados é ajustado ao modelo Gucumatz $\boldsymbol{\mu} = \beta_0 + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2$ mediante pequenas perturbações nos dados, e os modelos ajustados são avaliados com utilização de técnicas para diagnósticos apresentadas na Seção 4.6.

Uma descrição do conjunto de dados encontra-se na Tabela 5.

Tabela 5 – Descrição do conjunto de dados simulado.

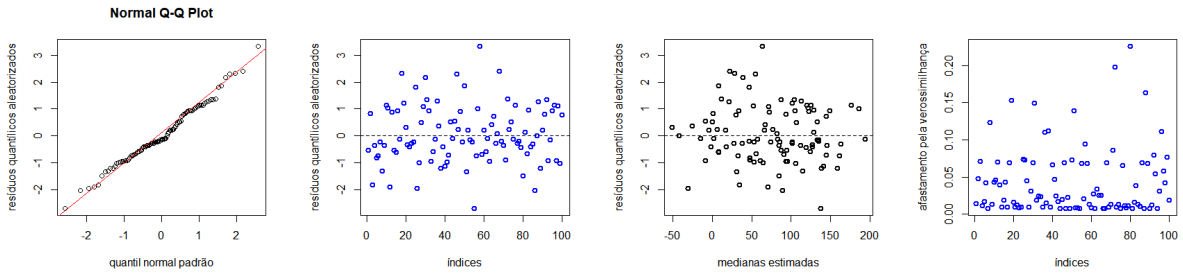
	Mín	1º quartil	Mediana	Média	3º quartil	Máx
$\boldsymbol{y}$	-50,41	34,13	82,08	78,66	121,67	456,00
$\boldsymbol{x}_1$	1,11	12,00	24,47	25,04	40,16	49,70
$\boldsymbol{x}_2$	26,45	41,13	49,33	48,23	54,81	77,23

O conjunto de dados descrito na Tabela 5 é ajustado ao modelo Gucumatz e as estimativas e intervalos de confiança são apresentados na Tabela 6. Utiliza-se $IC_{95\%}(\lambda = 2,5) = (2,176; 3,121)$, como na Subseção 4.7.1.

Tabela 6 – Ajuste do conjunto de dados ao modelo Gucumatz $(\mu_i, \sigma, \lambda, \alpha_1, \alpha_2)$.

$\boldsymbol{\theta}$	Estimativas	IC (95%)
$\beta_0 = 5$	5,696	(4,218; 7,174)
$\beta_1 = -3$	-2,997	(-3,169; -2,825)
$\beta_2 = 3$	2,988	(2,784; 3,192)
$\sigma = 1$	1,087	(0,507; 1,667)
$\lambda = 2,5$	2,345	(2,176; 3,121)

Para diagnóstico, são apresentados, na Figura 10, os gráficos quantil-quantil normal dos resíduos quantílicos aleatorizados (Figura 10a), resíduos quantílicos aleatorizados contra índices das observações (Figura 10b), resíduos quantílicos aleatorizados contra valores ajustados $\hat{\boldsymbol{\mu}}$ (Figura 10c) e afastamento pela verossimilhança das observações (Figura 10d).



(a) (b) (c) (d)

Figura 10 – (a) quantil-quantil normal dos resíduos quantílicos aleatorizados, (b) resíduos quantílicos aleatorizados contra índices das observações, (c) resíduos quantílicos aleatorizados contra $\hat{\boldsymbol{\mu}}$, e (d) afastamento pela verossimilhança das observações.

A Figura 10a indica que o modelo é verdadeiro, devido aparente normalidade dos resíduos. As Figuras 10b e 10c mostram dispersão aleatória dos resíduos ao redor da origem e a Figura 10d indica que não há observações influentes, visto que nenhum ponto é superior a $\chi^2_{5;0,95} = 11,0705$.

Efetua-se uma perturbação nestes dados, de modo que $y_{100} = y_{100} + 3 \cdot sd(\boldsymbol{y})$, $x_{50;1} = x_{50;1} + 3 \cdot sd(\boldsymbol{x}_1)$ e $x_{75;2} = x_{75;2} + 3 \cdot sd(\boldsymbol{x}_2)$, denominando-se *sd* por desvio padrão amostral. A descrição deste conjunto de dados sob perturbação está na Tabela 7.

Tabela 7 – Descrição do conjunto de dados simulados sob perturbação.

	Mín	1º quartil	Mediana	Média	3º quartil	Máx
$\boldsymbol{y}$	-50,41	36,66	82,51	80,65	122,94	456,00
$\boldsymbol{x}_1$	1,11	12,00	24,47	25,50	40,16	89,39
$\boldsymbol{x}_2$	26,45	41,13	49,47	48,54	54,94	77,23

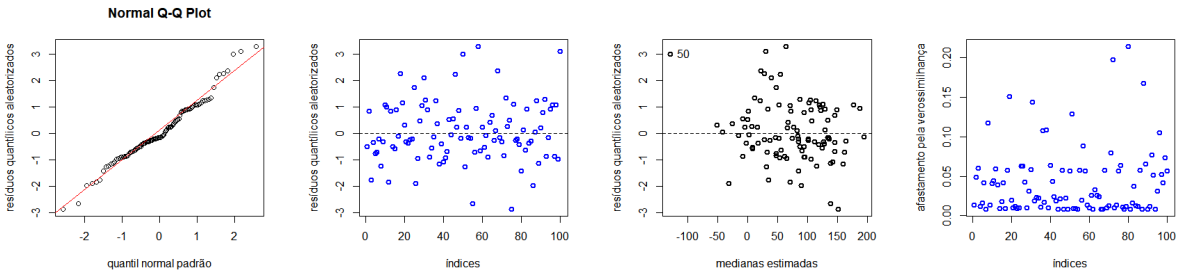
O ajuste deste conjunto de dados é apresentado na Tabela 8.

Tabela 8 – Ajuste do conjunto de dados sob primeira perturbação ao modelo Gucumatz $(\mu_i, \sigma, \lambda, \alpha_1, \alpha_2)$.

$\boldsymbol{\theta}$	Estimativas	IC (95%)
$\beta_0 = 5$	5,664	(4,164; 7,164)
$\beta_1 = -3$	-2,998	(-3,171; -2,824)
$\beta_2 = 3$	2,989	(2,782; 3,196)
$\sigma = 1$	1,111	(0,517; 1,706)
$\lambda = 2,5$	2,257	(2,176; 3,121)

A Tabela 8 indica que as mudanças nas estimativas foram irrelevantes, de modo que os intervalos de confiança ainda incluem os verdadeiros valores dos parâmetros $\beta_0, \beta_1, \beta_2$ e σ, e a estimativa do parâmetro λ está contida no respectivo intervalo de confiança.

Para diagnóstico, os seguintes gráficos são apresentados na Figura 11.



(a) (b) (c) (d)

Figura 11 – (a) quantil-quantil normal dos resíduos quantílicos aleatorizados, (b) resíduos quantílicos aleatorizados contra índices das observações, (c) resíduos quantílicos aleatorizados contra $\hat{\boldsymbol{\mu}}$, e (d) afastamento pela verossimilhança.

A Figura 11a indica que o modelo é verdadeiro, devido aparente normalidade dos resíduos. A Figura 11b apresenta dispersão aleatória dos pontos ao redor da origem, a Figura 11c apresenta dispersão aleatória dos pontos ao redor da origem, exceto pelo ligeiro padrão causado pelos resíduos da observação 50. Por sua vez, a Figura 11d indica que não há observações influentes, visto que nenhum ponto é superior a $\chi^2_{5;0,95} = 11,07$.

Efetua-se perturbação pela segunda vez nestes dados, como anteriormente, de modo que $y_{100} = y_{100} + 3 \cdot sd(\boldsymbol{y})$, $x_{50;1} = x_{50;1} + 3 \cdot sd(\boldsymbol{x}_1)$ e $x_{75;2} = x_{75;2} + 3 \cdot sd(\boldsymbol{x}_2)$, denominando-se *sd*

por desvio padrão. A descrição deste conjunto de dados sob segunda perturbação está na Tabela 9.

Tabela 9 – Descrição do conjunto de dados simulado após segunda perturbação.

	Mín	1º quartil	Mediana	Média	3º quartil	Máx
$\boldsymbol{y}$	-50,41	36,66	82,51	82,70	122,94	456,00
$\boldsymbol{x}_1$	1,11	12,00	24,47	25,99	40,16	138,59
$\boldsymbol{x}_2$	26,45	41,13	49,47	48,86	54,94	105,06

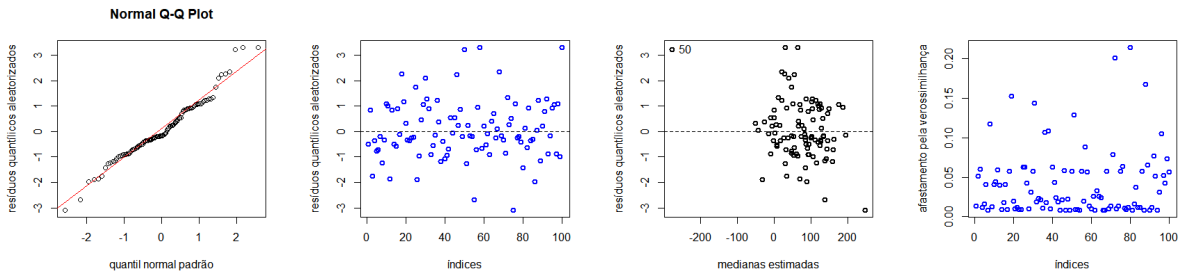
O ajuste deste conjunto de dados é apresentado na Tabela 10.

Tabela 10 – Ajuste do conjunto de dados sob segunda perturbação ao modelo Gucumatz $(\mu_i, \sigma, \lambda, \alpha_1, \alpha_2)$.

