# Modelo de Regressão Simplex Multivariado (Inferência, Diagnóstico, Aplicação)

Lucas Santos Vieira

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade Federal da Bahia
Para
obtenção do título
de
Mestre em Matemática

Área de concentração: Estatística

Orientador: Prof. Dr. Jalmar Manuel Farfan Carrasco

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da FAPESB.

Salvador/BA, fevereiro de 2024

**Modelo de Regressão Simplex Multivariado (Inferência, Diagnóstico, Aplicação)**

Esta é a versão final da dissertação elaborada
por Lucas Santos Vieira e aprovada
pela Comissão Julgadora.

Ficha catalográfica

Vieira, Lucas Santos.

Modelo de Regressão Simplex Multivariado (Inferência, Diagnóstico, Aplicação) / Lucas Santos Vieira. – Salvador/BA, fevereiro de 2024.

100 f. : il

Orientador: Prof. Dr. Jalmar Manuel Farfan Carrasco.

Dissertação (Mestrado - Programa de Pós-graduação em Matemática / Área de concentração Estatística) – Universidade Federal da Bahia, Instituto de Matemática e Estatística, 2024.

1. Modelo de Regressão Simplex Multivariado. 2. Inferência. 3. Diagnóstico. 4. Aplicação.

I. Carrasco, Jalmar Manuel Farfan.

II. Título.

# Modelo de Regressão Simplex Multivariado (Inferência, Diagnóstico, Aplicação)

Lucas Santos Vieira

Dissertação apresentada ao Colegiado do Curso de Pós-graduação em Matemática da Universidade Federal da Bahia, como requisito parcial para obtenção do Título de Mestre em Matemática.

## Banca examinadora

Documento assinado digitalmente
gov.br
JALMAR MANUEL FARFAN CARRASCO
Data: 28/02/2024 10:08:39-0300
Verifique em https://validar.iti.gov.br

---

Prof. Dr. Jalmar Manuel Farfan Carrasco (orientador - UFBA)

Patrícia Leone Espinheira Ospina

---

Profa. Dra. Patricia Leone Espinheira Ospina (UFBA)

Documento assinado digitalmente
gov.br
ALDO WILLIAM MEDINA GARAY
Data: 27/02/2024 21:53:16-0300
Verifique em https://validar.iti.gov.br

---

Prof. Dr. Aldo William Medina Garay (UFPE)

# Agradecimentos

À Deus, primeiramente.

Ao professor Dr. Jalmar Manuel Farfan Carrasco, pela excelente orientação, paciência e confiança dedicada a mim na elaboração desta dissertação.

Aos meus pais Davi e Josefina pelo amor, educação e apoio que sempre me ofereceram.

Ao professor Dr. Anderson Ara e a professora Dra. C. C. Costa pelos comentários e sugestões na fase da apresentação do projeto que culminou na presente dissertação.

A Fundação de Amparo à Pesquisa do Estado da Bahia - FAPESB, pela bolsa de mestrado que foi de fundamental importância durante o andamento dos meus estudos.

Ao Departamento de Estatística (Instituto de Matemática e Estatística - UFBA) e ao seu corpo docente pelo comprometimento com a qualidade e excelência do ensino.

Aos membros da banca examinadora, pelo interesse e disponibilidade.

Aos colegas do curso da pós-graduação pelas trocas de ideias, ajuda mútua e desafios enfrentados com espírito colaborativo.

# Resumo

Vieira, Lucas Santos. **Modelo de Regressão Simplex Multivariado (Inferência, Diagnóstico, Aplicação)**. 2023. 100 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade Federal da Bahia, Salvador/BA, 2024.

Os modelos de regressão Beta e Simplex tem sido amplamente utilizados para analisar variáveis que representam taxas, proporções ou índices, isto é, variáveis mensuráveis no intervalo aberto (0,1). Em alguns fenômenos estas variáveis estão correlacionadas, o que requer a obtenção de distribuição multivariada, em particular o caso bivariado, segundo uma determinada abordagem. Nesse sentido, o presente trabalho tem como objetivo principal propor o modelo de regressão Simplex Multivariado (MRSM) via função cópula. Estimadores para os parâmetros são encontrados via o método de máxima verossimilhança (MV) e, via um estudo de simulação estuda-se seus respectivos comportamentos assintóticos. Uma análise de diagnostico, tais como: análise de resíduos e influência global (distância de Cook generalizada e afastamento da verossimilhança), são desenvolvidos com o intuito de identificar possíveis pontos atípicos e/ou influentes e a adequabilidade do modelo aos dados. Por fim, os resultados são aplicados a dois conjuntos de dados reais para exemplificar a metodologia desenvolvida.
**Palavras-chave:** Modelo de regressão Simplex multivariado, inferência, diagnóstico e aplicações.

# Abstract

Vieira, Lucas Santos. **Multivariate Simplex Regression Model (Inference, Diagnosis, Application)**. 2024. 100 f. Dissertation. Institute of Mathematics and Statistics, Federal University of Bahia, Salvador/BA, 2024.

Beta and Simplex regression models have been widely used to analyze variables that represent rates, proportions or index, that is, variables measurable in the open interval (0.1). In some phenomena these variables are correlated, which requires obtaining a multivariate distribution, in particular the bivariate case, according to a certain approach. In this sense, the main objective of this work is to propose the Multivariate Simplex regression model (MRSM) via the copula function. Estimators for the parameters are found via the maximum likelihood (MV) method and, via a simulation study, their respective asymptotic behaviors are studied. A diagnostic analysis, such as: residual analysis and global influence (generalized Cook's distance and likelihood departure), are developed with the aim of identifying possible atypical and/or influential points and the suitability of the model to the data. Finally, the results are applied to two sets of real data to exemplify the developed methodology.
**Keywords:** Multivariate Simplex regression model, inference, diagnosis and applications.

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Capítulo 1

# Introdução

Conforme elucidado em Liu *et al.* (2020), estudos científicos de diferentes áreas produzem dados contínuos proporcionais que estão relacionados a várias situações práticas, quer sejam experimentais ou observacionais. Assim, existe o interesse de investigar e modelar variáveis que representam taxas, proporções ou índices, isto é, variáveis mensuráveis no intervalo aberto (0,1).

Por muito tempo, os modelos lineares tem sido utilizados para descrever a maioria dos fenômenos aleatórios, entretanto, quando o objetivo é modelar dados restritos ao intervalo (0,1), o modelo linear torna-se inadequado, pois, tais dados, em geral, apresentam assimetria e violam certas suposições básicas, como por exemplo heteroscendasticidade. (Silva, 2015). Por esta razão, diversos autores propuseram modelos considerando distribuições em que a variável resposta é limitada no intervalo aberto (0,1). Dentre os modelos mais estudados para esses tipos de dados temos o modelo de regressão Beta proposto por diversos autores, como por exemplo: Paolino (2001) aplicada a dados sobre ciência política; Vasconcellos e Cribari-Neto (2005) propuseram uma classe de modelos de regressão para modelar dados restritos ao intervalo (0,1); Rocha e Simas (2011) avaliaram a influencia das observações na classe geral dos modelos de regressão Beta introduzidos por Simas *et al.* (2010); Ferrari e Cribari-Neto (2004) propuseram um modelo de regressão cuja resposta corresponde a distribuição Beta por meio de uma reparametrização. Na literatura, alternativamente à distribuição Beta temos a distribuição Simplex, que pertence a uma classe de modelos alternativos, conhecido como a classe dos modelos de dispersão (Jorgensen, 1997).

Diversos autores também vem se dedicando aos modelos de regressão Simplex, tais como: Liu *et al.* (2020) que propuseram um modelo de regressão Simplex inflacionado de zeros-uns, contemplando a parte discreta (0's e 1's), bem como a parte contínua, o intervalo aberto (0,1); Miyashiro (2008) desenvolveu técnica de análise de diagnóstico para modelos de regressão Simplex a partir do trabalho de Espinheira *et al.* (2008); Espinheira e de Oliveira Silva (2019) propuseram técnicas de análise de influência local para a classe geral dos modelos de regressão Simplex; Carrasco e Reid (2021) consideraram o modelo de regressão Simplex na presença de erro de medida na covariável; Silva (2019) propôs uma extensão do modelo de regressão Simplex apresentado por Miyashiro (2008), em que o parâmetro da média e precisão estão relacionados às covariáveis por meio do preditores não lineares.

Na prática, dado a exigência de vários eventos cujos dados estão correlacionados, surge uma necessidade por modelos de regressão multivariados, tendo em vista que existem variáveis a serem modeladas de forma conjunta. Nesta vertente, Cepeda-Cuervo *et al.* (2014) propuseram o modelo de regressão Beta bivariada com modelagem conjunta dos parâmetros de média e dispersão via função cópula Farlie-Gumbel-Morgenstern (FGM); Souza e Moura (2016) propuseram o modelo de regressão Beta multivariado para diferentes funções cópulas, conduzindo o processo de inferência sob uma abordagem Bayesiana; Koochemeshkian *et al.* (2020) propuseram o modelo de regressão Beta bivariado e suas aplicações médicas, baseada em uma distribuição Beta bivariada flexível com três parâmetros de forma, etc. Assim o presente trabalho de dissertação tem em seu maior intuito e inovação estender o modelo de regressão Simplex univariado para um modelo de regressão Simplex multivariado via função cópula. Inferências via máxima verossimilhança são realizadas, intervalos

de confiança e testes hipóteses são desenvolvidos; estudos de simulação são conduzidos; propomos alguns método de diagnostico e utilizamos dois conjuntos de dados reais para validar a metodologia proposta.

## 1.1 Objetivos do trabalho

a - Objetivo geral

- O objetivo geral deste trabalho de pesquisa é propor o modelo de regressão Simplex multivariado (MRSM) via função cópula.

b - Objetivos específicos

- Estudar o modelo de regressão Simplex univariádo:
  - Método de estimação via máxima verossimilhança (MV);
  - Análise de diagnóstico.
- Definir o modelo de regressão Simplex multivariado:
  - Analisar métodos de associação via cópulas, a saber: FGM, Clayton, Frank;
  - Estudar o modelo de regressão Simplex bivariado;
  - Métodos de estimação via máxima verossimilhança (MV);
  - Via simulação de Monte Carlo (MC) estudar o comportamento assintótico dos estimadores de (MV) do modelo proposto;
  - Análise de diagnóstico, tais como: análise de resíduos (resíduos quantílicos), análise de influência global (distância de Cook-generalizada, afastamento da verossimilhança) são obtidos;
  - Avaliar a metodologia proposta mediante uso de conjunto de dados reais.

## 1.2 Suporte computacional

Os resultados alcançados neste trabalho de dissertação dar-se-á por meio da utilização da linguagem de programação R (R Core Team, 2024), em sua versão 4.1.2 para sistema operacional Microsoft Windows que pode ser obtido gratuitamente através do site http://www.R-project.org. Uma das principais vantagens deste software é o fato de ser de livre distribuição e possuir código fonte aberto, bem como, uma vasta concentração de librarias estatísticas para a análise de dados, dos quais pode-se citar: *maxLik* (Henningsen e Toomet, 2011), *maxlogL* (Gutiérrez e Hernández, 2020), *betareg* (Zeileis *et al.*, 2016), *simplexreg* (Zhang *et al.*, 2016), *copula* (Hofert *et al.*, 2014) e *gamlss* (Stasinopoulos e Rigby, 2008). Este trabalho é editado utilizando o sistema de tipografia LaTeX (maiores detalhes sobre o LaTeX podem ser encontrados no site http://www.tex.ac.uk/CTAN/latex).

# Capítulo 2

# Teoria dos Modelos de Dispersão

Neste capítulo é apresentado, brevemente, a classe dos modelos de dispersão, que estende os modelos lineares generalizados (Jorgensen, 1997); a distribuição Simplex (Barndorff-Nielsen e Jørgensen, 1991), suas propriedades e métodos de estimação dos parâmetros.

## 2.1 Modelos de Dispersão

Os modelos de dispersão podem ser visto como uma generalização da distribuição normal. Ou seja, consideremos a densidade de uma distribuição normal univariada com média $\mu \in R$ e variância $\sigma^2 \in R_+$, como

$$f(y;\mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}, y \in R.$$

Seja $d(y;\mu) = (y-\mu)^2$ e $a(y;\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}}$, pode-se expressar a densidade acima na forma

$$f(y;\mu,\sigma^2) = a(y;\sigma^2) \exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\}, y \in S, \tag{2.1}$$

em que $S$ é o suporte da distribuição, E[y] $= \mu \in \Omega$, $\Omega$ é o espaço paramétrico de $\mu$, $\sigma^2 > 0$ e $a(\cdot,\cdot) \geq 0$ é um termo de normalização adequado independente de $\mu$ e $d(y;\mu)$, tal que $(y;\mu) \in R \times \Omega$. A ideia dos modelos de dispersão é substituir o quadrado da distância euclidiana $d(y;\mu)$ em (2.1) por outra função adequada, chamada desvio unitário, que mede a distância de uma observação $y$ de um ponto referencia central $\mu$ da distribuição.

Dito isso, seja $S \in R$ o conjunto dos valores realizáveis das distribuições de probabilidade contidas na família dos modelos de dispersão. Denote o suporte convexo de $S$ (i.e., o menor intervalo contendo $S$) por $C$ e defina $\Omega \subseteq C$. Aqui $\Omega$ será o espaço paramétrico referencial do parâmetro $\mu$. Por fim, um modelo de dispersão DM$(\mu,\sigma^2)$ é uma família de distribuição de probabilidade parametrizada por um parâmetro de localização $\mu$ e um parâmetro de dispersão $\sigma^2$, cujo ponto de partida para sua construção é definir o conceito de desvio unitário (Cordeiro *et al.*, 2021).

**Definição 1** *Jorgensen (1997) Seja $\Omega \subseteq C \subseteq R$ intervalos com $\Omega$ aberto. Uma função $d : C \times \Omega \to R$ é chamada de desvio unitário se satisfaz:*

$$d(y;y) = 0 \quad \forall y \in \Omega \qquad e \qquad d(y;\mu) > 0 \quad \forall y \neq \mu.$$

Um desvio unitário é dito regular quando $d(y;\mu)$ é duas vezes diferenciável com respeito a $(y;\mu)$ sobre $\Omega \times \Omega$ e satisfaz

$$\frac{\partial^2 d(\mu;\mu)}{\partial \mu^2} > 0, \quad \forall \mu \in \Omega.$$

Para cada desvio de unidade regular $d : C \times R \to R_+$ definimos a função de variância unitária $V : \Omega \to R_+$ dada por

$$V(\mu) = \frac{2}{\dfrac{\partial^2 d(\mu;\mu)}{\partial \mu^2}}$$

para cada $\mu \in \Omega$. A função de variância unitária desempenha um papel importante na teoria dos modelos de dispersão, pois expressa a dependência da variância da expectativa em modelos de dispersão e caracteriza de forma única os elementos de algumas classes importantes desses modelos, ver Tabela 2.1.

**Tabela 2.1:** *Funções de desvio unitário e variância de alguns modelos de dispersão.*

| Distribuição | $d(y;\mu)$ | $C$ | $\Omega$ | $V(\mu)$ |
|---|---|---|---|---|
| Normal | $(y-\mu)^2$ | $(-\infty,\infty)$ | $(-\infty,\infty)$ | 1 |
| Poisson | $2\left(y\log\left(\frac{y}{\mu}\right)-y+\mu\right)$ | $\{0,1,2,...\}$ | $(0,\infty)$ | $\mu$ |
| Binomial | $2\left\{y\log\left(\frac{y}{\mu}\right)+(n-y)\log\left(\frac{n-y}{n-\mu}\right)\right\}$ | $\{0,1,2,...,n\}$ | $(0,1)$ | $\mu(1-\mu)$ |
| Binomial Negativa | $2\left\{y\log\left(\frac{y}{\mu}\right)+(1-y)\log\left(\frac{1-y}{1-\mu}\right)\right\}$ | $\{0,1,2,...\}$ | $(0,\infty)$ | $\mu(1+\mu)$ |
| Gama | $2\left(\frac{y}{\mu}-\log\left(\frac{y}{\mu}\right)-1\right)$ | $(0,\infty)$ | $(0,\infty)$ | $\mu^2$ |
| Normal inversa | $\frac{(y-\mu)^2}{y\mu^2}$ | $(0,\infty)$ | $(0,\infty)$ | $\mu^3$ |
| von Mises | $2(1-\cos(y-\mu))$ | $(0,2\pi)$ | $(0,1)$ | 1 |
| Simplex | $\frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}$ | $(0,1)$ | $(0,1)$ | $\mu^3(1-\mu)^3$ |

## 2.2 Propriedade

Apesar da simplicidade de sua definição, um desvio unitário regular possui algumas propriedades úteis em relação ao seu comportamento próximo ao seu mínimo.

**Lema 1** *Jorgensen (1997) Um desvio unitário regular satisfaz*

$$\frac{\partial^2 d(\mu;\mu)}{\partial y^2} = \frac{\partial^2 d(\mu;\mu)}{\partial \mu^2} = \frac{\partial^2 d(\mu;\mu)}{\partial \mu \partial y} \quad \mu \in \Omega$$

*Prova:*

*Por definição, um desvio de unidade regular satisfaz as condições*

$$d(y;y) = d(\mu;\mu) = 0 \quad e \quad d(y;\mu) > 0 \quad \forall y \neq \mu$$

*Assim, a função $d(y;\cdot)$ tem um único mínimo em $y$, e similarmente $d(\cdot;\mu)$ tem um único mínimo em $\mu$, implicando que para todo $\mu \in \Omega$*

$$\frac{\partial d(\mu;\mu)}{\partial \mu} = 0 \qquad \frac{\partial d(\mu;\mu)}{\partial y} = 0.$$

*Diferenciando a primeira das equações acima em relação a $\mu$ obtemos, pela regra da cadeia:*

$$\frac{\partial^2 d(\mu;\mu)}{\partial \mu^2} + \frac{\partial^2 d(\mu;\mu)}{\partial \mu \partial y} = 0,$$

*e da mesma forma da segunda equação,*

$$\frac{\partial^2 d(\mu;\mu)}{\partial y^2} + \frac{\partial^2 d(\mu;\mu)}{\partial \mu \partial y} = 0.$$

*O resultado agora segue combinando essas duas equações.*

O Lema 1 nos dá três expressões equivalentes para a função de variância unitária,

$$V(\mu) = \frac{2}{\dfrac{\partial^2 d(\mu;\mu)}{\partial \mu^2}} = \frac{2}{\dfrac{\partial^2 d(\mu;\mu)}{\partial y^2}} = -\frac{2}{\dfrac{\partial^2 d(\mu;\mu)}{\partial \mu \partial y}}.$$

## 2.3 Distribuição Simplex

Desenvolvida por Barndorff-Nielsen e Jørgensen (1991) e posteriormente introduzida no grupo de modelos de dispersão por Jorgensen (1997), estendendo a classe dos modelos lineares generalizados (Nelder e Wedderburn, 1972). A distribuição Simplex é bastante conveniente e muito flexível quando o assunto é modelar dados restritos ao intervalo contínuo (0,1), que podem ser interpretada como proporções, taxas ou índices. Assim, uma variável aleatória $y$ que segue uma distribuição Simplex com média $\mu \in (0,1)$ e parâmetro de dispersão $\sigma^2 > 0$, sendo $y \sim S^-(\mu, \sigma^2)$, tem função de densidade de probabilidade da forma

$$f(y;\mu;\sigma^2) = \{2\pi\sigma^2[y(1-y)]^3\}^{-1/2} \exp\left\{\frac{-1}{2\sigma^2} d(y;\mu)\right\}, 0 < y < 1, \tag{2.2}$$

em que

$$d(y;\mu) = \frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}, \tag{2.3}$$

com E(y) = $\mu$ e

$$\text{Var(y)} = \mu(1-\mu) - \sqrt{\frac{1}{2\sigma^2}} \exp\left\{\frac{1}{\sigma^2\mu^2(1-\mu)^2}\right\} \Gamma\left\{\frac{1}{2}, \frac{1}{2\sigma^2\mu^2(1-\mu)^2}\right\},$$

onde $\Gamma(a,b)$ é a função gama incompleta definida por $\Gamma(a,b) = \int_a^\infty x^{\alpha-1}e^{-x}dx$. Definimos a função de variância de $Y$ como $V(\mu) = \mu^3(1-\mu)^3$, que é uma função $V : \Omega \to (0,\infty)$.

Diversas formas (simétrica e assimétrica) da densidade da distribuição Simplex, para diferentes valores de $(\mu, \sigma^2)$ são ilustradas na Figura 2.1.

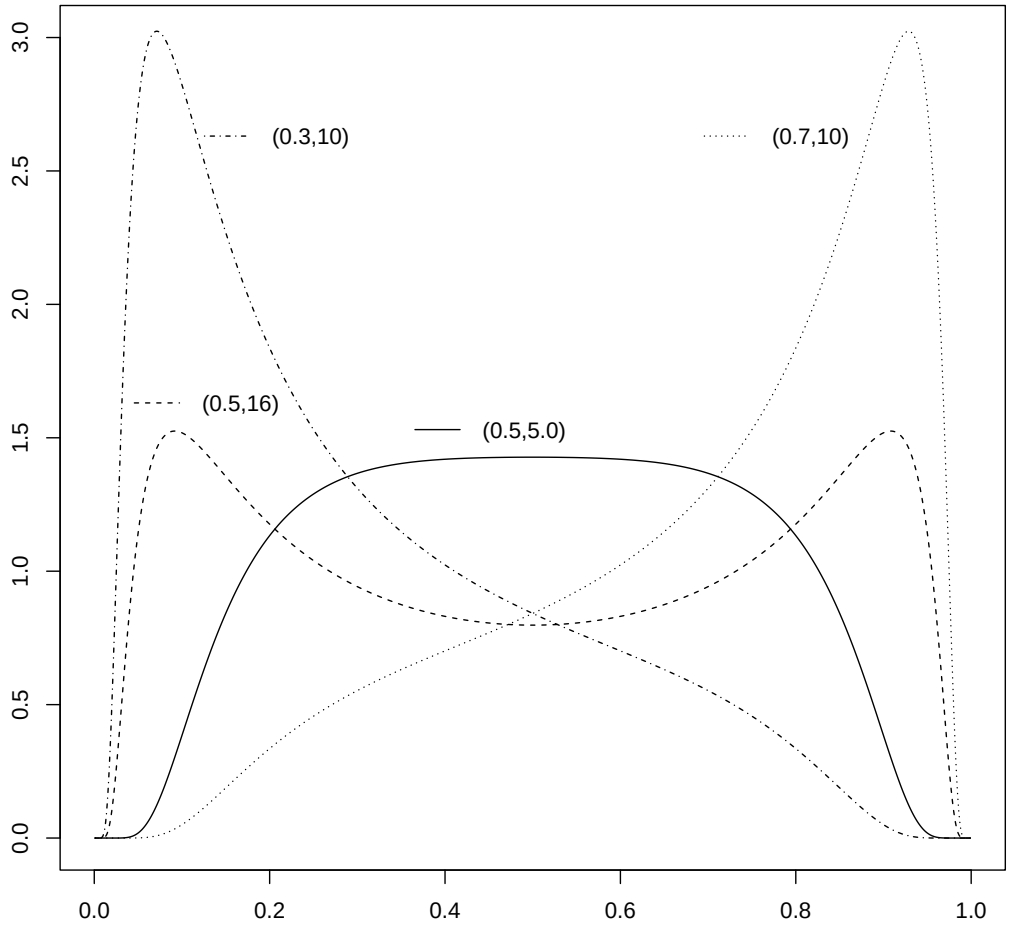



**Figura 2.1:** *Densidade da distribuição Simplex para valores dos parâmetros* $\mu = (0.5, 0.5, 0.7, 0.3)$ *e* $\sigma^2 = (\sqrt{5}, \sqrt{16}, \sqrt{10}, \sqrt{10})$.

### 2.3.1 Propriedades

Seja $y$ uma variável aleatória que segue uma distribuição Simplex com média $\mu$ e parâmetro de dispersão $\sigma^2$. As propriedades a seguir são definidas

1. $\mathrm{E}[\mathrm{d}'(\mathrm{y};\mu)] = 0$;
2. $\mathrm{Var}[\mathrm{d}(\mathrm{y};\mu)] = 2(\sigma^2)^2$;
3. $\mathrm{E}[\mathrm{d}(\mathrm{y};\mu)] = \sigma^2$;
4. $\mathrm{E}[(\mathrm{y}-\mu)\mathrm{d}(\mathrm{y};\mu)] = 0$;
5. $\mathrm{E}[(\mathrm{y}-\mu)\mathrm{d}^2(\mathrm{y};\mu)] = 0$;
6. $\mathrm{E}[(\mathrm{y}-\mu)\mathrm{d}''(\mathrm{y};\mu)] = -2\sigma^2$;
7. $\frac{1}{2}\mathrm{E}[(\mathrm{d}''(\mathrm{y};\mu)] = \frac{3\sigma^2}{\mu(1-\mu)} + \frac{1}{\mu^3(1-\mu)^3}$.

em que $d'(y;\mu) = \partial d(y;\mu)/\partial\mu$ e $d''(y;\mu) = \partial^2 d(y;\mu)/\partial\mu^2$. Maiores detalhes a respeito de tais propriedades podem ser visto, por exemplo, em Song e Tan (2000) e (Silva, 2016).