$\boldsymbol{\theta}$	Estimativas	IC (95%)
$\beta_0 = 5$	5,663	(4,163; 7,163)
$\beta_1 = -3$	-2,998	(-3,171; -2,824)
$\beta_2 = 3$	2,989	(2,782; 3,196)
$\sigma = 1$	1,111	(0,516; 1,706)
$\lambda = 2,5$	2,257	(2,176; 3,121)

A Tabela 10 mostra que as mudanças nas estimativas, em relação àquelas apresentadas na Tabela 8, foram quase inexistentes, de modo que os intervalos de confiança ainda incluem os verdadeiros valores dos parâmetros $\beta_0, \beta_1, \beta_2$ e σ, e a estimativa do parâmetro λ está contida no respectivo intervalo de confiança.

Para diagnóstico, os seguintes gráficos são apresentados na Figura 12.



(a) (b) (c) (d)

Figura 12 – (a) quantil-quantil normal dos resíduos quantílicos aleatorizados, (b) resíduos quantílicos aleatorizados contra índices das observações, (c) resíduos quantílicos aleatorizados contra $\hat{\boldsymbol{\mu}}$, e (d) afastamento pela verossimilhança.

A Figura 12a indica que o modelo é verdadeiro, devido aparente normalidade dos resíduos, apesar de que há maior inclinação dos pontos do que anteriormente. A Figura 12b apresenta dispersão aleatória dos pontos ao redor da origem, a Figura 12c apresenta dispersão aleatória dos pontos ao redor da origem, e um destaque do resíduo referente à observação 50. Por sua

vez, a Figura 12d indica que não há observações influentes, visto que nenhum ponto é superior a $\chi^2_{5;0,95} = 11,07$.

Após se efetuar terceira e quarta perturbações nestes dados, como anteriormente, e ajustá-los ao modelo Gucumatz, as estimativas e intervalos de confiança ficaram inalterados até o nível centesimal e os diagnósticos dos modelos ajustados pouco se alteraram. Estes resultados foram omitidos para evitar repetições.

Efetua-se uma quinta perturbação neste conjunto de dados, como anteriormente, de modo que $y_{100} = y_{100} + 3 \cdot sd(\boldsymbol{y})$, $x_{50;1} = x_{50;1} + 3 \cdot sd(\boldsymbol{x}_1)$ e $x_{75;2} = x_{75;2} + 3 \cdot sd(\boldsymbol{x}_2)$. A descrição deste conjunto de dados está na Tabela 11.

Tabela 11 – Descrição do conjunto de dados simulado após quinta perturbação.

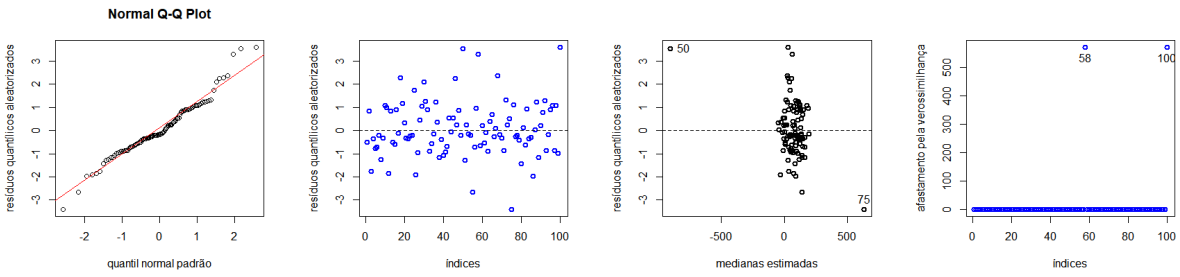
	Mín	1º quartil	Mediana	Média	3º quartil	Máx
$\boldsymbol{y}$	-50,41	36,66	82,51	90,84	122,94	1249,40
$\boldsymbol{x}_1$	1,11	12,00	24,47	28,08	40,16	348,17
$\boldsymbol{x}_2$	26,45	41,13	49,47	50,15	54,94	233,61

O resultado do ajuste dos dados ao modelo Gucumatz está registrado na Tabela 12.

Tabela 12 – Ajuste da amostra sob quinta perturbação ao modelo de regressão Gucumatz $(\mu_i, \sigma, \lambda, \alpha_1, \alpha_2)$.

$\boldsymbol{\theta}$	Estimativas	IC (95%)
$\beta_0 = 5$	5,663	(4,163; 7,163)
$\beta_1 = -3$	-2,998	(-3,171; -2,824)
$\beta_2 = 3$	2,989	(2,782; 3,196)
$\sigma = 1$	1,111	(0,517; 1,706)
$\lambda = 2,5$	2,257	(2,176; 3,121)

A Tabela 12 mostra que as mudanças, em relação àquelas apresentadas na Tabela 10, foram irrelevantes. Para diagnóstico, os seguintes gráficos são apresentados na Figura 13.



(a) (b) (c) (d)

Figura 13 – (a) quantil-quantil normal dos resíduos quantílicos aleatorizados, (b) resíduos quantílicos aleatorizados contra índices das observações, (c) resíduos quantílicos aleatorizados contra valores ajustados $\hat{\boldsymbol{\mu}}$, e (d) afastamento pela verossimilhança.

A Figura 13a, associada ao gráfico quantil-quantil normal dos resíduos, mostra inclinação um pouco maior dos pontos, e um teste shapiro-wilk aplicado aos resíduos apresentou p-valor $0,013$, confirmando que este modelo deixou de ser verdadeiro; por sua vez, a Figura 13d indica que as observações 58 e 100 são altamente influentes sobre o ajuste geral do modelo, com *LDi* acima de 500. Portanto, o diagnóstico deste modelo apresenta mudanças importantes, em relação aos anteriormente apresentados.

Efetua-se uma sexta perturbação neste conjunto de dados, como anteriormente, de modo que $y_{100} = y_{100} + 3 \cdot sd(\boldsymbol{y})$, $x_{50;1} = x_{50;1} + 3 \cdot sd(\boldsymbol{x}_1)$ e $x_{75;2} = x_{75;2} + 3 \cdot sd(\boldsymbol{x}_2)$. A descrição deste conjunto de dados está na Tabela 13.

Tabela 13 – Descrição do conjunto de dados simulado sob sexta perturbação.

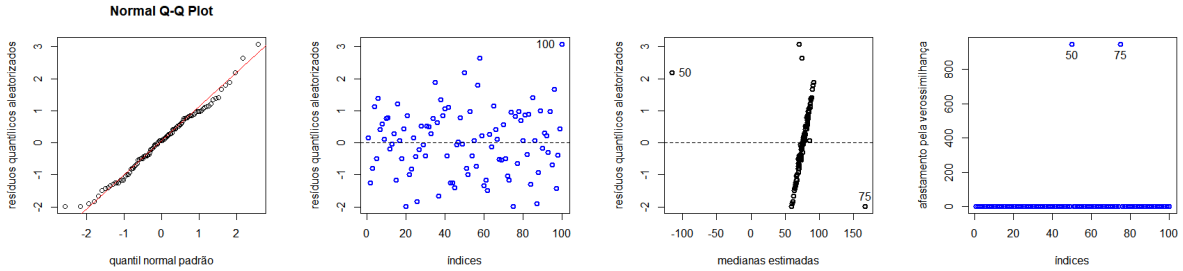
	Mín	1º quartil	Mediana	Média	3º quartil	Máx
$\boldsymbol{y}$	-50,41	36,66	82,51	94,88	122,94	1653,10
$\boldsymbol{x}_1$	1,11	12,00	24,47	29,15	40,16	455,19
$\boldsymbol{x}_2$	26,45	41,13	49,47	50,79	54,94	297,43

O ajuste destes dados ao modelo Gucumatz está registrado na Tabela 14.

Tabela 14 – Ajuste da amostra sob sexta perturbação ao modelo de regressão Gucumatz $(\mu_i, \sigma, \lambda, \alpha_1, \alpha_2)$.

$\boldsymbol{\theta}$	Estimativas	IC (95%)
$\beta_0 = 5$	69,86	(63,26; 76,45)
$\beta_1 = -3$	-0,442	(-1,095; 0,210)
$\beta_2 = 3$	0,364	(-0,484; 1,212)
$\sigma = 1$	50,25	(46,52; 53,98)
$\lambda = 2,5$	3,121	(2,176; 3,121)

Pela Tabela 14, as estimativas e intervalos de confiança dos parâmetros β_0, β_2 e σ sofreram alterações bruscas em relação aos apresentados na Tabela 12. Para diagnóstico, os seguintes gráficos são apresentados na Figura 14.



(a) (b) (c) (d)

Figura 14 – (a) quantil-quantil normal dos resíduos quantílicos aleatorizados, (b) resíduos quantílicos aleatorizados contra índices das observações, (c) resíduos quantílicos aleatorizados contra valores ajustados $\hat{\boldsymbol{\mu}}$, e (d) afastamento pela verossimilhança.

Pelas Figuras 14b e 14c, os resíduos referentes às observações 50 e 75 têm mais destaque do que anteriormente; por sua vez, a Figura 14d indica que as observações 50 e 75 são altamente influentes sobre o ajuste geral do modelo, com $LDi's$ superiores a 800.

Com base nesta simulação, entende-se que a estimação no modelo não é impactada quando os dados sofrem algumas perturbações, até o estado em que pode haver suposto afastamento da normalidade dos resíduos, ou alterações bruscas nas estimativas, ou observações que se tornam altamente influentes no ajuste geral do modelo. Contudo, é necessário testar a utilização de outras técnicas de diagnóstico em modelos Gucumatz.

4.8 Aplicação

Radial Velocity Experiment (RAVE) é uma pesquisa astronômica voltada a reconstruir a história de nossa Galáxia reunindo componentes-chave de movimento e composições químicas de estrelas. O objetivo principal do RAVE é derivar a velocidade radial das estrelas a partir de espectros observados, como temperatura efetiva, gravidade superficial, metalicidade, paralaxe fotométrica e dados de abundância elementar para as estrelas.