### 2.3.2 Estimação via Máxima Verossimilhança

Baseado em uma amostra de $n$ observações independentes, a partir de (2.2), podemos definir a função de verossimilhança da distribuição Simplex, que é dada por

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i \mid \mu, \sigma^2) = \prod_{i=1}^{n} \{2\pi\sigma^2[y_i(1-y_i)]^3\}^{-1/2} \exp\left(\sum_{i=1}^{n}\left\{\frac{-1}{2\sigma^2} d(y_i;\mu)\right\}\right).$$

onde $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$.

O logaritmo da função de verossimilhança para a distribuição Simplex, é dada por:

$$\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}), \tag{2.4}$$

em que

$$\begin{aligned} \ell_i(\boldsymbol{\theta}) &= -\frac{1}{2}\log[2\pi\sigma^2\{(y_i(1-y_i)\}^3) - \frac{1}{2\sigma^2} d(y,\mu)], \\ &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{3}{2}\log(y_i(1-y_i)) - \frac{1}{2\sigma^2} d(y_i;\mu). \end{aligned}$$

Para obter os estimadores de máxima verossimilhança para os parâmetros $\mu$ e $\sigma^2$, precisamos solucionar simultaneamente as equações de verossimilhança, a saber:

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \mu} = 0 \quad ; \quad \text{e} \quad \frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma^2} = 0. \tag{2.5}$$

Derivando $\ell(\boldsymbol{\theta})$ em relação a $\mu$, temos

$$\frac{\partial \ell(\theta)}{\partial \mu} = -\frac{1}{2\sigma^2} d'(y,\mu), \tag{2.6}$$

onde

$$d'(y,\mu) = -\frac{2(y-\mu)}{\mu(1-\mu)}\left[d(y,\mu) + \frac{1}{\mu^2(1-\mu)^2}\right].$$

Analogamente, derivando $\ell(\boldsymbol{\theta})$ em relação a $\sigma^2$, temos

$$\frac{\partial \ell(\theta)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{d(y,\mu)}{2(\sigma^2)^2}. \tag{2.7}$$

Igualando a Equação (2.7) a zero, temos que

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} d(y,\mu). \tag{2.8}$$

Faz-se necessário verificar se as segundas derivadas parciais do logaritmo da função de verossimilhança define uma matriz hessiana $\mathrm{H}(\mu,\sigma^2) = \partial^2\ell(\theta)/\partial\theta\partial\theta$,

$$\mathrm{H}(\mu,\sigma^2) = \begin{vmatrix} \dfrac{\partial^2\ell(\boldsymbol{\theta})}{\partial\mu^2} & \dfrac{\partial^2\ell(\boldsymbol{\theta})}{\partial\mu\partial\sigma^2} \\ \dfrac{\partial^2\ell(\boldsymbol{\theta})}{\partial\mu\partial\sigma^2} & \dfrac{\partial^2\ell(\boldsymbol{\theta})}{\partial(\sigma^2)^2} \end{vmatrix}.$$

é estritamente convexa (positiva), para garantir que a solução em (2.5) é ponto de máximo.

Diferentemente do estimador de máxima verossimilhança de $\widehat{\mu}$, que só pode ser obtido por métodos numéricos, o estimador de máxima verossimilhança de $\widehat{\sigma}^2$ pode ser obtido analiticamente, pois possui uma forma explicita.

As segundas derivadas de $\ell(\boldsymbol{\theta})$ em relação a $\mu$ é dada por

$$\frac{\partial^2 \ell(\theta)}{\partial \mu^2} = -\frac{1}{2\sigma^2} d''(y,\mu), \quad \text{sendo} \quad d''(y,\mu) = \frac{\partial^2 d(y,\mu)}{\partial \mu^2}. \tag{2.9}$$

Pela propriedade propriedade 7 dada na Seção 2.3.1, temos que

$$\mathrm{E}\left[\frac{\partial^2 \ell(\theta)}{\partial \mu^2}\right] = -\frac{1}{\sigma^2}\left\{\frac{1}{2}\mathrm{E}[\mathrm{d}''(\mathrm{y},\mu)]\right\} = -\frac{3}{\mu(1-\mu)} - \frac{1}{\sigma^2\mu^3(1-\mu)^3}.$$

Analogamente, a segunda derivada de $\ell\boldsymbol{\theta})$ em relação a $\sigma^2$ é dada por

$$\begin{aligned}\frac{\partial^2 \ell(\theta)}{\partial (\sigma^2)^2} &= \frac{\partial}{\partial \sigma^2}\left(-\frac{1}{2\sigma^2} + \frac{d(y,\mu)}{2(\sigma^2)^2}\right) \\ &= \frac{1}{2\sigma^4} - \frac{d(y,\mu)}{\sigma^6},\end{aligned}$$

e pela propriedade 3, dada na Seção 2.3.1, temos que

$$\mathrm{E}\left[\frac{\partial^2 \ell(\theta)}{\partial (\sigma^2)^2}\right] = -\frac{\mathrm{n}}{2\sigma^4}. \tag{2.10}$$

Finalmente, a derivada de segunda ordem de $\ell(\boldsymbol{\theta})$ com respeito a $\mu$ e $\sigma^2$ é dada por:

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \sigma^2 \partial \mu} = \frac{\partial}{\partial \sigma^2}\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \mu}\right) = -\frac{1}{\sigma^4} d'(y,\mu).$$

Aplicando a propriedade 1 dada na Seção 2.3.1, segue que

$$\mathrm{E}\left[\frac{\partial^2 \ell(\mu,\sigma^2)}{\partial \mu \partial \sigma^2}\right] = 0.$$

Sob condições de regularidade e para grande amostras, a distribuição aproximada dos estimadores de máxima verossimilhança é normal com média $\mu$ e matriz de variância e covariância $K^{-1}(\mu,\sigma^2)$, matriz da informação de Fisher, definida como

$$K(\mu,\sigma^2) = \begin{bmatrix} K_{\mu\mu} & 0 \\ 0 & K_{\sigma^2\sigma^2} \end{bmatrix}, \tag{2.11}$$

onde $K_{\mu\mu}$ e $K_{\sigma^2\sigma^2}$ são os estimadores de máxima verossimilhança definido em (2.6) e (2.8), respectivamente.

## 2.4 Critério de informação e seleção

Nas equações subsequentes os Critério de Informação Akaike (AIC), Critério de Informação Bayesiano (BIC) e Pseudo-$\mathrm{R}^2$ são expressos, respectivamente.

O Critério de Informação Akaike Akaike (1974) é dado por

$$\begin{aligned} \text{AIC} &= -2\sum_{i=1}^{n} \log f(x_n|\widehat{\theta}) + 2p \\ &= -2\log L(\widehat{\theta}) + 2p, \end{aligned} \tag{2.12}$$

onde $p$ é o número de parâmetros a serem estimados no modelo, $L(\widehat{\theta})$ é a função de verossimilhança. O termo $2p$ é o termo de penalidade e atua como uma compensação pelo viés na falta de ajuste quanto os estimadores de máxima verossimilhança são usados.

Proposto por Schwarz (1978), o Critério de Informação Bayesiano (BIC) é dado por

$$\begin{aligned} \text{BIC} &= -2\sum_{i=1}^{n} \log f(x_n|\widehat{\theta}) + p\log(n) \\ &= -2\log L(\widehat{\theta}) + p\log(n), \end{aligned} \tag{2.13}$$

em que $n$ é o número de observações da amostra, $p$ é o número de parâmetros a serem estimados no modelo e $L(\widehat{\theta})$ é a função de verossimilhança.

Segundo Burnham e Anderson (2004), o AIC e BIC são critérios assintóticos para comparar modelos encaixados, mas também podem ser aplicados em modelos não encaixados (Moura *et al.*, 2021).

O coeficiente de determinação para um modelo de regressão linear é a redução proporcional no erro de previsão ao quadrado usando a previsão do modelo em vez da média. Ou seja, se tivermos n observações, um vetor $X$ de $p$ preditores e um modelo $\mathrm{E[y]} = \mu = \alpha + \mathrm{x}\beta$, definimos

$$\mathrm{R}^2 = 1 - \left( \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - \widehat{\mu})^2 / \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - \overline{\mathrm{y}})^2 \right).$$

Em paralelo as ideias do coeficiente de determinação, Nagelkerke *et al.* (1991) definem o Pseudo-$\mathrm{R}^2$, que corresponde a uma nova versão melhorada do Pseudo-$\mathrm{R}^2$ de Cox e Snell (1989), uma vez que restringe o seu valor entre 0 e 1, expresso pela equação:

$$\mathrm{R}^2_{\mathrm{N}} = 1 - \left[ \left( \mathrm{L}(0)/\mathrm{L}(\widehat{\beta}) \right)^{2/\mathrm{n}} / \left( 1 - \mathrm{L}(0)^{2/\mathrm{n}} \right) \right], \tag{2.14}$$

em que $L(0)$ é o valor máximo da função de verossimilhança do modelo somente com intercepto, $L(\widehat{\beta})$ é o valor máximo da função de verosimilhança do modelo completo e $n$ é o tamanho da amostra (Veall e Zimmermann, 1996).

## 2.5 Aplicação

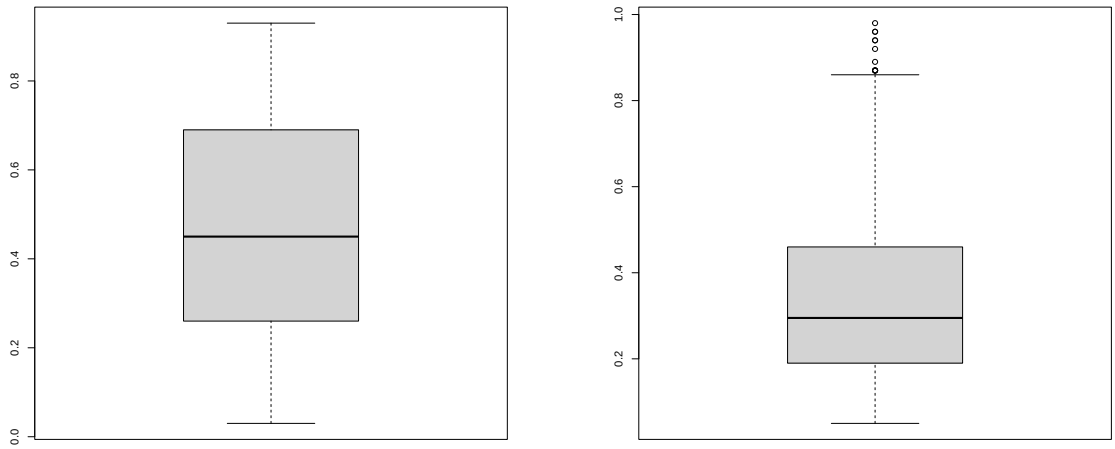
Consideremos as variáveis: proporção de pobres $y_1$ e taxa de mortalidade infantil $y_2$ dos 5.506 municípios brasileiros referente ao ano de 2000, disponível em (Municipal, 2023). Nosso objetivo aqui é encontrar as estimativas de máxima verossimilhança para os parâmetros da distribuição Simplex. Alternativamente, é encontrada as estimativas de máxima verossimilhança (MV) dos parâmetros para distribuição Beta, esta comumente utilizada na literatura. A Tabela 2.2 apresenta as medidas descritivas dos dados, ou seja, a proporção média de pobres entre os municípios brasileiros é de 46%, com proporção mínima de pobres, cerca de 0,03, apresentada pelos municípios de Fernando de Noronha/PE e São Caetano do Sul/SP, enquanto que a proporção máxima, cerca de 0,93, compreendendo os municípios de Belágua/MA, Jordão/AC e Manai/PE. Já a taxa média de mortalidade infantil é de 0,34 entre os municípios brasileiros, com taxa mínima de mortalidade

infantil, cerca de 0,05, apresentada pelo município de Quatro Pontes/PR, enquanto que a taxa máxima de mortalidade infantil é de 0,98 e está presente no município de Águas Belas/PE.

**Tabela 2.2:** *Resumo descritivo das variáveis proporção de pobres e taxa de mortalidades infantil.*

| Estatística | Variáveis | |
|---|---|---|
| | Proporção de pobres | Taxa de mortalidade infantil |
| Mínimo | 0,03 | 0,05 |
| Máximo | 0,93 | 0,98 |
| 1º Quantil | 0,26 | 0,19 |
| 3º Quantil | 0,69 | 0,46 |
| Média | 0,46 | 0,34 |
| Desvio padrão | 0,23 | 0,18 |
| Assimetria | 0,02 | 0,71 |
| Curtose | -1,36 | -0,35 |

Na Figura 2.2, constata-se, pelo gráfico de *boxplot*, que a variável taxa de mortalidade infantil apresenta assimetria a direita com a presença de alguns *outliers* que compreendem os municípios de Águas Belas, Lago dos Gatos, Jurema, Jucati e Iati no estado de Pernambuco.




**(a)** **(b)**

**Figura 2.2:** *Boxplot das variáveis proporção de pobres (a) e Taxa de mortalidade infantil (b).*

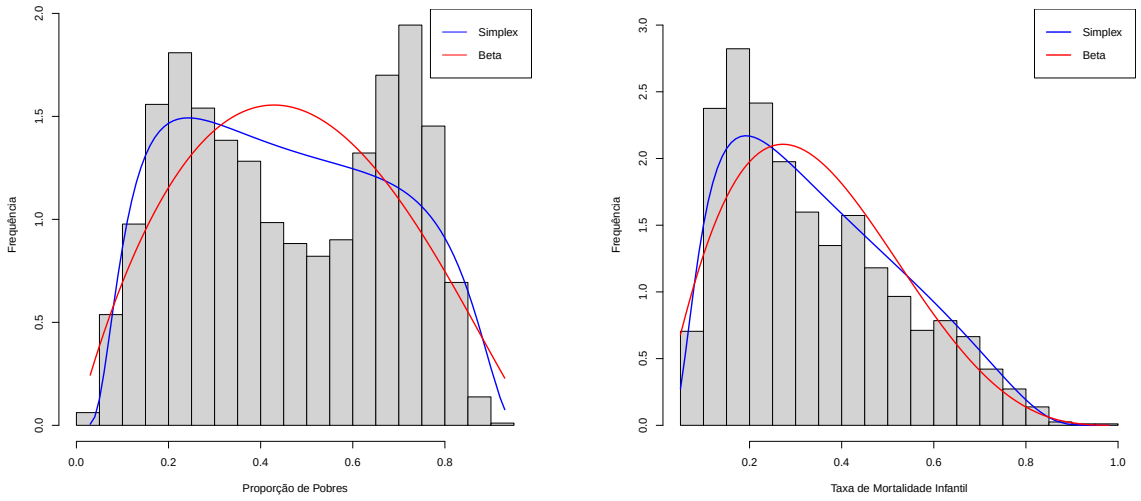
Na Tabela 2.3 são apresentados as estimativas, obtidas a partir do pacote *"gamlss"*, para os parâmetros das distribuições Simplex e Beta[1] com seus respectivos erros padrão (SE) e valor-$p$ para as variáveis taxa de mortalidade infantil e proporção de pobres.

**Tabela 2.3:** *Estimativa, SE e valor-p para os parâmetros das distribuições Simplex e Beta.*

| Variável\Distribuição | | Parâmetro | Estimativa | SE | valor-$p$ |
|---|---|---|---|---|---|
| $y_1$ | Simplex | $\mu$ | 0,455 | 0,003 | 0,000 |
| | | $\sigma^2$ | 2,407 | 0,023 | 0,000 |
| | Beta | $\mu$ | 0,514 | 0,002 | 0,000 |
| | | $\sigma^2$ | 0.901 | 0,004 | 0,000 |
| $y_2$ | Simplex | $\mu$ | 0,352 | 0,002 | 0,000 |
| | | $\sigma^2$ | 2,094 | 0,019 | 0,000 |
| | Beta | $\mu$ | 0,481 | 0,002 | 0,000 |
| | | $\sigma^2$ | 0,717 | 0,004 | 0,000 |

Proporção de pobres: [Simplex (AIC = -2120,16; BIC = -2106,94), Beta (AIC = -1608,88; BIC = -1595,65)]; Taxa de mortalidade infantil: [Simplex (AIC = -4302,62; BIC = -4289,39), Beta (AIC = -3972,56; BIC = -3959,33)].

Mediante estimação de parâmetros e análise gráfica das densidades estimadas para ambos os modelos, nota-se que a distribuição Simplex se prevalece a distribuição Beta, pois, conforme apresentado na Figura 2.3, vemos que a distribuição Simplex apresenta um melhor ajuste aos dados tanto para taxa de mortalidade infantil quanto para a proporção de pobres. É valido ressaltar, que o aspecto gráfico apresentado pela variável proporção de pobres ($y_1$) em seu histograma, na Figura 2.3 (a), tende a ser melhor compreendido pela distribuição Simplex, uma vez que, pela Figura 2.1, temos que uma das formas da função de densidade Simplex remete ao padrão gráfico especificado.




**(a)** **(b)**

**Figura 2.3:** *Histograma e curvas de densidades ajustadas para as variáveis proporção de pobres (a) e taxa de mortalidade infantil (b) para as distribuições Simplex e Beta, respectivamente.*

Por fim, pelos critério de informação de Akaike (AIC) e critério de informação Bayesiano (BIC), fica evidente que em ambas as situação a distribuição Simplex ajusta-se melhor aos dados, apresenta menor AIC quanto BIC, conforme consta na Tabela 2.3. Em outras palavras, a distribuição Simplex tende a representar melhor os dados comparado a distribuição Beta, principalmente quando

[1]Na família *GAMLSS*, a distribuição Beta tem reparametrização: $\sigma^2 = \frac{1}{1+\phi} \rightarrow \phi = \frac{(1-\sigma^2)}{\sigma^2}$, sendo $\phi > 1$ e $0 < \sigma^2 < 1$.

os dados apresentam uma certa ondulação (aspecto bimodal), pois é uma das características da distribuição Simplex.

# Capítulo 3

# Modelo de Regressão Simplex Univariada

Neste capítulo é apresentado o modelo de regressão baseado na distribuição Simplex, o vetor escore, matriz Hessiana e a matriz de informação de Fisher para o vetor de parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$.

## 3.1 Modelo de Regressão Simplex univariada

Sejam $y_1, y_2, y_3, ..., y_n$ variáveis aleatórias independentes, sendo cada $y_i \sim S^-(\mu_i, \sigma_i^2)$, $i = 1, 2, 3, ..., n$. O modelo de regressão Simplex é definido pela função de densidade (2.2), em que os componentes sistemáticos para o parâmetro $\mu$ e $\sigma^2$, são dados por:

$$g(\mu_i) = \sum_{l=0}^{k} x_{il}\beta_l = \eta_i \qquad \text{e} \qquad h(\sigma_i^2) = \sum_{j=0}^{p} z_{ij}\gamma_j = \xi_i,$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)^\top$ e $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, ..., \gamma_q)^\top$ são vetores de parâmetros desconhecidos, $\boldsymbol{\beta} \in R^k$ e $\boldsymbol{\gamma} \in R^q$, $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_k)^\top$ e $\boldsymbol{\xi} = (\xi_1, \xi_2, ..., \xi_q)^\top$ são preditores lineares, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ik})^\top$ e $\boldsymbol{z}_i = (z_{i1}, z_{i2}, ..., z_{iq})\top$ são observações em $k$ e $q$ covariáveis conhecidas, $g(\cdot)$ e $h(\cdot)$ funções de ligação estritamente monótonas e duas vezes diferenciáveis, sendo $g : (0, 1) \to R$ e $h : (0, \infty) \to R$. Para $g(\cdot)$ e $h(\cdot)$ diferentes funções de ligações podem ser utilizadas: para $\mu$, por exemplo, a função *Logit* $g(\mu) = \log(\mu/1 - \mu)$ e para $\sigma^2$ temos, por exemplo, a função logarítmica $h(\sigma^2) = \log(\sigma^2)$. Para todo $i = 1, 2, ..., k$ temos que

$$g(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right) = x_{il}^\top \beta_l,$$

consequentemente,

$$\mu_i = \frac{\exp\left(x_i^\top \beta_l\right)}{1 + \exp\left(x_i^\top \beta_l\right)}. \tag{3.1}$$

A equação (3.1) é uma função inversa de $g(\mu_i)$. Aqui, os parâmetros têm uma importante interpretação. Suponha que o valor da $i$-esima variável regressora é aumentado por $\omega$ unidades e todas as outras variáveis independentes permanecem inalteradas. Seja $\mu^*$ a média de $y$ sob este novo valor das covariáveis, uma vez que $\mu$ denota a média de $y$ sob o valor original das covariáveis. Temos então que

$$\frac{\mu^*}{1 - \mu^*} = \exp(x_{i1}\beta_1 + ... + (x_{il} + \omega)\beta_l + ... + x_{ik}\beta_l),$$

facilmente podemos verificar que

$$\exp(\omega\beta_i) = \frac{\mu^*/(1-\mu^*)}{\mu/(1-\mu)} \tag{3.2}$$

isto é, $\exp(\omega\beta_i)$ é a *odds ratio*, ou seja, a razão de chance (de Oliveira, 2004).

Dada uma amostra de $n$ observações independentes e identicamente distribuídas com uma distribuição Simplex, temos que a função de verossimilhança é

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i \mid \mu_i, \sigma_i^2) = \prod_{i=1}^{n} \{2\pi\sigma_i^2[y_i(1-y_i)]^3\}^{-1/2} \exp\left\{\sum_{i=1}^{n}\left(-\frac{1}{2\sigma_i^2}d(y_i;\mu_i)\right)\right\}, \tag{3.3}$$

em que $\boldsymbol{\theta} = (\beta, \gamma)^\top$.

O logaritmo da função de verossimilhança é dada por:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\mu_i, \sigma_i^2), \tag{3.4}$$

com

$$\ell_i(\boldsymbol{\theta}) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma_i^2 - \frac{3}{2}\log[y_i(1-y_i)] - \frac{1}{2\sigma_i^2}d(y_i;\mu_i),$$

Derivando o logaritmo da função de verossimilhança em relação a cada $\boldsymbol{\beta}_l$, para $l = 1, 2, ..., k$, temos

$$\frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}_l} = \sum_{i=1}^{n} \frac{\partial\ell_i(\mu_i,\sigma_i^2)}{\partial\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial\eta_i}{\partial\beta_l},$$

em que $d\mu_i/d\eta_i = 1/g'(\mu_i)$, $\partial\eta_i/\partial\boldsymbol{\beta}_i = x_{il}$ e

$$\frac{\partial\ell_i(\mu_i,\sigma_i^2)}{\partial\mu_i} = \sum_{i=1}^{n} -\frac{1}{2\sigma_i^2}d'(y_i,\mu_i), \tag{3.5}$$

tal que

$$d'(y_i;\mu_i) = -\frac{2(y_i-\mu_i)}{\mu_i(1-\mu_i)}\left(d(y_i;\mu_i) + \frac{1}{\mu_i^2(1-\mu_i)^2}\right).$$

Diferenciando o logaritmo da função de verossimilhança com relação a $\gamma_j$, para $j = 1, 2, ..., q$, temos

$$\frac{\partial\ell(\boldsymbol{\theta})}{\partial\gamma_j} = \sum_{i=1}^{n} \frac{\partial\ell_i(\mu_i,\sigma_i^2)}{\partial\sigma_i^2}\frac{d\sigma_i^2}{d\xi_i}\frac{\partial\xi_i}{\partial\gamma_j},$$

em que $\partial\sigma_i^2/\partial\xi_i = 1/h'(\sigma_i^2)$, $\partial\xi_i/\partial\gamma_i = z_{ij}$ e

$$\frac{\partial\ell_i(\mu_i,\sigma_i^2)}{\partial\sigma_i^2} = -\frac{1}{2\sigma_i^2} + \frac{d(y_i;\mu_i)}{2(\sigma_i^2)^2}. \tag{3.6}$$

Matricialmente o vetor escore para o parâmetro $\boldsymbol{\beta}$ é expressa da seguinte forma

$$U_\beta(\boldsymbol{\theta}) = X^\top\Sigma TU(y_i - \mu_i),$$

em que $X$ é uma matriz $n \times k$ cuja $t$-ésima linha é $\boldsymbol{x}_i = (x_{i1}, ..., x_{ik})$, $\Sigma = \text{diag}(1/\sigma_1^2, ..., 1/\sigma_n^2)$, $T$ e

$U$ são dadas, respectivamente, por

$$T = \text{diag}(1/g'(\mu_1), ..., 1/g'(\mu_n)) \quad e \quad U = (u_1, ..., u_n),$$

sendo

$$u_i = \frac{1}{\mu_i(1-\mu_i)}\left(d(y_i;\mu_i) + \frac{1}{\mu_i^2(1-\mu_i)^2}\right), \forall i = 1, 2, ..., n.$$