Neste trabalho, o conjunto de dados completo da 5ª pesquisa RAVE (RAVE Collaboration, 2020), disponibilizado por Solorzano (2019) com 36 variáveis explicativas, foi reduzido para n=500 primeiras observações, e a quantidade de covariáveis foi reduzida para 15 consideradas mais importantes, com base na descrição da pesquisa. O objetivo neste trabalho é modelar a mediana da velocidade radial (em km/s) em função de 15 medidas espectrais, a fim de exemplificar o uso da regressão Gucumatz.

y_i: *radial velocity* (velocidade radial em km/s);
x_{1i}: *metallicity* (proporção da sua matéria constituída de elementos químicos diferentes do hidrogênio e hélio)
x_{2i}: *spectrophotometric distance* (distância espectométrica);
x_{3i}: *spectrophotometric parallax* from RAVE DR5 (distância espectométrica parallax);
x_{4i}: 2MASS J magnitude; x_{5i}: 2MASS H magnitude; x_{6i}: 2MASS K magnitude;
x_{7i}: *abundance of* Mg; x_{8i}: *abundance of* Si; x_{9i}: *abundance of* Fe;
x_{10i}: *right ascension* (ascensão direta em graus);
x_{11i}: *declination* (declinação em graus);
x_{12i}: APASS DR9 B magnitude; x_{13i}: APASS DR9 V magnitude;
x_{14i}: APASS DR9 RP magnitude; x_{15i}: APASS DR9 IP magnitude;

Radial velocity (velocidade radial) é a velocidade com que o astro se aproxima ou se afasta do ponto observador, que no caso, é o planeta Terra. Quando positiva, o astro se aproxima do ponto observador; quando negativa, o astro se afasta.

O modelo de regressão Gucumatz $(\mu_i, \sigma, \lambda, \alpha_1, \alpha_2)$ relaciona a mediana μ_i da variável resposta velocidade radial com respeito à estrela i, e componente sistemático formado por 15 covariáveis $\boldsymbol{x}_i = (1, \boldsymbol{x}_{1i}, ..., \boldsymbol{x}_{15i})^T$ referentes à matriz de planejamento $\boldsymbol{X}$, e coeficientes $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{15})^T$, por meio da função de ligação identidade. O modelo de regressão é dado por:

$$\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 \boldsymbol{x}_{1i} + ... + \beta_{15}\boldsymbol{x}_{15i}, \quad i = 1, ..., 500. \tag{4.18}$$

Na Figura a seguir são apresentadas a dostribuição de frequência da variável resposta $\boldsymbol{y}$ *radial velocity* e a sua relação descritiva com as covariáveis.

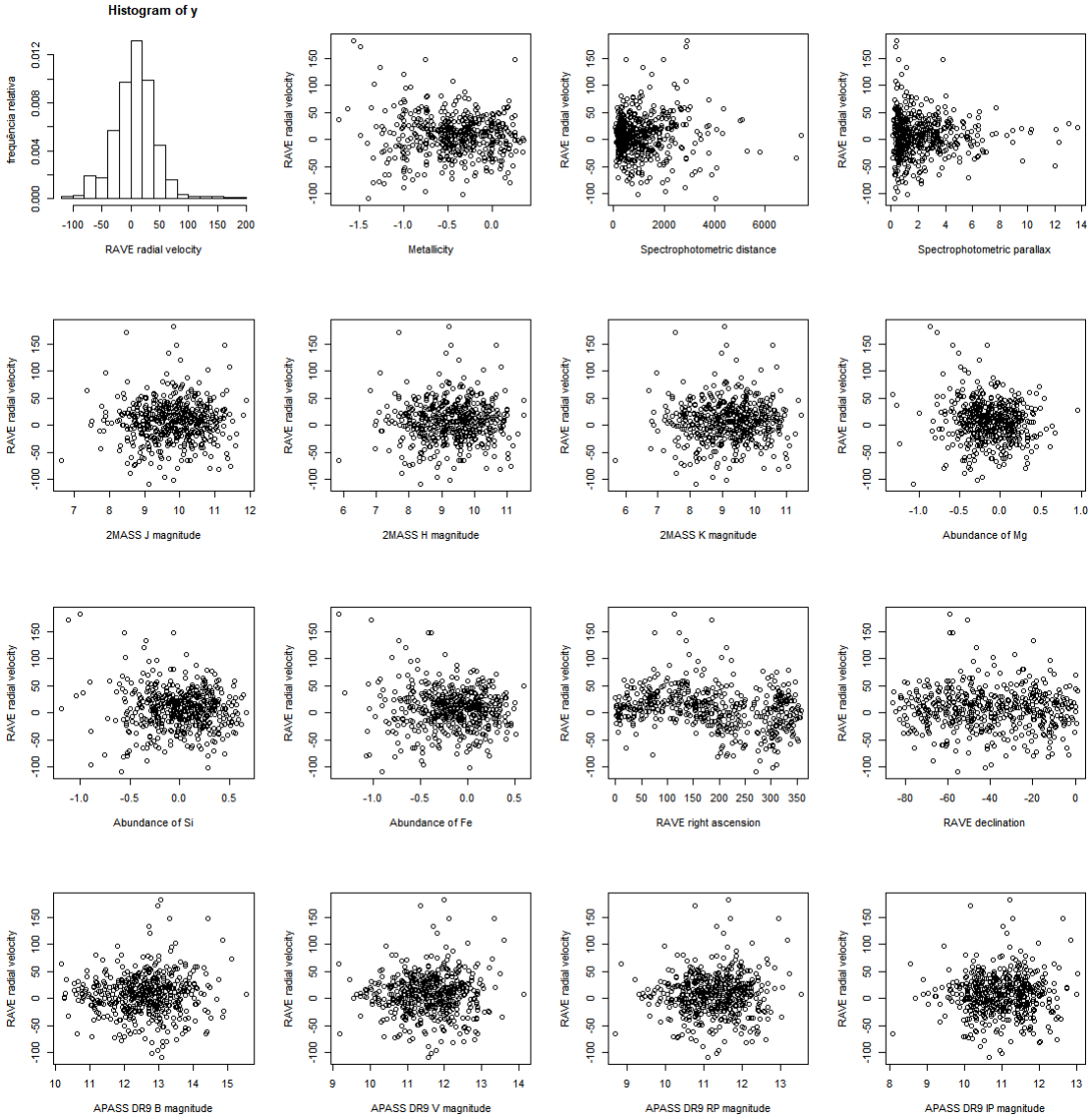


Figura 15 – Relação descritiva da variável resposta *y* (velocidade radial) com as covariáveis.

A figura 15 indica que quase todas as covariáveis tem pouca relação com a variável resposta.

O ajuste dos dados ao modelo de regressão Gucumatz $(\mu_i, \sigma, \lambda, \alpha_1, \alpha_2)$ é apresentado na Tabela 15, no qual somente a covariável *right ascension* é selecionada, com utilização da estatística da razão de verossimilhanças, com significância $p < 0,10$.

Tabela 15 – Ajuste do conjunto de dados RAVE ao modelo Gucumatz.

Covariável	Parâmetro	EMV	Erro padrão
right ascension	β_0	20,56	2,983
	β_1	-0,078	0,014
	σ	0,030	0,093
	λ	3,246	-
$AIC = 5.004,344$	$BIC = 5.014,989$		

Right ascension é uma medida angular horizontal do astro, que, juntamente com a medida angular vertical *declination*, determinam a posição do astro em um sistema de coordenadas celestes.

O ajuste do conjunto de dados RAVE ao modelo de regressão Normal está na Tabela 16.

Tabela 16 – Ajuste do conjunto de dados RAVE ao modelo Normal.

Covariável	Parâmetro	EMV	Erro padrão
right ascension	β_0	23,242	3,277
	β_1	-0,085	0,015
	σ^2	36,52	-
$AIC = 5.020,715$	$BIC = 5.033,359$		

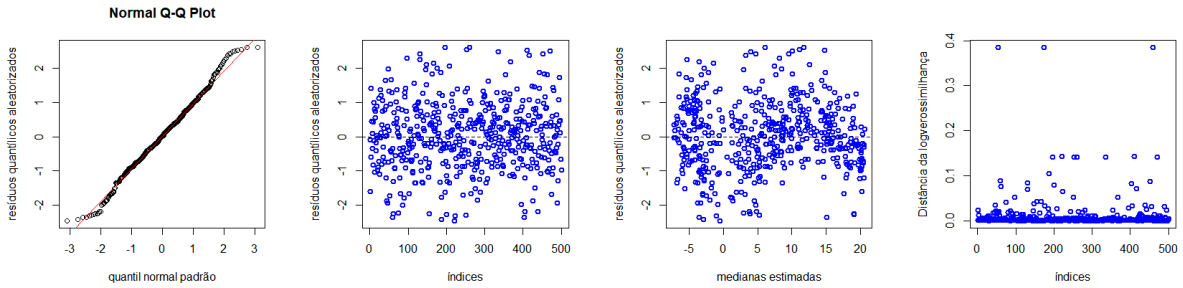
O ajuste do conjunto de dados RAVE ao modelo de regressão Cauchy está na Tabela 17.

Tabela 17 – Ajuste do conjunto de dados RAVE ao modelo Cauchy.

Covariável	Parâmetro	EMV	Erro padrão
right ascension	β_0	17,877	2,492
	β_1	-0,059	0,013
	σ	19,334	1,108
$AIC = 5.103,177$	$BIC = 5.115,821$		

Pelas Tabelas 15, 16 e 17, associadas aos ajustes do conjunto de dados aos modelos Gucumatz, Normal e Cauchy, respectivamente, observa-se que os critérios AIC e BIC do modelo Gucumatz apresentam menor perda de informação.

Para diagnóstico, os seguintes gráficos são apresentados na Figura 16: os gráficos quantil-quantil normal dos resíduos quantílicos aleatorizados (Figura 16a), resíduos quantílicos aleatorizados contra índices das observações (Figura 16b), resíduos quantílicos aleatorizados contra valores ajustados $\hat{\boldsymbol{\mu}}$ (Figura 16c) e afastamento pela verossimilhança das observações (Figura 16d).



(a) (b) (c) (d)

Figura 16 – (a) quantil-quantil normal dos resíduos quantílicos aleatorizados, (b) resíduos quantílicos aleatorizados contra índices das observações, (c) resíduos quantílicos aleatorizados contra $\hat{\boldsymbol{\mu}}$, e (d) afastamento pela verossimilhança.