Analogamente, temos que

$$U_\gamma(\boldsymbol{\theta}) = Z^\top Q\boldsymbol{a},$$

em que $Q = \text{diag}\{1/h'(\sigma_1^2), ..., 1/h'(\sigma_n^2)\}$, $Z$ é uma matriz $n \times p$ cuja $t$-ésima linha é $\boldsymbol{z}_i = (z_{i1}, ..., z_{iq})^\top$ e $\boldsymbol{a} = (a_1, ..., a_n)^\top$, com

$$a_i = -\frac{1}{2\sigma_i^2} + \frac{d(y_i;\mu_i)}{2(\sigma_i^2)^2}.$$

As segundas derivadas do logaritmo da função de verossimilhança com respeito a $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$, para $i = 1, ..., k$ e $s = 1, 2, ..., k$, é

$$\begin{aligned}
\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}_i\partial\boldsymbol{\beta}_s} &= \frac{\partial}{\partial\boldsymbol{\beta}_s}\left(\sum_{i=1}^{n}\frac{\partial\ell_i(\mu_i,\sigma_i^2)}{\partial\mu_i}\frac{d(\mu_i)}{d\eta_i}\frac{\partial\eta_i}{\partial\boldsymbol{\beta}_i}\right), \\
&= \sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\beta}_s}\left(\frac{\partial\ell_i(\mu_i,\sigma_i^2)}{\partial\mu_i}\frac{d\mu_i}{d\eta_i}\right)x_{il}, \\
&= \sum_{i=1}^{n}\left(\frac{\partial^2\ell_i(\mu_i,\sigma_i^2)}{\partial\mu_i^2}\frac{d\mu_i}{d\eta_i} + \frac{\partial\ell_i(\mu_i,\sigma_i^2)}{\partial\mu_i}\frac{\partial}{\partial\mu_i}\frac{d\mu_i}{d\eta_i}\right)\frac{d\mu_i}{d\eta_i}x_{il}x_{is}.
\end{aligned} \tag{3.7}$$

É possível mostrar que $\text{E}[\partial\ell_\text{i}(\mu_\text{i},\sigma_\text{i}^2)/\partial\mu_\text{i}] = 0$. Logo,

$$\text{E}\left[\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}_\text{i}\partial\boldsymbol{\beta}_\text{s}}\right] = \text{E}\left[\sum_{\text{i}=1}^{\text{n}}\text{E}\left[\frac{\partial^2\ell_\text{i}(\boldsymbol{\theta})}{\partial\mu_\text{i}^2}\right]\right]\left(\frac{\text{d}\mu_\text{i}}{\text{d}\eta_\text{i}}\right)^2 x_{il}x_{is}.$$

De 3.5 temos que

$$\frac{\partial\ell_i(\mu_i,\sigma_i^2)}{\partial\mu_i} = -\frac{1}{2\sigma_i^2}d''(y_i,\mu_i),$$

e pela propriedade 7, na Seção 2.3.1, temos

$$\frac{1}{2}\text{E}[d''(y_i;\mu_i)] = \frac{3\sigma_i^2}{\mu_i(1-\mu_i)} + \frac{1}{\mu_i^3(1-\mu_i)^3},$$

logo, a esperança de 3.7 será

$$\text{E}\left[\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}_\text{i}\partial\boldsymbol{\beta}_\text{s}}\right] = -\sum_{i=i}^{n}\frac{1}{\sigma_i^2}w_\text{t}x_{il}x_{is},$$

sendo

$$w_t = \left( \frac{3\sigma_i^3}{\mu_i(1-\mu_i)} + \frac{1}{\mu_i^3(1-\mu_i)^3} \right) \frac{1}{[g'(\mu_i)]^2}.$$

Matricialmente, temos que a matriz de informação de Fisher para $\boldsymbol{\beta}$ é dada por

$$K_{\beta\beta} = -\mathrm{E}\left[ \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_\mathrm{i} \partial \boldsymbol{\beta}_\mathrm{s}} \right] = \mathrm{X}^\top \Sigma \mathrm{W} \mathrm{X}, \tag{3.8}$$

em que $W = \mathrm{diag}(w_1, ..., w_n)$ e $\Sigma = \mathrm{diag}(1/\sigma_1^2, ..., 1/\sigma_n^2)$.

Por fim, a derivada de segunda ordem de $\ell(\boldsymbol{\theta})$ com relação a $\boldsymbol{\gamma}_j$ e $\boldsymbol{\gamma}_r$, para $j, r = 1, ..., q$ é dada por

$$\begin{aligned}
\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_j \partial \boldsymbol{\gamma}_r} &= \frac{\partial}{\partial \boldsymbol{\gamma}_r} \left( \sum_{i=1}^{n} \frac{\partial \ell_i(\mu_i, \sigma_i^2)}{\partial \sigma_i^2} \frac{d(\sigma_i^2)}{d\xi_i} \frac{\partial \xi_i}{\partial \boldsymbol{\gamma}_j} \right), \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\gamma}_r} \left( \frac{\partial \ell_i(\mu_i, \sigma_i^2)}{\partial \sigma_i^2} \frac{d\sigma_i^2}{d\xi_i} \right) z_{ij}, \\
&= \sum_{i=1}^{n} \left( \frac{\partial^2 \ell_i(\mu_i, \sigma_i^2)}{\partial (\sigma_i^2)^2} \frac{d\sigma_i^2}{d\xi_i} + \frac{\partial \ell_i(\mu_i, \sigma_i^2)}{\partial \sigma_i^2} \frac{\partial}{\partial \sigma_i^2} \frac{d\sigma_i^2}{d\xi_i} \right) \frac{d\sigma_i^2}{d\xi_i} z_{ij} x_{ir}.
\end{aligned}$$

Temos que $\mathrm{E}[\partial \ell_\mathrm{i}(\mu_\mathrm{i}, \sigma_\mathrm{i}^2)/\partial \sigma_\mathrm{i}^2] = 0$, logo

$$\mathrm{E}\left[ \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_\mathrm{j} \partial \boldsymbol{\gamma}_\mathrm{r}} \right] = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{E}\left[ \frac{\partial^2 \ell_\mathrm{i}(\mu_\mathrm{i}, \sigma_\mathrm{i}^2)}{\partial (\sigma_\mathrm{i}^2)^2} \right] \left( \frac{\mathrm{d}\sigma_\mathrm{i}^2}{\mathrm{d}\xi_\mathrm{i}} \right) z_{ij} z_{ir}.$$

A partir da equação (3.6), temos que

$$\frac{\partial \ell_i(\mu_i, \sigma_i^2)}{\partial (\sigma_i^2)^2} = \frac{1}{2(\sigma_i^2)^2} + \frac{d(y_i; \mu_i)}{(\sigma_i^2)^3},$$

logo

$$\mathrm{E}\left[ \frac{\partial \ell_\mathrm{i}(\mu_\mathrm{i}, \sigma_\mathrm{i}^2)}{\partial (\sigma_\mathrm{i}^2)^2} \right] = \frac{1}{2(\sigma_i^2)^2} - \frac{1}{(\sigma_i^2)^3} \mathrm{E}[d(y_i; \mu_i)].$$

A partir da propriedade 3 na seção 2.3.1, segue que

$$\mathrm{E}\left[ \frac{\partial \ell_\mathrm{i}(\mu_\mathrm{i}, \sigma_\mathrm{i}^2)}{\partial (\sigma_\mathrm{i}^2)^2} \right] = \frac{1}{2(\sigma_\mathrm{i}^2)^2}.$$

e, portanto,

$$\mathrm{E}\left[ \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_\mathrm{j} \partial \boldsymbol{\gamma}_\mathrm{r}} \right] = -\sum_{i=1}^{n} v_i z_{ij} z_{ir},$$

sendo

$$v_i = \frac{1}{2(\sigma_i^2)^2}\frac{1}{[h'(\sigma_i^2)]^2}.$$

Matricialmente, temos que a matriz de informação de Fisher para $\gamma$ é dada por

$$K_{\gamma\gamma} = -\mathrm{E}\left[\frac{\partial^2 \ell(\theta)}{\partial\gamma_{\mathrm{j}}\partial\gamma_{\mathrm{r}}}\right] = \mathrm{Z}^\top \mathrm{V}\mathrm{Z}, \tag{3.9}$$

em que $V = \mathrm{diag}(v_1, ..., v_n)$.

Finalmente, a matriz de informação de Fisher para o vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ é dada por

$$K = K(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \begin{pmatrix} K_{\boldsymbol{\beta}\boldsymbol{\beta}} & 0 \\ 0 & K_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \end{pmatrix},$$

em que $K_{\boldsymbol{\beta}\boldsymbol{\beta}}$ e $K_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$ estão definidas em (3.8) e (3.9), respectivamente.

Sob condições gerais de regularidade e para grandes amostras, a distribuição aproximada dos estimadores de máxima verossimilhança segue uma distribuição normal com média $\mu$ e matriz de variâncias e covariâncias $K^{-1}$.

## 3.2 Aplicação

Temos agora por interesse conhecer a relação entre as variáveis: $y_1$: proporção de pobres e $y_2$: taxa de mortalidade infantil, com respeito a variável $x$: Índice de Desenvolvimento Humano (IDH) por município (como uma variável independente). Os dados podem ser acessados em (Municipal, 2023). Na Figura 3.1 é apresentado as características gráficas da variável $x$, através do gráfico de histograma e *boxplot*, onde podemos notar uma leve assimetria a esquerda. A Tabela 3.1 apresenta as medidas descritiva para a variável IDH por município, ou seja, os municípios brasileiros apresentam em média IDH por município de aproximadamente 0,699, com IDH por município máximo de aproximadamente de 0,920 e mínimo de aproximadamente de 0,480. Notamos também que a distribuição dos dados é platicúrtica, ou seja, têm caudas mais "leves" comparado a distribuição normal, pois apresenta coeficiente de curtose negativo, ver Tabela 3.1.

**Tabela 3.1:** *Medidas descritivas da variável Índice de Desenvolvimento Humano (IDH) municipal.*

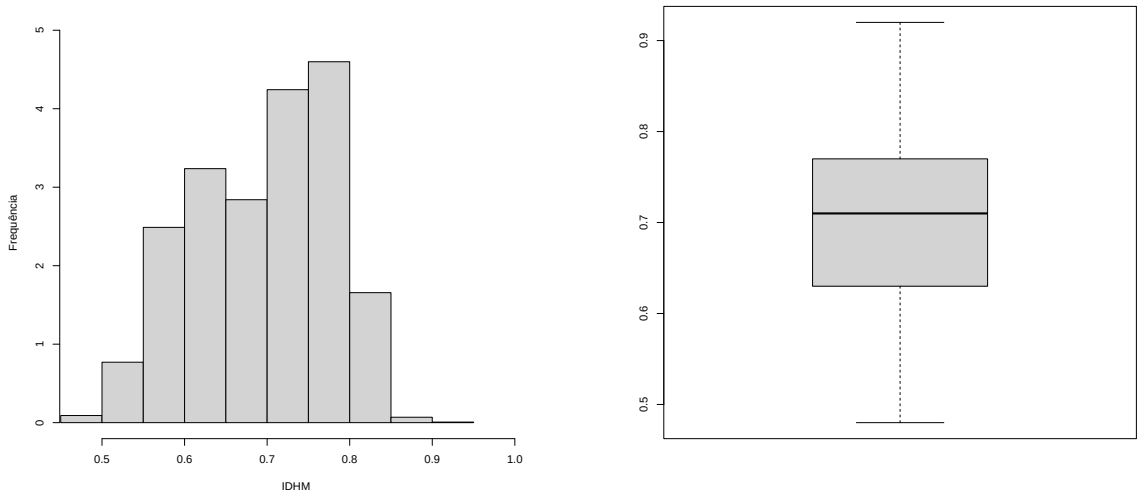
| Estatística | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Máximo | Mínimo | 1º Quartil | 3º Quartil | Média | Mediana | Desvio Padrão | Assimetria | Curtose |
| 0,920 | 0,480 | 0,630 | 0,770 | 0,699 | 0,710 | 0,083 | −0,295 | −0,925 |

O modelo de regressão Simplex é definido pela densidade em 2.2, cujas componentes sistemáticas para a regressão Simplex é definida por

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1}, \quad \text{e} \quad h(\sigma_i^2) = \gamma_0 + \gamma_1 x_{i1} \quad \text{para} \quad i = 1, ..., 5.506,$$

onde $g(\mu_i) = \log[\mu_i/(1-\mu_i)]$, e $h(\sigma_i^2) = \log(\sigma_i^2)$. De forma análoga podemos considerar para o modelo de regressão Beta.