A Figura 16a, associada ao gráfico quantil-quantil normal dos resíduos quantílicos aleatorizados, indica que o modelo é verdadeiro, devido aparente normalidade dos resíduos. A Figura 16b, associada ao gráfico dos resíduos contra índices das observações, demonstra dispersão aleatória dos pontos em torno da origem. A Figura 16c, associada ao gráfico dos resíduos contra os valores ajustados $\hat{\boldsymbol{\mu}}$, demonstra dispersão aleatória dos pontos em torno da origem. A Figura 16d, associada ao gráfico de afastamentos pela verossimilhança para cada observação excluída, indica que não há observações globalmente influentes, visto que nenhum ponto é superior a $\chi^2_{5;0,95} = 11,07$. O modelo estimado, portanto, é dado por

$$\hat{\mu}_i = 20,56 - 0,078 x_{i(right\ ascension)}, \quad i = 1,...,500 \tag{4.19}$$

Interpreta-se que a cada aumento de um grau na ascenção direta, a mediana da velocidade radial diminui em 0,078 km/s.

Neste capítulo, apresentou-se o modelo de regressão Gucumatz $(\mu_i, \sigma, \lambda, \alpha_1, \alpha_2)$, no qual a mediana μ é modelada por covariáveis, de maneira similar ao que ocorre em modelos lineares generalizados. Realizou-se estimação por máxima verossimilhança de dados completos e apresentou-se intervalos de confiança assintóticos para os parâmetros $\boldsymbol{\beta}$ e σ. Foram realizados testes de hipóteses por meio de estatística da razão de verossimilhança para seleção de covariáveis, bem como seleção de modelos com utilização do Critério de Informação de Akaike (AIC) e Critério de Informação Bayesiano (BIC). Por meio de estudos de simulação, foi observada consistência de todos os estimadores e normalidade assintótica dos estimadores $\hat{\boldsymbol{\beta}}$ e $\hat{\sigma}$, foi observado que o critério AIC identifica melhores ajustes Gucumatz sobre dados desta distribuição e aplicaram-se técnicas para diagnóstico do modelo com utilização de resíduos quantílicos aleatorizados e medida de afastamento pela verossimilhança. Um conjunto de dados reais foi analisado utilizando-se modelo de regressão Gucumatz, o qual apresentou menor perda de informação, em relação aos modelos de regressão Normal e Cauchy sobre os mesmos dados.

CAPÍTULO

5

ESTIMAÇÃO ALTERNATIVA

Uma alternativa de estimação para o vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma, \lambda)$ é considerar λ conhecido. A escolha de λ é determinada em uma situação em que o modelo Gucumatz é identificável, em um *grid* de valores entre 1 e 4. Diferentes modelos Gucumatz são ajustados para cada valor do *grid*, e um critério de seleção de modelos é adotado para determinar o valor de λ mais apropriado. Para a seleção dos modelos, pode-se utilizar o Critério de Informação de Akaike, de Akaike (1974) ou o Critério de Informação Bayesiano, de Schwarz (1978).

Com o valor do parâmetro λ pré-definido, o vetor de parâmetros $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^T, \sigma)$ é estimado por máxima verossimilhança de dados completos, como realizado na Seção 4.2, sendo que a esperança da log-verossimilhança dos dados completos, reescrita a partir de (4.5), a menos das constantes,

$$E_{K|Y,\boldsymbol{\theta}^{*(t)}}[l(\boldsymbol{\theta}^{*(t)}|\boldsymbol{y},\boldsymbol{k})] = l(\boldsymbol{\theta}^{*(t)}|\boldsymbol{y},\boldsymbol{\kappa}^{(t)}) = \sum_{i=1}^{n}\left[\frac{-\kappa_i}{2}\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}^{(t)}}{\sigma^{(t)}}\right)^2\right] - n\log(\sigma^{(t)}) - \sum_{i=1}^{n}(1-\kappa_i)\log\left[1+\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}^{(t)}}{\sigma^{(t)}}\right)^2\right], \quad (5.1)$$

é maximizada apenas em uma etapa, no vetor de parâmetros $\boldsymbol{\theta}^{*(t+1)}$.

Dos elementos do vetor escore $\boldsymbol{U}_{\boldsymbol{\theta}^*}$ dados em (4.6) e (4.7), a matriz de informação observada é dada por

$$\boldsymbol{J}(\boldsymbol{\theta}^*) = -\begin{bmatrix} \boldsymbol{J}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \boldsymbol{J}_{\boldsymbol{\beta}\sigma} \\ & J_{\sigma\sigma} \end{bmatrix}, \quad (5.2)$$

em que $\boldsymbol{J}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ é a matriz bloco diagonal de dimensão $(p+1)$ x $(p+1)$, cujo elemento $J_{\beta_r\beta_s}$ é dado em (4.9), $\boldsymbol{J}_{\boldsymbol{\beta}\sigma}$ é um vetor de dimensão $p+1$, cujo elemento $J_{\beta_r\sigma}$ é dado em (4.10) e o elemento $J_{\sigma\sigma}$ é dado em (4.11). A matriz de informação observada (5.2), mediante pré-determinação do

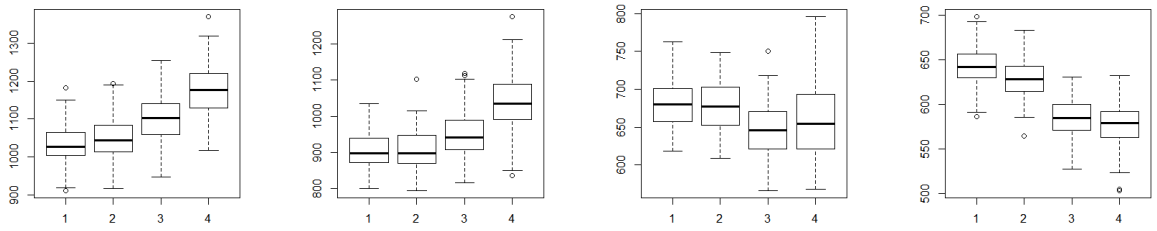
parâmetro λ, apresenta maior facilidade de cálculo em relação àquela, pela ausência do termo $J_{\lambda\lambda}$. De (4.12), se admite, para $\hat{\boldsymbol{\theta}}^* = (\hat{\boldsymbol{\beta}}, \hat{\sigma})$, que

$$\hat{\boldsymbol{\theta}}^* \sim^a N(\boldsymbol{\theta}^*, [\boldsymbol{J}(\hat{\boldsymbol{\theta}}^*)^{-1}]), \tag{5.3}$$

em que $[\boldsymbol{J}(\hat{\boldsymbol{\theta}}^*)^{-1}])$ é a inversa da matriz de informação observada de $\boldsymbol{\theta}^*$, mostrada em (5.2), avaliada em $\hat{\boldsymbol{\theta}}^*$. Assume-se (5.3) devido a distribuição Gucumatz apresentar componente da distribuição Normal, para possibilitar a utilização de técnicas de inferências e diagnósticos mais conhecidas, ainda que não satisfaz todas as condições de regularidade, como discutido na Subseção 3.1.2.

5.1 Estudos de simulação

Com o objetivo de comparar modelos da distribuição Gucumatz com diferentes valores do parâmetro λ pré-determinados, simulam-se 4 grupos de 100 conjuntos de dados do modelo Gucumatz de tamanho $n = 200$, cada grupo mediante $\lambda = 1, \lambda = 2, \lambda = 3, \lambda = 4$, com valores pré-determinados de parâmetros $\beta_0 = 5; \beta_1 = -3; \beta_2 = 3; \sigma = 1$. Cada um dos 400 conjunto de dados foi ajustado à distribuição Gucumatz mediante valores de λ pré-determinados $\lambda = 1, \lambda = 2, \lambda = 3, \lambda = 4$ e os AIC's destes ajustes são apresentados em 4 diagramas de caixa (1,2,3,4), respectivamente.



(a) Simulação com $\lambda = 1$ (b) Simulação com $\lambda = 2$ (c) Simulação com $\lambda = 3$ (d) Simulação com $\lambda = 4$

Figura 17 – Cada figura apresenta 4 diagramas de caixa (1, 2, 3, 4) formados por valores de AIC's de ajustes Gucumatz em que são pré-determinados $\lambda = 1, \lambda = 2, \lambda = 3, \lambda = 4$ respectivamente, sobre amostras Gucumatz simuladas mediante (a) $\lambda = 1$, (b) $\lambda = 2$, (c) $\lambda = 3$ e (d) $\lambda = 4$.

Seleciona-se visualmente os modelos Gucumatz ajustados com $\lambda = 1, \lambda = 2, \lambda = 3, \lambda = 4$ pré-determinados, pela observação dos comportamentos dos diagramas de caixa 1, 2, 3 e 4 formados pelos AIC's dos modelos. A Figura 17b mostra que os diagramas 1 e 2 são quase iguais. Assim, nota-se claramente que o critério AIC pode selecionar modelos com valores de λ incorretos em até uma unidade.

Por isso, avalia-se a consistência e normalidade assintótica dos estimadores de máxima verossimilhança em modelos Gucumatz cujo valor de λ é pré-determinado incorretamente com

diferença em até uma unidade em relação ao verdadeiro. Para isto, são simulados $m = 1.000$ conjuntos de dados com variável resposta que segue distribuição Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$ de tamanhos n, mediante valores reais pré-determinados $\beta_0 = 5, \beta_1 = -3, \beta_2 = 3, \sigma = 1$ e $\lambda = 2$. Duas covariáveis, $X_1 \sim U(1,100)$ e $X_2 \sim N(50,10)$ são consideradas. Neste estudo, são avaliados erro quadrático médio (EQM), variância, vício e probabilidades de cobertura (PC) dos estimadores $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}$. Seja o vetor de parâmetros $\boldsymbol{\theta}^* = (\beta_0, \beta_1, \beta_2, \sigma)$ cujo elemento é θ_j em que $1 \leq j \leq 5$. São avaliados, para $1 \leq j \leq 4$,

$$EQM(\hat{\theta}_j) = \frac{\sum_{l=1}^{m}(\hat{\theta}_{jl} - \theta_j)^2}{m-1}, \quad var(\hat{\theta}_j) = \frac{\sum_{l=1}^{m}(\hat{\theta}_{jl} - \bar{\hat{\theta}}_j)^2}{m-1} \quad \text{e vício}(\hat{\theta}_j) = \frac{\sum_{l=1}^{m}(\hat{\theta}_{jl} - \theta_j)}{m-1}, \tag{5.4}$$

em que $\bar{\hat{\theta}}_j = \dfrac{\sum_{l=1}^{m}\hat{\theta}_{jl}}{m}$, $l = 1, 2, ..., m$.