Na Tabela 3.2, apresenta-se as estimativas dos parâmetros para os modelos de regressões Simplex e Beta, com o intuito de explicar a relação entre a proporção de pobres e o IDH por município (Modelo 1) e a relação entre a taxa de mortalidade infantil e o IDH por município (Modelo 2). Em ambos os modelos as conclusões são similares, ou seja, tanto o intercepto ($\beta_0$) quanto o parâmetro ($\beta_1$) associado a variável IDH por município são significativas ao nível de 1%. Pela equação




**(a)** **(b)**

**Figura 3.1:** *Gráficos de histograma (a) e boxplot (b) da variáveis Índice de desenvolvimento humano - IDHM.*

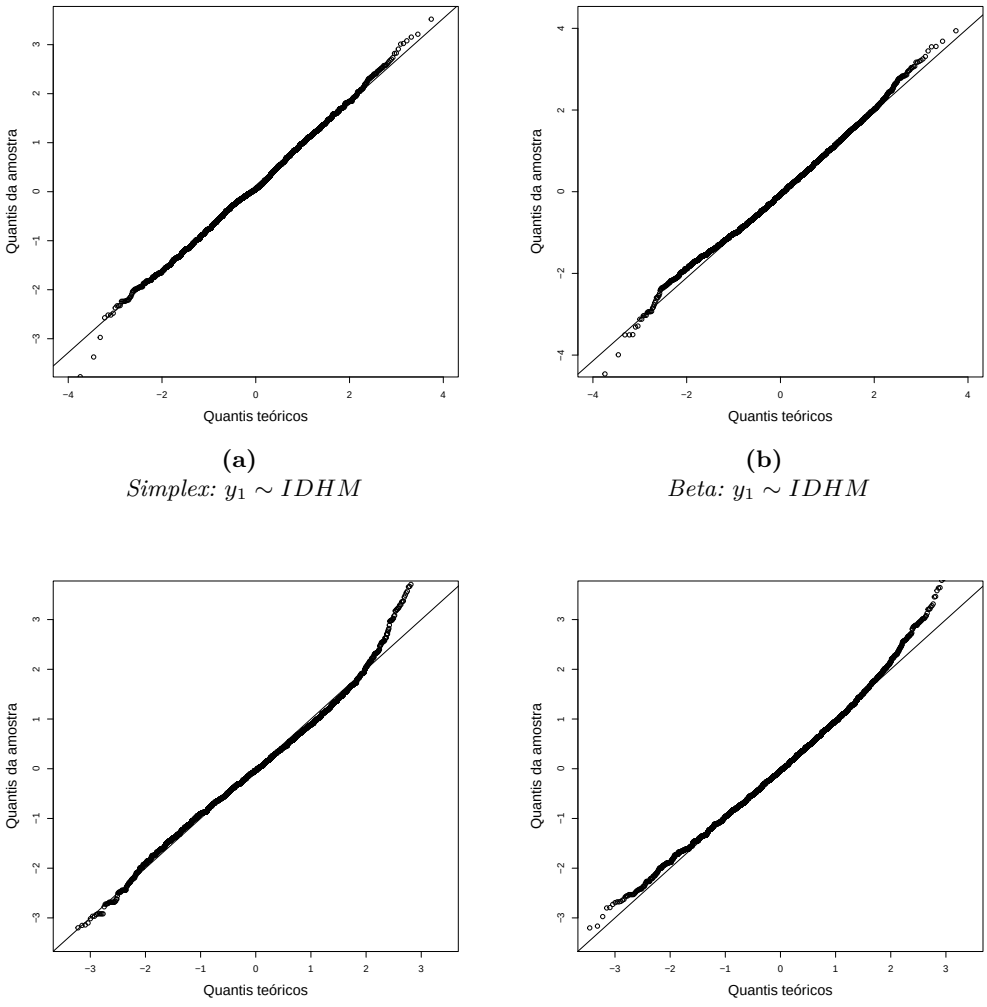
(3.2), podemos concluir que $\exp(0,01 \times \widehat{\beta}_{11}) = \exp(-11,229) \simeq 0,894$, ou seja, com o aumento de 1% no IDH por município, a chance de redução da proporção de pobres é de aproximadamente 10,62% através do modelo de regressão Simplex e $\exp(0,01 \times \widehat{\beta}_{11}) = \exp(-11,709) \simeq 0,894$, ou seja, com o aumento de 1% no IDH por município, a chance de redução da proporção de pobres é de aproximadamente 11,05% através do modelo de regressão Beta - (Modelo 1). Respectivamente, $\exp(0,01 \times \widehat{\beta}_{11}) = \exp(-9,491) \simeq 0,909$, ou seja, com o aumento de 1% no IDH por município, a chance de redução da taxa de mortalidade infantil é de aproximadamente 9,05% através do modelo de regressão Simplex e $\exp(0,01 \times \widehat{\beta}_{11}) = \exp(-9,573) \simeq 0,908$, ou seja, com o aumento de 1% no IDH por município, a chance de redução da taxa de mortalidade infantil é de aproximadamente 9,13% através do modelo de regressão Beta - (Modelo 2).

**Tabela 3.2:** *Estimativas, erros padrão e valor-p dos parâmetros do modelo de regressão univariado das distribuições Simplex e Beta.*

| Modelo | Distribuição | Parâmetro | Estimativa | SE | $p$-valor |
|---|---|---|---|---|---|
| $y_1 \sim IDHM$ | Simplex | $\beta_0$ | 7,656 | 0,040 | 0,000 |
| | | $\beta_1$ | −11,229 | 0,060 | 0,000 |
| | | $\gamma_0$ | −1,369 | 0,079 | 0,000 |
| | | $\gamma_1$ | 1,737 | 0,113 | 0,000 |
| | Beta | $\beta_0$ | 8,042 | 0,049 | 0,000 |
| | | $\beta_1$ | −11,709 | 0,072 | 0,000 |
| | | $\gamma_0$ | −2,324 | 0,107 | 0,000 |
| | | $\gamma_1$ | 0,911 | 0,153 | 0,000 |
| $y_2 \sim IDHM$ | Simplex | $\beta_0$ | 5,900 | 0,040 | 0,000 |
| | | $\beta_1$ | −9,491 | 0,056 | 0,000 |
| | | $\gamma_0$ | 0,093 | 0,067 | 0,000 |
| | | $\gamma_1$ | −0,340 | 0,095 | 0,000 |
| | Beta | $\beta_0$ | 5,952 | 0,043 | 0,000 |
| | | $\beta_1$ | −9,573 | 0,060 | 0,000 |
| | | $\gamma_0$ | 0,372 | 0,088 | 0,000 |
| | | $\gamma_1$ | −3,003 | 0,124 | 0,000 |

Proporção de pobres: [Simplex (AIC = −13470,63; BIC = −13444,18; $R^2$-ajustado = 0,8728), Beta (AIC = −13871,21; BIC = −13844,75; $R^2$-ajustado = 0,8922)]; Taxa de mortalidade infantil: [Simplex (AIC = −14034,27; BIC = −14007,81; $R^2$-ajustado = 0,8293), Beta (AIC = −14510,95; BIC = −14484,50; $R^2$-ajustado = 0,8526)].

Na Figura 3.2 é apresentado os gráficos dos resíduos quantílicos para os Modelos 1 e 2, evidenciando a não existência de indícios quanto a falta de qualidade de ajuste, ou seja, ambos os modelos ajustem bem os dados. Contudo, enquanto que pelos critérios AIC, BIC e Pseudo-$R^2$ o modelo de regressão Beta demostra um melhor ajustar os dados, podemos observar nos gráficos dos resíduos que os pontos estão bem mais alinhado ao longo da reta para o modelo de regressão Simplex. Com isso, tendo em vista a análise dos resíduos, é preferível a escolha do modelo de regressão Simplex ao modelo de regressão Beta como o que melhor ajusta os dados, ao menos para o caso do Modelo 1 em que a variável resposta tem aspecto bimodal.




**(a)**
*Simplex:* $y_1 \sim IDHM$

**(b)**
*Beta:* $y_1 \sim IDHM$

**(c)**
*Simplex:* $y_2 \sim IDHM$

**(d)**
*Beta:* $y_2 \sim IDHM$

**Figura 3.2:** *Gráfico de resíduos quantílicos - QQ-plot.*

# Capítulo 4

# Modelo de Regressão Simplex Multivariado via Cópulas

Neste capítulo será apresentado o conceito de função cópula, medidas de dependência, distribuição conjunta via cópulas. Iremos definir o modelo de regressão Simplex multivariado (MRSM) via cópula; finalmente, definiremos o modelo de regressão Simplex bivariado (MRSB).

## 4.1 Função de cópulas

A função de cópulas foi presentada pela primeira vez na literatura estatística por Sklar (1959), muito embora ideias e resultados similares possam ser encontrados em (Hoeffding, 1940). A função de cópula é uma dentre várias formas de se gerar distribuições multivariadas, tal técnica consiste em unir funções de distribuição marginais através de uma função de cópula, sendo um instrumento de grande utilidade quando as marginais são dadas ou conhecidas, com isso, possibilitando a representação de diversos tipos de dependências entre as variáveis. Uma função cópula $p$-dimensional é uma função $C : [0,1]^p \longrightarrow [0,1]$, que satisfaz as seguintes propriedades (Nadarajah *et al.*, 2017):

I) $C(u_1, ..., u_{i-1}, 0, u_{i+1}, ..., u_p) = 0$ para todo $1 \leq i \leq p$ e $0 \leq u_t \leq 1$, $t = 1, 2, ..., p$.

II) $C(1, ..., u, 1, ...., 1) = u$ para todo $0 < u < 1$ em cada um dos argumentos $p$. Ou seja, a cópula é igual a $u$ se um argumento é $u$ e todos os outros são iguais a 1.

III) Para qualquer $a_i$ e $b_i$ ordenados com $a_i \leq b_i$, $i = 1, 2, ..., p$, tem-se

$$\sum_{i_1=1}^{2} ... \sum_{i_p=1}^{2} (-1)^{i_1+...+i_p} C(u_{1,i_1}, ..., u_{p,i_1}) \geq 0,$$

onde $u_{j,1} = a_j$ e $u_{j,2} = b_j$ para $j = 1, 2, ..., p$. Logo, a cópula não é decrescente em $p$ dimensões.

**Definição 2** *(Anjos* et al., *2004) Uma cópula é definida como uma função de distribuição conjunta*

$$C(u_1, u_2, ..., u_p) = P(U_1 \leq u_1, U_2 \leq u_2, ..., U_p \leq u_p), \text{ para todo } 0 \leq u_i \leq 1,$$

*com* $U_i \sim U(0,1), i = 1, 2, ..., p$.

**Teorema 1** *(Nelsen, 2006) Seja* $B(\cdot)$ *uma função de distribuição acumulada p-dimensional com marginais* $F_1, ..., F_p$. *Então, existe uma cópula p-dimensional, digamos* $C$, *tal que,* $\forall\ (y_1, ..., y_p) \in R^p$,

$$B(y_1, ..., y_p) = C(F_1(y_1), ..., F_p(y_p)).$$

Será $C(\cdot)$ única, se e somente se, $F_1(\cdot), ..., F_p(\cdot)$ forem todas contínuas, caso contrário, podemos determinar $C(\cdot)$ por $Im(F_1(\cdot)) \times ... \times Im(F_p(\cdot))$, ou seja, considerando as imagens das $F_{p's}$.