A probabilidade de cobertura é a proporção de vezes em que o valor real do parâmetro é coberto pelos intervalos de confiança assintóticos (4.13) e (4.14).

Algumas amostras simuladas apresentam observações muito discrepantes, fazendo com que haja números negativos na diagonal principal das inversas das matrizes de informação observada (5.2). Além disso, em alguns casos, para $n = 10$, não ocorre convergência. Nestas situações as amostras são substituídas.

As medidas erro quadrático médio (EQM), variância, vício e probabilidades de cobertura (PC) dos estimadores $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}$ a partir de conjuntos de dados simulados mediante $\lambda = 2$, ajustados ao modelo Gucumatz com valor pré-determinado $\lambda = 1$, são apresentadas na Tabela 18.

Tabela 18 – comportamento assintótico dos estimadores $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ e $\hat{\sigma}$ com $\lambda = 1$ pré-determinado.

n	θ	Média	$EQM(\hat{\theta})$	Var$(\hat{\theta})$	Vício$(\hat{\theta})$	PC
10	β_0	4,778	$1,6 \cdot 10$	$1,6 \cdot 10$	-0,223	0,564
50	β_0	5,063	1,607	1,603	0,063	0,882
100	β_0	4,999	0,631	0,631	-0,001	0,912
200	β_0	4,976	0,217	0,216	-0,024	0,920
10	β_1	-2,998	0,007	0,007	0,002	0,508
50	β_1	-3,002	$2,2 \cdot 10^{-4}$	$2,2 \cdot 10^{-4}$	-0,002	0,886
100	β_1	-3,000	$9,5 \cdot 10^{-5}$	$9,5 \cdot 10^{-5}$	$3,0 \cdot 10^{-5}$	0,900
200	β_1	-3,000	$4,8 \cdot 10^{-5}$	$4,8 \cdot 10^{-5}$	$1,6 \cdot 10^{-4}$	0,924
10	β_2	3,005	0,006	0,005	0,005	0,562
50	β_2	2,999	$5,8 \cdot 10^{-4}$	$5,8 \cdot 10^{-4}$	$-5,3 \cdot 10^{-4}$	0,884
100	β_2	3,000	$2,6 \cdot 10^{-4}$	$2,6 \cdot 10^{-4}$	$5,4 \cdot 10^{-5}$	0,904
200	β_2	3,001	$8,2 \cdot 10^{-5}$	$8,2 \cdot 10^{-5}$	$5,1 \cdot 10^{-4}$	0,922
10	σ	0,438	0,406	0,089	-0,563	0,328
50	σ	0,725	0,094	0,019	-0,275	0,462
100	σ	0,740	0,076	0,008	-0,261	0,246
200	σ	0,750	0,067	0,004	-0,251	0,080

Na Tabela 18, observa-se que as medidas EQM, variância, vício dos estimadores $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$ decrescem com o aumento de n e a medida PC cresce, o que indica que estes estimadores são consistentes e assintoticamente normais. Portanto, pode-se realizar inferências, com base em normalidade assintótica, sobre estes estimadores. Por outro lado, o estimador $\hat{\sigma}$ é consistente, mas a medida PC diminui com o aumento de n, o que indica que este estimador não tem normalidade assintótica. Contudo, considera-se que (5.3) é válido, para possibilitar o emprego de técnicas de diagnósticos mais conhecidas.

A opção de se pré-determinar o parâmetro λ tem como vantagens a diminuição dos passos da estimação por máxima verossimilhança e maior facilidade na realização de inferências sobre os estimadores dos coeficientes, ainda que o estimador $\hat{\sigma}$ possa não ser assintoticamente normal. Contudo, pode-se testar outro critério de informação que possibilite selecionar o parâmetro λ mais próximo ao verdadeiro. Além disso, o espaço paramétrico relativamente pequeno $1 \leq \lambda \leq 4$ facilita a seleção de modelos para escolha do parâmetro λ.

CAPÍTULO

6

CONCLUSÃO

Há dados cujo comportamento indica a presença de duas distribuições de probabilidade que não se misturam, mas se combinam, de forma articulada. Para modelar dados com tais características, este trabalho propôs uma possível nova linha de pesquisa sobre distribuições combinadas, e foi desenvolvida, de forma ampliada, uma distribuição combinada Gucumatz $(\mu, \sigma, \lambda, \alpha_1, \alpha_2)$ para ajustar dados que apresentam comportamento de distribuição Normal no centro e Cauchy nas caudas, e incentivar pesquisas sobre outras distribuições combinadas.

No desenvolvimento da distribuição Gucumatz, foram definidos os parâmetros, funções densidade de probabilidade e função de distribuição acumulada; verificou-se que esta distribuição apresenta média e variância indefinidas, como ocorre com a distribuição de Cauchy, constatou-se que a mediana da distribuição é o parâmetro μ, enquanto que σ é o parâmetro de dispersão quantilica que possibilita uma ordenação, em variabilidade, entre distribuições Gucumatz; realizou-se estimação por máxima verossimilhança de dados completos e apresentou-se um possível caminho para estimação bayesiana.

Apresentou-se modelo de regressão Gucumatz $(\mu_i, \sigma, \lambda, \alpha_1, \alpha_2)$, no qual a mediana μ é modelada por covariáveis, em uma estrutura similar a que ocorre em modelos lineares generalizados. Por meio de estudos de simulação, constatou-se que todos os estimadores são consistentes, e que os estimadores $\hat{\boldsymbol{\beta}}$ e $\hat{\sigma}$ tem distribuição assintoticamente normal; admitiu-se que $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^T, \hat{\sigma}, \hat{\lambda})$ apresente normalidade assintótica, para possibilitar realização de inferências e emprego de técnicas de diagnósticos mais conhecidas, e utilizou-se intervalo de confiança bootstrap para o parâmetro λ. Realizou-se seleção de covariáveis com auxílio da estatística da razão de verossimilhanças, bem como seleção de modelos com utilização de critérios AIC e BIC; diagnósticos de modelos foram realizados por meio de técnicas que envolvem resíduos quantílicos aleatorizados e afastamento pela verossimilhança. Um conjunto de dados reais foi analisado utilizando-se modelos de regressão Gucumatz, Normal e Cauchy, e, constatou-se que o modelo Gucumatz apresentou menor perda de informação. Por fim, foi apresentado um método alternativo de estimação do vetor de parâmetros em que λ é escolhido a partir de um grid de

valores entre 1 e 4, mediante critérios de seleção de modelos.

Portanto, neste trabalho, estendemos a distribuição Gucumatz. O modelo Gucumatz pode ser útil para analisar conjuntos de dados que, de certa forma, se encontram entre as distribuições Normal e Cauchy.

Propõe-se pesquisas sobre outras distribuições combinadas, que sejam úteis para modelar dados que apresentam comportamento similar ao que ocorre na distribuição Gucumatz.

REFERÊNCIAS

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control vol. 19, nº6**, p. 716–723, 1974. Citado nas páginas 64 e 79.

BOLFARINE, H.; SANDOVAL, M. C. **Introdução à inferência estatística**. Rio de Janeiro: SBM, 2001. Citado na página 52.

BRENT, R. **Algorithms for Minimization Without Derivatives**. USA: Prentice Hall, 1973. Citado na página 50.

BROYDEN, C. G. The convergence of a class of double-rank minimization algorithms 1. general considerations. **IMA Journal of Applied Mathematics, Volume 6, Issue 1**, p. 76–90, 1970. Citado na página 50.

CASELLA, G.; BERGER, R. **Inferência estatística**. São Paulo: Cengage Learning, 2016. Citado nas páginas 28 e 52.

CHATTERJEE, S.; HADI, A. S. **Sensitivity Analysis in linear regression**. Canadá: Wiley, 1988. Citado na página 65.

COOK, R. D.; PENA, D.; WEISBERG, S. The likelihood displacement: a unifying principle for influence measures. **Communications in Statistics-Theory and Methods, Taylor Francis, v. 17**, p. 623–640, 1988. Citado na página 65.

COOK, R. D.; WEISBERG, S. **Residuals and Influence in Regression**. New York and London: Chapman and Hall, 1982. Citado na página 66.

CORDEIRO, G. M.; DEMETRIO, C. G. B. **Modelos Lineares Generalizados e Extensões**. Piracicaba-SP: [s.n.], 2008. Citado na página 64.

DEMETRIO, C. G. B.; ZOCCHI, S. S. **Modelos de regressão**. São Paulo: [s.n.], 2011. Citado na página 65.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. **Journal of the Royal Statistical Society**, v. 39, n. 1, p. 1–38, 1977. Citado nas páginas 49 e 50.

DOBSON, A. J. **An Introduction to Generalized Linear Models**. USA: Chapman Hall, 2002. Citado na página 63.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and Graphical Statistics, Taylor Francis, v. 5, n. 3**, p. 236–244, 1996. Citado nas páginas 64 e 65.

ERISEN, B. **Wind Turbine Scada Dataset**. 2019. [Online; accessed 16-march-2021]. Disponível em: <https://www.kaggle.com/berkerisen/wind-turbine-scada-dataset>. Citado na página 23.

FLETCHER, R. A new approach to variable metric algorithms. **The Computer Journal, Volume 13, Issue 3**, p. 317–322, 1970. Citado na página 50.

GEMAN, S.; GEMAN, D. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. **IEEE Transactions on Patern Analysis and Machine Intelligence, 6**, p. 721–741, 1987. Citado na página 55.

GOLDFARB, D. A family of variable-metric methods derived by variational means. **Mathematics of Computation 24**, p. 23–26, 1970. Citado na página 50.

HASTINGS, W. K. Monte carlo sampling methods using markov chains and their applications. **Biometrika, 57**, p. 97–109, 1953. Citado na página 55.

KUBO, M.; NAKAHARA, H.; HISHIKAWA, H.; HISHIKAWA, H. A curve-fitting method of combined distribution in probabilistic modeling of random variables. **First International Conference of Computational Stochastic Mechanics**, p. 47–58, 1991. Citado na página 24.

LEITHOLD, L. **O cálculo com geometria analítica, v.1, 3ª ed.** São Paulo: Harbra, 1994. Citado na página 52.

MITNIK, P. A.; BAEK, S. The kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation. **Springer nº54**, p. 177–192, 2012. Citado na página 43.

PAULA, G. A. **Modelos de regressão com apoio computacional**. São Paulo: [s.n.], 2013. Citado na página 65.

RAVE Collaboration. **RAVE**. 2020. [Online; accessed 16-December-2020]. Disponível em: <https://www.rave-survey.org/>. Citado na página 75.

ROGERS, W.; TUKEY, J. Understanding some long-tailed symmetrical distributions. **Statistica Neerlandica**, v. 26, p. 211–226, 1972. Citado nas páginas 24, 25 e 26.

ROSS, S. **A First Course in Probability**. USA: Pearson, 8ª edição, 2010. Citado nas páginas 15 e 27.

RStudio Team. **RStudio: Integrated Development Environment for R**. Boston, MA, 2018. Disponível em: <http://www.rstudio.com/>. Citado nas páginas 44, 50, 55 e 61.

SADEGHPOUR, A. **Empirical Investigation of Randomized Quantile Residuals for Diagnosis of Non-Normal Regression Models**. Tese (Doutorado) — University of Saskatchewan, 2016. Citado na página 65.

SCHWARZ, G. Estimating the dimension of a model. **The annals of statistics v.6 nº2**, p. 461–464, 1978. Citado nas páginas 64 e 79.

SHAKED, M.; SHANTHIKUMAR, J. G. **Stochastic orders and their applications**. San Diego, SA: Academic Press, 1994. Citado na página 43.

SHANNO, D. F. Conditioning of quasi-newton methods for function minimization. **Mathematics of Computation 24**, p. 647–656, 1970. Citado na página 50.

SOLORZANO, J. **Stars from Gaia DR2 and RAVE DR5**. 2019. [Online; accessed 16-December-2020]. Disponível em: <https://www.kaggle.com/solorzano/rave-dr5-gaia-dr2-consolidated>. Citado na página 75.

TOWNSEND, J.; COLONIUS, H. Variability of the max and min statistic: A theory of the quantile spread as a function of sample size. **PSYCHOMETRIKA VOL. 70, Nº 4**, p. 759–772, 2005. Citado nas páginas 42 e 43.

WASSERMAN, L. **All of Statistics: A Concise course in statistical inference**. New York: Springer-Verlag, 2004. Citado na página 63.

CAPÍTULO

A

ALGORITMOS

Algoritmo 4 – Algoritmo EM

1) Tome uma amostra $\boldsymbol{y}$ de tamanho n;
2) Tentativas iniciais dos parâmetros:
$\mu^{(0)}$ = mediana de $\boldsymbol{y}$;
$\sigma^{(0)}$ =|84º quantil amostral - $\mu^{(0)}$|;
$\lambda^{(0)} \sim U(2,3)$;
$\kappa_i^{(0)} = 1 \quad se \quad \mu^{(0)} - \lambda^{(0)}\sigma^{(0)} \leq y_i \leq \mu^{(0)} + \lambda^{(0)}\sigma^{(0)}$ e $\kappa_i^{(0)} = 0$ caso contrário;

Na iteração $t+1$:
3) Escreva $E_{K|Y,\boldsymbol{\theta}^{(t)}}[l(\boldsymbol{\theta}^{(t)}|\boldsymbol{y},\boldsymbol{k})]$;
4) Maximize $E_{K|Y,\boldsymbol{\theta}^{(t)}}[l(\boldsymbol{\theta}^{(t)}|\boldsymbol{y},\boldsymbol{k})]$ em $\sigma^{(t+1)}, \mu^{(t+1)}, \lambda^{(t+1)}$;
5) $t = t+1$

Enquanto $E_{K|Y,\boldsymbol{\theta}^{(t+1)}}[l(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{y},\boldsymbol{k})] - E_{K|Y,\boldsymbol{\theta}^{(t)}}[l(\boldsymbol{\theta}^{(t)}|\boldsymbol{y},\boldsymbol{k})] \geq 1e-6$, faça os passos 3, 4 e 5
Retorne $\hat{\mu}, \hat{\sigma}, \hat{\lambda}$
Calcule $\hat{\alpha}_1$ e $\hat{\alpha}_2$

Algoritmo 5 – Algoritmo EM Regressão

1) Tome uma amostra $\boldsymbol{y}$ de tamanho n;
2) As tentativas iniciais dos parâmetros são obtidas de ajuste dos dados à regressão linear Normal:
$\boldsymbol{\beta}^{(0)}$ = coeficientes estimados de ajuste da regressão Normal;
$\sigma^{(0)}$ = mín (500, raiz quadrada do Quadrado Médio dos Resíduos);
$\lambda^{(0)} \sim U(2,3)$;
$\kappa_i^{(0)} = 1 \quad se \quad \boldsymbol{x}^T\boldsymbol{\beta}^{(0)} - \lambda^{(0)}\sigma^{(0)} \leq y_i \leq \boldsymbol{x}^T\boldsymbol{\beta}^{(0)} + \lambda^{(0)}\sigma^{(0)}$ e $\kappa_i^{(0)} = 0$ caso contrário;

3) Obtenha a distribuição de $K|Y, \boldsymbol{\theta}^{(t)}$, cuja probabilidade de sucesso é igual a $\kappa^{(t)}$;
4) Calcule $E_{K|Y,\boldsymbol{\theta}^{(t)}}[l(\boldsymbol{\theta}^{(t)}|\boldsymbol{y},\boldsymbol{k})]$;
5) Maximize $E_{K|Y,\boldsymbol{\theta}^{(t)}}[l(\boldsymbol{\theta}^{(t)}|\boldsymbol{y},\boldsymbol{k})]$ em $1/\sigma^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \lambda^{(t+1)}$;
6) $t = t+1$

Enquanto $E_{K|Y,\boldsymbol{\theta}^{(t+1)}}[l(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{y},\boldsymbol{k})] - E_{K|Y,\boldsymbol{\theta}^{(t)}}[l(\boldsymbol{\theta}^{(t)}|\boldsymbol{y},\boldsymbol{k})] \geq 1e-6$, faça os passos 4, 5 e 6
Retorne $\hat{\boldsymbol{\beta}}, \hat{\sigma}, \hat{\lambda}$;
Calcule $\hat{\alpha}_1$ e $\hat{\alpha}_2$