Do Teorema 1, podemos encontrar a distribuição conjunta de $p$-variáveis aleatórias $y_1, ..., y_p$. Isto é, dado um conjunto de variáveis aleatórias contínuas com funções de distribuição marginais $F_1(\cdot), ..., F_p(\cdot)$ e função de distribuição $B(y_1, ..., y_p)$. Então, a função de densidade conjunta é dada por

$$\begin{aligned} b(y_1, ..., y_p) &= \frac{\partial^p B(y_1, ..., y_p)}{\partial y_1, ..., \partial y_p}, \\ &= \frac{\partial^p C(F_1(y_1), ..., F_p(y_p))}{\partial F_1(y_1), ..., \partial F_p(y_p)} \frac{\partial F_1(y_1)}{\partial y_1} \times ... \times \frac{\partial F_p(y_p)}{\partial y_p}, \\ &= c(F_1(y_1), ..., F_p(y_p)) \prod_{i=1}^{p} f_i(y_i), \end{aligned}$$

em que

$$c(F_1(y_1), ..., F_p(y_p)) = \frac{\partial^p C(F_1(y_1), ..., F_p(y_p))}{\partial F_1(y_1), ..., \partial F_p(y_p)} \quad \text{e} \quad f_i(y_i) = \frac{\partial F_i(y_i)}{\partial y_i}, \quad i = 1, 2, .., p.$$

### 4.1.1 Medidas de dependência

O grau e a forma de dependência entre variáveis estão associadas a um parâmetro que as associam. Como exemplo, consideremos a família de função Farlie-Gumbel-Morgenstern (FGM) que, no caso bivariado, tem a forma

$$C_\lambda(u, v) = uv + \lambda uv(1 - u)(1 - v),$$

em que $\lambda \in [-1, 1]$. Para cada $U$ e $V$ funções de distribuição acumulada, em que $U \sim U(0, 1)$ e $V \sim (0, 1)$ (Anjos *et al.*, 2004). E, conforme o Teorema 1, a função de distribuição é a cópula $C(\cdot)$. Logo, para a família FGM, a densidade de $U$ e $V$ é dada por

$$\frac{\partial^2 C(u, v)}{\partial u \partial v} = 1 + \lambda(1 - 2u)(1 - 2v).$$

Com $U \sim U(0, 1)$ e $V \sim (0, 1)$, temos que E(U) = E(V) = 1/2, Var(U) = Var(V) = 1/12 e

$$\mathrm{E(UV)} = \int_0^1 \int_0^1 \mathrm{uv}[1 + \lambda(1 - 2\mathrm{u})(1 - 2\mathrm{v})]\mathrm{dudv} = \frac{1}{4} + \lambda\frac{1}{36}.$$

Assim, a correlação para a família FGM é

$$Cor(U, V) = \frac{1}{3}\lambda,$$

em que $\lambda \in [-1, 1]$, e a correlação restrita ao intervalo $[-1/3, 1/3]$. Observa-se que, ao utilizar esta cópula para construir distribuições conjuntas, não importa quais sejam as distribuições marginais, a correlação entre as variáveis envolvidas sempre estará no intervalo $[-1/3, 1/3]$. Está cópula pode não ser adequada para situações em que a correlação é alta, em módulo. Contudo, levar em consideração o coeficiente de correlação linear pode nos levar a resultados equivocados, pois o mesmo não é o ideal na mensuração de associação entre variáveis de modelos via cópulas. Alternativamente, pode-se adotar a estatística $\tau$-Kendall, associada ao conceito de concordância e cuja medida está relacionada ao parâmetro de ligação $\lambda$ da cópula (Souza, 2011).

Dado pares de observações $(x, y)$ e $(x', y')$ de um certo vetor $(X, Y)$ serão ditos concordantes quando os pares $(x, x')(y, y') > 0$ e discordantes quando $(x, x')(y, y') < 0$. Calculando uma diferença entre probabilidade de concordância, cujo realização envolve a aplicação do Teorema 2.

**Teorema 2** *(Anjos et al., 2004) Sejam $(y_1, y_2)$ e $(y_1', y_2')$ vetores de variáveis aleatórias contínuas, con função de distribuição conjunta $B(y_1, y_2)$ e $B'(y_1', y_2')$, respectivamente. Sejam $C(\cdot)$ e $C(\cdot)'$ as cópulas de $(y_1, y_2)$ e $(y_1', y_2')$, respectivamente, de modo que $B(y_1, y_2) = C(F_1(y_1), F_2(y_2))$ e $B'(y_1', y_2') = C'(F_1(y_1'), F_2(y_2'))$. Seja $Q(\cdot)$ a diferença entre a probabilidade de concordância e a discordância de $(y_1, y_2)$ e $(y_1', y_2')$. Então,*

$$Q(C, C') = 4 \int_0^1 \int_0^1 C'(u, v) dC(u, v) - 1.$$

Sejam $(y_1, y_2)$ e $(y_1', y_2')$ vetores aleatórios independentes com uma função de distribuição conjunta $B(y_1, y_2)$, com marginais $F_1(\cdot)$ e $F_2(\cdot)$, e cópula $C$. Então, a medida $\tau(y_1, y_2)$ de Kendall é tal que

$$\tau(y_1, y_2) = P[(y_1 - y_1')(y_2 - y_2') > 0] - P[(y_1 - y_1')(y_2 - y_2') < 0].$$

Comprova-se, que o valor da estatística de $\tau$ depende unicamente da função de cópula utilizada e será o mesmo para qualquer distribuição marginal de $(y_1, y_2)$. A estatística de $\tau$ é uma medida de concordância ou associação e não de dependência, se $\tau = 0$ não significa falta de dependência, ou seja, independência. Uma outra medida de associação alternativa a de $\tau$-Kendall é a estatística de $\rho$ de Spearman (Souza, 2011, p.31).

### 4.1.2 Distribuição conjunta bivariada via cópulas

Consideremos que a estrutura de dependência é dada pela cópula de Farlie-Gumbel-Morgenstern (FGM), ou seja, dados $y_1$ e $y_2$, variáveis aleatórias com funções de distribuição $F_1$ e $F_2$, a distribuição conjunta de $F_1$ e $F_2$ é

$$B(y_1, y_2) = C(F_1(y_1), F_2(y_2)) = F_1(y_1)F_2(y_2) + \lambda F_1(y_1)F_2(y_2)[1 - F_1(y_1)][1 - F_2(y_2)].$$

Partindo da suposição de que $y_1$ e $y_2$ são variáveis aleatórias contínuas, sua função de densidade conjunta $b(\cdot)$ dar-se-á por

$$b(y_1, y_2) = \frac{\partial^2 B(y_1, y_2)}{\partial y_1 \partial y_2} = \frac{\partial}{\partial y_2}\left(\frac{\partial B(y_1, y_2)}{\partial y_1}\right) = f_1(y_1)f_2(y_2)\{1 + \lambda[1 - 2F_1(y_1)][1 - 2F_2(y_2)]\},$$

em que $f_1(\cdot)$ e $f_2(\cdot)$ são funções de densidade marginais de $y_1$ e $y_2$, respectivamente.

Na Tabela 4.1 é apresentado diferentes cópulas bivariada bastante conhecidas e utilizadas na prática, o domínio de variação do parâmetro de definição das cópulas $\lambda$, a medida de associação $\tau$ de Kendall e sua relação com $\lambda$.

**Tabela 4.1:** *Cópulas bivariadas, $\lambda$ e $\tau$ de Kendall.*

| Cópula | $C(u_1, u_2\|\lambda)$ | $\lambda$ | $\tau = h(\lambda)$ |
|---|---|---|---|
| Clayton | $(u_1^{-\lambda} + u_2^{-\lambda} - 1)^{-1/\lambda}$ | $(0, \infty)$ | $1 - \frac{2}{2+\lambda}$ |
| FGM | $u_1 u_2[1 + \lambda(1-u_1)(1-u_2)]$ | $[-1, 1]$ | $\frac{2\lambda}{9}$ |
| Frank | $-\frac{1}{\lambda}\ln\left(1 + \frac{(e^{-\lambda u_1}-1)(e^{-\lambda u_2}-1)}{e^{-\lambda}-1}\right)$ | $(-\infty, \infty) \setminus \{0\}$ | $1 - \frac{4}{\lambda}\left(1 - \frac{1}{\lambda}\int_0^\lambda \frac{t}{e^t-1}dt\right)$ |
| Gaussiana | $\int_{-\infty}^{\Phi^{-1}_{(u_1)}} \int_{-\infty}^{\Phi^{-1}_{(u_2)}} \frac{1}{2\pi\sqrt{1-\lambda^2}} \exp\left\{\frac{2\lambda st - s^2 - t^2}{2(1-\lambda^2)} dsdt\right\}$ | $[-1, 1]$ | $\frac{2}{\pi}\arcsin(\lambda)$ |

## 4.2 Modelo de regressão Simplex multivariada (MRSM) via cópulas

Dado $n$ observações independentes de $p$ variáveis aleatórias dependentes $(\boldsymbol{y_1}, \boldsymbol{y_2}, ..., \boldsymbol{y_p})$ e um conjunto de $k$ de variáveis regressoras $(\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_k})$. De modo geral, a estrutura de dependência para se obter a distribuição conjunta de $(\boldsymbol{y_1}, \boldsymbol{y_2}, ..., \boldsymbol{y_p})$ é dada por uma função cópula $C(\cdot)$ convencionalmente escolhida.

Matricialmente, o modelo de regressão Simplex multivariado é dado por

$$\begin{cases} \boldsymbol{Y} & \sim & S_p^-(\boldsymbol{\mu}, \boldsymbol{\sigma}) \\ g(\boldsymbol{\mu}) & = & \boldsymbol{\beta}_j^\top \boldsymbol{X}_i \\ h(\boldsymbol{\sigma}^2) & = & \boldsymbol{\gamma}_j^\top \boldsymbol{Z}_i \end{cases} , \quad \forall i = 1, 2, ..., n \quad \text{e} \quad j = 1, 2, ..., p,$$

de modo que

$$\begin{cases} \boldsymbol{Y} & = & (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_p)^\top & \rightarrow & \boldsymbol{y}_i & = & (\boldsymbol{y}_{1i}, \boldsymbol{y}_{2i}, ..., \boldsymbol{y}_{pi})^\top \\ \boldsymbol{\mu} & = & (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_p)^\top & \rightarrow & \boldsymbol{\mu}_i & = & (\boldsymbol{\mu}_{1i}, \boldsymbol{\mu}_{2i}, ..., \boldsymbol{\mu}_{pi})^\top \\ \boldsymbol{\sigma}^2 & = & (\boldsymbol{\sigma}_1^2, \boldsymbol{\sigma}_2^2, ..., \boldsymbol{\sigma}_p^2)^\top & \rightarrow & \boldsymbol{\sigma}_i^2 & = & (\boldsymbol{\sigma}_{1i}^2, \boldsymbol{\sigma}_{2i}^2, ..., \boldsymbol{\sigma}_{pi}^2)^\top \end{cases},$$

para todo $i = 1, 2, ..., n$. Em que

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_p \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{p1} \\ y_{21} & y_{22} & \cdots & y_{p2} \\ \cdots & \cdots & \cdots & \cdots \\ y_{1n} & y_{2n} & \cdots & y_{pn} \end{bmatrix}_{[p \times n]},$$

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_p \end{bmatrix} = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{p1} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{p2} \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{1n} & \mu_{2n} & \cdots & \mu_{pn} \end{bmatrix}_{[p \times n]},$$

$$\boldsymbol{\sigma}^2 = \begin{bmatrix} \boldsymbol{\sigma}_1^2 \\ \boldsymbol{\sigma}_2^2 \\ \vdots \\ \boldsymbol{\sigma}_p^2 \end{bmatrix} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{p1}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{p2}^2 \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{1n}^2 & \sigma_{2n}^2 & \cdots & \sigma_{pn}^2 \end{bmatrix}_{[p \times n]}.$$

Para cada variável resposta, o modelo de regressão Simplex multivariado é dado por

$$\begin{cases} g(\mu_{ji}) & = & \beta_{oj} X_{i0} + \beta_{1j} X_{i1} + \beta_{2j} X_{i2} + ... + \beta_{k_1 j} X_{ik_1} & = & \sum_{l=0}^{k_1} X_{il} \beta_{lj} \\ h(\sigma_{ji}^2) & = & \gamma_{oj} Z_{i0} + \gamma_{1j} Z_{i1} + \gamma_{2j} Z_{i2} + ... + \gamma_{k_1 j} Z_{ik_2} & = & \sum_{l=0}^{k_2} Z_{il} \gamma_{lj} \end{cases}.$$

em que $\boldsymbol{\beta}_j = (\beta_{oj}, \beta_{1j}, \beta_{2j}, ..., \beta_{k_1 p})^\top$, $\boldsymbol{\gamma}_j = (\gamma_{oj}, \gamma_{1j}, \gamma_{2j}, ..., \gamma_{k_2 p})^\top$, $\boldsymbol{X}_i = (X_{i0}, X_{i1}, X_{i2}, ..., X_{nk_1})^\top$, $\boldsymbol{Z}_i = (Z_{i0}, Z_{i1}, Z_{i2}, ..., Z_{nk_2})^\top$, onde