CAPÍTULO

B

PROGRAMAÇÕES EM R

```
######################## FUNÇÃO DENSIDADE DE PROBABILIDADE GUCUMATZ (mi,sig,tau,alp1,alp2)


fdp.GUCUMATZ <- function(x){
  ifelse(x > -tau & x < tau, alp1*dnorm(x,mi,sig), alp2*dcauchy(x,mi,sig))
}


########################  FUNÇÃO DE DISTRIBUIÇÃO ACUMULADA GUCUMATZ (mi,sig,tau,alp1,alp2)


fda.GUCUMATZ <- function(x){
  probs <- rep(NA, length(x))
  for (i in 1:length(x)) {
    if(x[i] < -tau){
      probs[i] <- alp2*(atan(x[i])/pi + 1/2)
    }
    if( x[i]>=-tau && x[i] <tau){
      probs[i] <-  alp2*(atan(-tau)/pi + 1/2) + alp1*(pnorm(x[i],mi,sig)-pnorm(-tau,mi,sig))
    }
    if(x[i]>=tau){
      probs[i] <- alp2*(atan(x[i])/pi + 1/2 + atan(-tau)/pi - atan(tau)/pi) +

      alp1*(pnorm(tau,mi,sig)-pnorm(-tau,mi,sig))
    }
  }
  return(probs)
}
```

```
###################### FUNÇÃO QUANTIL GUCUMATZ (mi,sig,tau,alp1,alp2)

QUANTIL <- function(v){
  Q <- rep(NA, length(v))
  for (i in 1:length(v)) {
    if(v[i] <  alp2*(1/pi*atan(-tau)+0.5)){
      Q[i] <- sig*tan(pi*((1/alp2)*v[i]-0.5))+mi[i]
    }
    if(v[i]>= alp2*(1/pi*atan(-tau)+0.5) && v[i] < alp2*(1/pi*atan(-tau)+0.5)
    +alp1*(pnorm(tau)-pnorm(-tau))){
      Q[i] <-  sig*qnorm((1/alp1)*v[i]-alp2/alp1*(1/pi*atan(-tau)+0.5)+pnorm(-tau))+mi[i]
    }
    if(v[i]> alp2*(1/pi*atan(-tau)+0.5)+alp1*(pnorm(tau)-pnorm(-tau))){
      Q[i] <- sig*tan((pi/alp2)*v[i]-atan(-tau)+pi/2-pi*alp1/alp2*(pnorm(tau)-pnorm(-tau))
      +atan(tau))+mi[i]
    }
  }
  return(Q)
}



###################### ESTIMAÇÃO FREQUENTISTA
REGRESSÃO GUCUMATZ (mi,sigma,tau,alpha1,alpha2), mi=beta0+beta1*x1+beta2*x2

###tentativas iniciais: mi0,sigma0,tau0, em que mi0=betain0+betain1*x1+betain2*x2
fit=lm(y~x1+x2) #regressão normal para obter tentativas iniciais
k=numeric()
k0=numeric()
tau=numeric()
betain0=fit$coefficients[1]
betain1=fit$coefficients[2]
betain2=fit$coefficients[3]
beta0=numeric()
beta1=numeric()
beta2=numeric()
mi0=numeric()
mi=numeric()
sigma=numeric()
tau0=runif(1,2,3);tau0

mi0=betain0+x1*betain1+x2*betain2
sigma0<-min(500,sqrt(sum((fit$residuals)^2)/fit$df.residual))

for (i in 1:n) {

  if(y[i]>(mi0[i]-tau0*sigma0) && y[i]<(mi0[i]+tau0*sigma0)){
```

```
    k0[i]<-1
  } else {
    k0[i]<-0
  }

}

Elogv<-0   #Esperança da logverossimilhança na distribuição de k|mi0,sigma0
Elogv[1]<-0

j<-2

alpha1=1/(exp(-1/2*tau0^2)*(1+tau0^2)/sqrt(2*pi)*(pi-2*atan(tau0))+(2*pnorm(tau0)-1));
rho=2*pnorm(tau0)-1;

Elogv[j]<- sum(-k0/2*((y-(mi0))/sigma0)^2)-sum(k0*log(sigma0*sqrt(2*pi)))
-sum((1-k0)*log(sigma0*pi))-sum((1-k0)*log(1+((y-(mi0))/sigma0)^2))+sum(k0)*2*log(alpha1)
+sum(k0)*log(rho)+sum(1-k0)*2*log(1-alpha1*rho)-sum(1-k0)*log(1-2/pi*atan(tau0))

### ALGORITMO EM

while(abs(Elogv[j]-Elogv[j-1])>=0.0001){

  L<-function(par){
    sigma=exp(par[1])
    beta0=par[2]
    beta1=par[3]
    beta2=par[4]

    mi=beta0+x1*beta1+x2*beta2

    EL<-sum(-k0/2*((y-(mi))/sigma)^2)-sum(k0*log(sigma*sqrt(2*pi)))-sum((1-k0)*log(sigma*pi))
    -sum((1-k0)*log(1+((y-(mi))/sigma)^2))

    -EL

  }

  mod=optim(par = c(sigma0,betain0,betain1,betain2), fn = L, method = "BFGS")

  Ltau<-function(par){
    tau=(par[1])


    alpha1=1/(exp(-1/2*tau^2)*(1+tau^2)/sqrt(2*pi)*(pi-2*atan(tau))+(2*pnorm(tau)-1))
    rho=2*pnorm(tau)-1
```

```
    -1*(sum(k0)*2*log(alpha1)+sum(k0)*log(rho)+sum(1-k0)*2*log(1-alpha1*rho)
    -sum(1-k0)*log(1-2/pi*atan(tau)))

  }

  modtau<-optim(par = c(tau0), fn=Ltau, method = "Brent", lower = 0, upper = 4)

  tau<-modtau$par[1]
  sigmaexp<-mod$par[1]
  sigma<-exp(sigmaexp)
  beta0<-mod$par[2]
  beta1<-mod$par[3]
  beta2<-mod$par[4]


  v<-c(beta0,beta1,beta2,sigma,tau) #estimativas
  print(v)
  mi=beta0+x1*beta1+x2*beta2

  for (i in 1:n) {

    if(y[i]>(mi[i]-tau*sigma) && y[i]<(mi[i]+tau*sigma)){

      k[i]<-1
    } else {
      k[i]<-0
    }

  }

  alpha1=1/(exp(-1/2*tau^2)*(1+tau^2)/sqrt(2*pi)*(pi-2*atan(tau))+(2*pnorm(tau)-1))
  rho=2*pnorm(tau)-1

  Elogv[j+1]<- sum(-k/2*((y-(mi))/sigma)^2)-sum(k*log(sigma*sqrt(2*pi)))
  -sum((1-k)*log(sigma*pi))-sum((1-k)*log(1+((y-(mi))/sigma)^2))
  +sum(k)*2*log(alpha1)+sum(k)*log(rho)+sum(1-k)*2*log(1-alpha1*rho)-sum(1-k)*log(1-2/pi*atan(tau))

  j<-j+1
  k0<-k

}

alpha2=(1-alpha1*rho)/(1-2/pi*atan(taus));

#log-verossimilhança avaliada nas estimativas
DLOGV=sum(log(k*alpha1*exp(-1/2*((y-mi)/sigma)^2)/(sigma*sqrt(2*pi))
```

```
+(1-k)*(alpha2/(sigma*pi*(1+((y-mi)/sigma)^2))))) #log-verossimilhança


####################### ESTIMAÇÃO FREQUENTISTA REGRESSÃO GUCUMATZ (mi,sigma,tau,alpha1,alpha2),
mi=beta0+beta1*x1+beta2*x2 COM TAU PRÉ-DETERMINADO

###tentativas iniciais: mi0,sigma0, em que mi0=betain0+betain1*x1+betain2*x2

#FUNÇÃO ALPHA1
falpha1<-function(tau){
  1/(exp(-tau^2/2)*pi*(1+tau^2)/sqrt(2*pi)*(1+1/pi*(atan(-tau)-atan(tau)))+
  pnorm(tau)-pnorm(-tau))
}

#FUNÇÃO ALPHA2
falpha2<-function(tau){
  (1/(exp(-tau^2/2)*pi*(1+tau^2)/sqrt(2*pi)*(1+1/pi*(atan(-tau)-atan(tau)))+
  pnorm(tau)-pnorm(-tau)))/sqrt(2*pi)*exp(-(tau^2)/2)*pi*(1+tau^2)
}

tau=1 #pré-determinar um valor entre 1 e 4

alpha1=falpha1(tau)
alpha2=falpha2(tau)

k=numeric()
k0=numeric()
beta0=numeric()
beta1=numeric()
beta2=numeric()
mi0=numeric()
mi=numeric()
sigma=numeric()

  fit=lm(y~x1+x2) #regressão normal para obter tentativas iniciais
  betain0=fit$coefficients[1]
  betain1=fit$coefficients[2]
  betain2=fit$coefficients[3]

  mi0=betain0+x1*betain1+x2*betain2
  sigma0<-min(500,sqrt(sum((fit$residuals)^2)/fit$df.residual))

  for (i in 1:n) {

  if(y[i]>(mi0[i]-tau*sigma0) && y[i]<(mi0[i]+tau*sigma0)){

    k0[i]<-1
```

```
} else {
  k0[i]<-0
}


}

Elogv<-0  #Esperança da logverossimilhança na distribuição de k|mi0,sigma0,tau
Elogv[1]<-0
j<-2
Elogv[j]<-sum(-k0/2*((y-(mi0))/sigma0)^2)-n*log(sigma0)-sum((1-k0)*log(1+((y-(mi0))/sigma0)^2))

while(abs(Elogv[j]-Elogv[j-1])>=0.0001 & j<50){

L<-function(par){
  sigma=exp(par[1])
  beta0=par[2]
  beta1=par[3]
  beta2=par[4]

  mi=beta0+x1*beta1+x2*beta2

  EL<-sum(-k0/2*((y-(mi))/sigma)^2)-n*log(sigma)-sum((1-k0)*log(1+((y-(mi))/sigma)^2))
  -EL

}

mod=optim(par = c(sigma0,betain0,betain1,betain2), fn = L, method = "BFGS")


sigmaexp<-mod$par[1]
sigma<-exp(sigmaexp)
beta0<-mod$par[2]
beta1<-mod$par[3]
beta2<-mod$par[4]


v<-c(beta0,beta1,beta2,sigma)
print(v) #estimativas

mi=beta0+x1*beta1+x2*beta2

for (i in 1:n) {

  if(y[i]>(mi[i]-tau*sigma) && y[i]<(mi[i]+tau*sigma)){

    k[i]<-1
```

```
    } else {
      k[i]<-0
    }

  }

  #Esperança da logverossimilhança na distribuição de k|mi,sigma
  Elogv[j+1]<-sum(-k/2*((y-(mi))/sigma)^2)-n*log(sigma)-sum((1-k)*log(1+((y-(mi))/sigma)^2))
  j<-j+1
  k0<-k

}


################## AFASTAMENTO PELA VEROSSIMILHANÇA -
REGRESSÃO GUCUMATZ (mi,sigma,tau,alpha1,alpha2), mi=beta0+beta1*x1+beta2*x2


dc=numeric()  #afastamento pela verossimilhança aproximado pela distancia de cook
x1s=numeric()  #x1 com retirada da observação s
x2s=numeric()  #x2 com retirada da observação s
ys=numeric()  #y com retirada da observação s
mi0s=numeric()  #mi0 com retirada da observação s
mis=numeric()  #mi com retirada da observação s
k0s=numeric()  #k0 com retirada da observação s
ks=numeric()  #k com retirada da observação s
km=numeric()  #indicadora usada no final
Elogvs=numeric()  #Esperança da logverossimilhança dos dados completos
com retirada da observação s
LD=numeric()  #afastamento pela verossimilhança
DLOGVM=numeric() #logverossimilhança mediante retirada da observação s
beta0s=numeric()  #beta0 com retirada da observação s
beta1s=numeric()  #beta1 com retirada da observação s
beta2s=numeric()  #beta2 com retirada da observação s
sigmas=numeric()  #sigma com retirada da observação s
taus=numeric()  #tau com retirada da observação s

s=1
for (s in 1:n) {
  print(s)
  x1s<-x1[-s]
  x2s<-x2[-s]
  ys<-y[-s]

  #fit=lm(ys~x1s+x2s)
  #betain0=fit$coefficients[1]
  #betain1=fit$coefficients[2]
```

```
#betain2=fit$coefficients[3]

mi0s=betain0+x1s*betain1+x2s*betain2
#sigma0<-min(500,sqrt(sum((fit$residuals)^2)/fit$df.residual))


for (i in 1:(n-1)) {

  if(ys[i]>(mi0s[i]-tau0*sigma0) && ys[i]<(mi0s[i]+tau0*sigma0)){

    k0s[i]<-1
  } else {
    k0s[i]<-0
  }

}

Elogvs<-0
Elogvs[1]<-0
h<-2
alpha1=1/(exp(-1/2*tau0^2)*(1+tau0^2)/sqrt(2*pi)*(pi-2*atan(tau0))+(2*pnorm(tau0)-1));
rho=2*pnorm(tau0)-1;

Elogvs[h]<- sum(-k0s/2*((ys-(mi0s))/sigma0)^2)-sum(k0s*log(sigma0*sqrt(2*pi)))
-sum((1-k0s)*log(sigma0*pi))-sum((1-k0s)*log(1+((ys-(mi0s))/sigma0)^2))+sum(k0s)*2*log(alpha1)
+sum(k0s)*log(rho)+sum(1-k0s)*2*log(1-alpha1*rho)-sum(1-k0s)*log(1-2/pi*atan(tau0))

while(abs(Elogvs[h]-Elogvs[h-1])>=0.0001){

  L<-function(par){
    sigmas=exp(par[1])
    beta0s=par[2]
    beta1s=par[3]
    beta2s=par[4]

    mis=beta0s+x1s*beta1s+x2s*beta2s

    EL<-sum(-k0s/2*((ys-(mis))/sigmas)^2)-(n-1)*log(sigmas)
    -sum((1-k0s)*log(1+((ys-(mis))/sigmas)^2))

    -EL

  }

  mod=optim(par = c(sigma0,betain0,betain1,betain2), fn = L, method = "BFGS")
```

```
Ltau<-function(par){
  taus=(par[1])


  alpha1=1/(exp(-1/2*taus^2)*(1+taus^2)/sqrt(2*pi)*(pi-2*atan(taus))+(2*pnorm(taus)-1))
  rho=2*pnorm(taus)-1

  -1*(sum(k0s)*2*log(alpha1)+sum(k0s)*log(rho)+sum(1-k0s)*2*log(1-alpha1*rho)
  -sum(1-k0s)*log(1-2/pi*atan(taus)))

}

modtau<-optim(par = c(tau0), fn=Ltau, method = "Brent", lower = 0, upper = 4)

taus<-modtau$par[1]
sigmaexp<-mod$par[1]
sigmas<-exp(sigmaexp)
beta0s<-mod$par[2]
beta1s<-mod$par[3]
beta2s<-mod$par[4]

vobs<-c(beta0s,beta1s,beta2s,sigmas,taus)
print(vobs)


#distribuição de k|mi,sigma
mis=beta0s+x1s*beta1s+x2s*beta2s

for (i in 1:(n-1)) {

  if(ys[i]>(mis[i]-taus*sigmas) && ys[i]<(mis[i]+taus*sigmas)){

    ks[i]<-1
  } else {
    ks[i]<-0
  }

}

alpha1=1/(exp(-1/2*taus^2)*(1+taus^2)/sqrt(2*pi)*(pi-2*atan(taus))+(2*pnorm(taus)-1))
rho=2*pnorm(taus)-1

Elogvs[h+1]<- sum(-ks/2*((ys-(mis))/sigmas)^2)-sum(ks*log(sigmas*sqrt(2*pi)))
-sum((1-ks)*log(sigmas*pi))-sum((1-ks)*log(1+((ys-(mis))/sigmas)^2))+sum(ks)*2*log(alpha1)
+sum(ks)*log(rho)+sum(1-ks)*2*log(1-alpha1*rho)
-sum(1-ks)*log(1-2/pi*atan(taus))
h<-h+1
```

```
    k0s<-ks

  }

  thetaobs= matrix(vobs,nrow = 5,ncol = 1); #vetor de estimativas mediante retirada da observação s

  dcook= t(thetaobs-theta)%*%jj%*%(thetaobs-theta);dcook
  dc[s]<-dcook #afastamento pela verossimilhança aproximado pela distância de cook
  mediante retirada da observação s

  mm=beta0s+x1*beta1s+x2*beta2s #mi com observações completas

  for (i in 1:n) {

    if(y[i]>(mm[i]-taus*sigmas) && y[i]<(mm[i]+taus*sigmas)){

      km[i]<-1
    } else {
      km[i]<-0
    }

  }

  alpha2=(1-alpha1*rho)/(1-2/pi*atan(taus));

  #logverossimilhança mediante retirada da observação s
  DLOGVM[s]=sum(log(km*(alpha1*exp(-1/2*((y-mm)/sigmas)^2)/(sigmas*sqrt(2*pi)))+
  (1-km)*(alpha2/(sigmas*pi*(1+((y-mm)/sigmas)^2)))))

  LD[s]=2*(DLOGV-DLOGVM[s]) #afastamento pela verossimilhança da observação i


}


################### SIMULAÇÃO ESTIMAÇÃO BAYESIANA -
DISTRIBUIÇÃO GUCUMATZ (mi,sigma,tau,alpha1,alpha2)

niter=100.000
mi<-numeric()
sigma<-numeric()
tau<-numeric()
k<-numeric()
vmi=numeric()
vsigma=numeric()
vtau=numeric()
```

```
s<-seq(500,100.000,by=50)
r<-1


#começam as 500 simulações
m=500
while (r <= m) {

  #gerar a amostra e tentativas iniciais

  print(r)
  v<-runif(n)
  x<-QUANTIL(v)
  y<-c(sort(x))

  mi0<-median(y)
  sigma0<-quantile(y,0.84)-mi0
  tau0<-runif(1,1,4)
  mi[1]<-mi0
  sigma[1]<-sigma0
  tau[1]<-tau0


  #começa o gibbs sampling

  for (i in 2:niter) {

    mik<-mi[i-1]
    tauk<-tau[i-1]
    sigmak<-sigma[i-1]

    #prob condicional do vetor k

    for (j in 1:n) {
      if(y[j]>(mik-tauk*sigmak) && y[j]<(mik+tauk*sigmak)){

        k[j]<-1
      } else {
        k[j]<-0
      }

    }

    #prob condicional de mi

    p1=rnorm(1,0,2)
    pmi=mi[i-1]+p1 #gerando candidatos
```

```
  minum=exp(sum(-k/2*((y-pmi)/sigma[i-1])^2))*prod((1/(1+((y-pmi)/sigma[i-1])^2))^(1-k))
  *exp((-1/2)*((pmi)^2/sigma[i-1]^2))
  miden=exp(sum(-k/2*((y-mi[i-1])/sigma[i-1])^2))*prod((1/(1+((y-mi[i-1])/sigma[i-1])^2))^(1-k))
  *exp((-1/2)*((mi[i-1])^2/sigma[i-1]^2))
  tmi=minum/miden #teste para aceitação de pmi na cadeia
  aceitami = min(1, tmi)
  u = runif(1)
  mi[i] = ifelse(u < aceitami, pmi, mi[i-1])

  #prob condicional de sigma


  p2=rnorm(1,0,2)
  psigma=0.01+abs(p2) #gerando candidatos
  sigmanum=(1/psigma)^sum(k)*exp(-sum(k/2*((y-mi[i])/psigma)^2))*(1/psigma)^sum(1-k)
  *prod((1/(1+((y-mi[i])/psigma)^2))^(1-k))*exp((-1/2)*((mi[i])^2/psigma^2))*(1/(100*psigma)
  *exp(-(log(psigma)^2)/20000))
  sigmaden=(1/sigma[i-1])^sum(k)*exp(-sum(k/2*((y-mi[i])/sigma[i-1])^2))*(1/sigma[i-1])^sum(1-k)
  *prod((1/(1+((y-mi[i])/sigma[i-1])^2))^(1-k))
  *exp((-1/2)*((mi[i])^2/sigma[i-1]^2))*(1/(100*sigma[i-1])*exp(-(log(sigma[i-1])^2)/20000))
  tsigma=sigmanum/sigmaden #teste para aceitação de psigma na cadeia
  acsigma = min(1, tsigma)
  u = runif(1)
  sigma[i] = ifelse(u < acsigma, psigma, sigma[i-1])


  #prob condicional completa de tau


  ptau=runif(1,1,4) #gerando candidatos

  Ltnum = ((pnorm(ptau)-pnorm(-ptau))/(exp(-ptau^2/2)*pi*(1+ptau^2)/sqrt(2*pi)
  *(1+1/pi*(atan(-ptau)-atan(ptau)))+(pnorm(ptau)-pnorm(-ptau))))^sum(k)
  *(1-(pnorm(ptau)-pnorm(-ptau))/(exp(-ptau^2/2)*pi*(1+ptau^2)/sqrt(2*pi)
  *(1+1/pi*(atan(-ptau)-atan(ptau)))+(pnorm(ptau)-pnorm(-ptau))))^sum(1-k)*(1/(100*ptau)
  *exp(-(log(ptau)^2)/20000))
  Ltden = ((pnorm(tau[i-1])-pnorm(-tau[i-1]))/(exp(-tau[i-1]^2/2)*pi*(1+tau[i-1]^2)/sqrt(2*pi)
  *(1+1/pi*(atan(-tau[i-1])-atan(tau[i-1])))+(pnorm(tau[i-1])-pnorm(-tau[i-1]))))^sum(k)
  *(1-(pnorm(tau[i-1])-pnorm(-tau[i-1]))/(exp(-tau[i-1]^2/2)*pi*(1+tau[i-1]^2)/sqrt(2*pi)
  *(1+1/pi*(atan(-tau[i-1])-atan(tau[i-1])))+(pnorm(tau[i-1])-pnorm(-tau[i-1]))))^sum(1-k)
  *(1/(100*tau[i-1])*exp(-(log(tau[i-1])^2)/20000))
  ttau=Ltnum/Ltden #teste para aceitação de ptau na cadeia
  aceitaptau = min(1, ttau)
  u = runif(1)
  tau[i] = ifelse(u < aceitaptau, ptau, tau[i-1])

}
```

```
  mi<-c(mi[s])
  sigma<-c(sigma[s])
  tau<-c(tau[s])

  vmi[r]<-mean(mi); print(vmi[r])
  vsigma[r]<-mean(sigma); print(vsigma[r])
  vtau[r]<-mean(tau); print(vtau[r])

  r<-r+1

}
```


ICMC
USP
SÃO CARLOS