$$
\begin{aligned}
g(\boldsymbol{\mu}) &= \begin{bmatrix} g(\boldsymbol{\mu}_1) \\ g(\boldsymbol{\mu}_2) \\ \vdots \\ g(\boldsymbol{\mu}_p) \end{bmatrix} = \begin{bmatrix} g(\mu_{11}, \mu_{21}, ..., \mu_{p1}) \\ g(\mu_{12}, \mu_{22}, ..., \mu_{p2}) \\ \cdots \\ g(\mu_{1n}, \mu_{2n}, ..., \mu_{pn}) \end{bmatrix}_{[p\times n]}, \\
h(\boldsymbol{\sigma}^2) &= \begin{bmatrix} h(\boldsymbol{\sigma}_1^2) \\ h(\boldsymbol{\sigma}_2^2) \\ \vdots \\ h(\boldsymbol{\sigma}_p^2) \end{bmatrix} = \begin{bmatrix} h(\sigma_{11}^2, \sigma_{21}^2, ..., \sigma_{p1}^2) \\ h(\sigma_{12}^2, \sigma_{22}^2, ..., \sigma_{p2}^2) \\ \cdots \\ h(\sigma_{1n}^2, \sigma_{2n}^2, ..., \sigma_{pn}^2) \end{bmatrix}_{[p\times n]}, \\
\boldsymbol{\beta}^\top \boldsymbol{X} &= [\beta_{0j}, \beta_{1j}, \beta_{2j}, ..., \beta_{k_1 p}] \begin{bmatrix} X_{i0} \\ X_{i1} \\ X_{i2} \\ \vdots \\ X_{nk_1} \end{bmatrix}, \\
\boldsymbol{\gamma}^\top \boldsymbol{Z} &= [\gamma_{0j}, \gamma_{1j}, \gamma_{2j}, ..., \gamma_{k_2 p}] \begin{bmatrix} Z_{i0} \\ Z_{i1} \\ Z_{i2} \\ \vdots \\ Z_{nk_2} \end{bmatrix},
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{X} &= \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_n \end{bmatrix} = \begin{bmatrix} X_{10} & X_{11} & \cdots & X_{1k_1} \\ X_{20} & X_{21} & \cdots & X_{2k_1} \\ \cdots & \cdots & \cdots & \cdots \\ X_{n0} & X_{n1} & \cdots & X_{nk_1} \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k_1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k_1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk_1} \end{bmatrix}_{[n\times k_1]}, \\
\boldsymbol{Z} &= \begin{bmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \\ \vdots \\ \boldsymbol{Z}_n \end{bmatrix} = \begin{bmatrix} Z_{10} & Z_{11} & \cdots & Z_{1k_2} \\ Z_{20} & Z_{21} & \cdots & Z_{2k_2} \\ \cdots & \cdots & \cdots & \cdots \\ Z_{n0} & Z_{n1} & \cdots & Z_{nk_2} \end{bmatrix} = \begin{bmatrix} 1 & Z_{11} & Z_{12} & \cdots & Z_{1k_2} \\ 1 & Z_{21} & Z_{22} & \cdots & Z_{2k_2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & Z_{n1} & Z_{n2} & \cdots & Z_{nk_2} \end{bmatrix}_{[n\times k_2]}.
\end{aligned}
$$

Dado $n$ observações independentes de $p$ variáveis dependentes $(\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_p)$ e um conjunto de $k$ variáveis regressoras $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k)$ , em que $\boldsymbol{y} \sim S_p^-(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. A partir do Teorema 1, temos que a função de verossimilhança é definida como

$$
L(\boldsymbol{\theta}) = \prod_{j=1}^{p} \prod_{i=1}^{n} c(F_1(\boldsymbol{y}_{1i}), F_2(\boldsymbol{y}_{2i}), ..., F_p(\boldsymbol{y}_{pn})) f(\boldsymbol{y}_{ji}), \tag{4.1}
$$

em que $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})^\top$, $\boldsymbol{\lambda}$ o parâmetro que interliga as variáveis aleatórias de dependentes através da função cópula e

$$
c(F_1(\boldsymbol{y}_{1i}), ..., F_p(\boldsymbol{y}_{pn})) = \frac{\partial^k C(F_1(\boldsymbol{y}_{1i}), ..., F_p(\boldsymbol{y}_{pn}))}{\partial F_1(\boldsymbol{y}_{1i}), ..., \partial F_p(\boldsymbol{y}_{pn})} \quad \text{e} \quad f(\boldsymbol{y}_{ji}) = \frac{\partial F_j(\boldsymbol{y}_{ji})}{\partial \boldsymbol{y}_{ji}}.
$$

Sem perda de generalidade, na Seção 4.2.1 é apresentado a estrutura do modelo de regressão Simplex bivariado (MRSB) via cópulas, um caso particular do (MRSM) vía cópulas.

### 4.2.1 Modelo de regressão Simplex bivariada (MRSB) via cópulas

Dado o par de variáveis aleatórias dependentes $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ e um conjunto $k$ de variáveis regressoras $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k)$. A estrutura de dependência para se obter a distribuição conjunta de $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ é dada por uma função cópula $C(F(\boldsymbol{y}_1), F(\boldsymbol{y}_2))$ convencionalmente escolhida.

Matricialmente, o modelo de regressão Simplex bivariado é dado por

$$\begin{cases} \boldsymbol{Y} & \sim & S_2^-(\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \\ g(\boldsymbol{\mu}) & = & \boldsymbol{\beta}_j \boldsymbol{X}_i \\ h(\boldsymbol{\sigma}^2) & = & \boldsymbol{\gamma}_j \boldsymbol{Z}_i \end{cases}, \quad \forall i = 1, 2, ..., n \quad \text{e} \quad j = 1, 2,$$

de modo que

$$\begin{cases} \boldsymbol{Y} & = & (\boldsymbol{y}_1, \boldsymbol{y}_2)^\top & \to & \boldsymbol{y}_i & = & (\boldsymbol{y}_{1i}, \boldsymbol{y}_{2i})^\top, \\ \boldsymbol{\mu} & = & (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)^\top & \to & \boldsymbol{\mu}_i & = & (\boldsymbol{\mu}_{1i}, \boldsymbol{\mu}_{2i})^\top \\ \boldsymbol{\sigma}^2 & = & (\boldsymbol{\sigma}_1^2, \boldsymbol{\sigma}_2^2)^\top & \to & \boldsymbol{\sigma}_i^2 & = & (\boldsymbol{\sigma}_{1i}^2, \boldsymbol{\sigma}_{2i}^2)^\top, \end{cases},$$

para todo $i = 1, 2, ..., n$, onde

$$\begin{aligned} \boldsymbol{Y} &= \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \end{bmatrix}_{[2\times n]}, \\ \boldsymbol{\mu} &= \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1n} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2n} \end{bmatrix}_{[2\times n]}, \\ \boldsymbol{\sigma}^2 &= \begin{bmatrix} \boldsymbol{\sigma}_1^2 \\ \boldsymbol{\sigma}_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2n}^2 \end{bmatrix}_{[2\times n]}. \end{aligned}$$

Para cada variável resposta, o modelo de regressão simplex bivariado é dado por

$$\begin{cases} g(\mu_{ji}) & = & \beta_{oj} X_{i0} + \beta_{1j} X_{i1} + \beta_{2j} X_{i2} + ... + \beta_{k_1 j} X_{ik_1} & = & \sum_{l=0}^{k_1} X_{il} \beta_{lj}, \\ h(\sigma_{ji}^2) & = & \gamma_{oj} Z_{i0} + \gamma_{1j} Z_{i1} + \gamma_{2j} Z_{i2} + ... + \gamma_{k_1 j} Z_{ik_2} & = & \sum_{l=0}^{k_2} Z_{il} \gamma_{lj}. \end{cases} \tag{4.2}$$

em que $\boldsymbol{\beta_j} = (\beta_{oj}, \beta_{1j}, \beta_{2j}, ..., \beta_{k_1 p})^\top$, $\boldsymbol{\gamma_j} = (\gamma_{oj}, \gamma_{1j}, \gamma_{2j}, ..., \gamma_{k_2 p})^\top$, $\boldsymbol{X_i} = (X_{i0}, X_{i1}, X_{i2}, ..., X_{nk_1})^\top$, $\boldsymbol{Z_i} = (Z_{i0}, Z_{i1}, Z_{i2}, ..., Z_{nk_2})^\top$, sendo

$$\begin{aligned} g(\boldsymbol{\mu}) &= \begin{bmatrix} g(\boldsymbol{\mu}_1) \\ g(\boldsymbol{\mu}_2) \end{bmatrix} = \begin{bmatrix} g(\mu_{11}, \mu_{21}, ..., \mu_{1n}) \\ g(\mu_{12}, \mu_{22}, ..., \mu_{2n}) \end{bmatrix}_{[2\times n]}, \\ h(\boldsymbol{\sigma}^2) &= \begin{bmatrix} h(\boldsymbol{\sigma_1^2}) \\ h(\boldsymbol{\sigma_2^2}) \end{bmatrix} = \begin{bmatrix} h(\sigma_{11}^2, \sigma_{21}^2, ..., \sigma_{1n}^2) \\ h(\sigma_{12}^2, \sigma_{22}^2, ..., \sigma_{2n}^2) \end{bmatrix}_{[2\times n]}, \\ \boldsymbol{\beta}^\top \boldsymbol{X} &= [\beta_{0j}, \beta_{1j}, \beta_{2j}, ..., \beta_{k_1 p}] \begin{bmatrix} X_{i0} \\ X_{i1} \\ X_{i2} \\ \vdots \\ X_{nk_1} \end{bmatrix}, \\ \boldsymbol{\gamma}^\top \boldsymbol{Z} &= [\gamma_{0j}, \gamma_{1j}, \gamma_{2j}, ..., \gamma_{k_2 p}] \begin{bmatrix} Z_{i0} \\ Z_{i1} \\ Z_{i2} \\ \vdots \\ Z_{nk_2} \end{bmatrix}, \end{aligned}$$

$$\begin{aligned}
\boldsymbol{X} &= \begin{bmatrix}\boldsymbol{X_1}\\ \boldsymbol{X_2}\end{bmatrix} = \begin{bmatrix}X_{10} & X_{11} & \cdots & X_{1k_1}\\ X_{20} & X_{21} & \cdots & X_{2k_1}\end{bmatrix} = \begin{bmatrix}1 & X_{11} & X_{12} & \cdots & X_{1k_1}\\ 1 & X_{21} & X_{22} & \cdots & X_{2k_1}\end{bmatrix}_{[n\times k_1]},\\
\boldsymbol{Z} &= \begin{bmatrix}\boldsymbol{Z}_1\\ \boldsymbol{Z}_2\end{bmatrix} = \begin{bmatrix}Z_{10} & Z_{11} & \cdots & Z_{1k_2}\\ Z_{20} & Z_{21} & \cdots & Z_{2k_2}\end{bmatrix} = \begin{bmatrix}1 & Z_{11} & Z_{12} & \cdots & Z_{1k_2}\\ 1 & Z_{21} & Z_{22} & \cdots & Z_{2k_2}\end{bmatrix}_{[n\times k_2]}.
\end{aligned}$$

Dado $n$ observações independentes do par de variáveis aleatórias dependentes $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ e um conjunto de $k$ variáveis regressoras $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k)$ , tal que $\boldsymbol{y}_1 \sim S^-(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$ e $\boldsymbol{y}_2 \sim S^-(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2)$. A partir do Teorema 1, temos que a função de verossimilhança é definida como

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{y}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} f(\boldsymbol{y}_{1i}, \boldsymbol{y}_{2i}; \boldsymbol{\beta}, \boldsymbol{\gamma}),$$

em que $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})^\top$.

A função de máxima verossimilhança é definida como

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} c(F_1(\boldsymbol{y}_{1i}), F_2(\boldsymbol{y}_{2i})) f(\boldsymbol{y}_{1i}) f(\boldsymbol{y}_{2i}) \tag{4.3}$$

em que $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda)^\top$, $\lambda$ o parâmetro que interliga as variáveis aleatórias dependentes através da função cópula e

$$c(F_1(\boldsymbol{y}_{1i}), F_2(\boldsymbol{y}_{2i})) = \frac{\partial^2 C(F_1(\boldsymbol{y}_{1i}), F_2(\boldsymbol{y}_{2i}))}{\partial F_1(\boldsymbol{y}_{1i}), \partial F_2(\boldsymbol{y}_{2i})},$$

de modo que,

$$\begin{aligned}
\boldsymbol{\beta} &= \begin{bmatrix}\boldsymbol{\beta}_0\\ \boldsymbol{\beta}_1\end{bmatrix} = \begin{bmatrix}\beta_{01} & \beta_{02} & \cdots & \beta_{0p}\\ \beta_{11} & \beta_{12} & \cdots & \beta_{1p}\end{bmatrix}_{[k_1\times p]},\\
\boldsymbol{\gamma} &= \begin{bmatrix}\boldsymbol{\gamma}_0\\ \boldsymbol{\gamma}_1\end{bmatrix} = \begin{bmatrix}\gamma_{01} & \gamma_{02} & \cdots & \gamma_{0p}\\ \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p}\end{bmatrix}_{[k_2\times p]}.
\end{aligned}$$

Nas Subseções 4.2.2, 4.2.3 e 4.2.4 são apresentados as estruturas do modelo de regressão Simplex bivariado para cópulas FGM, Clayton e Frank, como casos particulares do modelo de regressão Simplex multivariado.

### 4.2.2 MRSB via cópula FGM

Seja $\boldsymbol{y}_1 \sim S^-(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$ e $\boldsymbol{y}_2 \sim S^-(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2)$ variáveis aleatórias com função de densidade dada em (2.2). Para $n$ observações independentes de variáveis aleatórias dependentes $y_{1i}, y_{2i})$, $i = 1, 2, ..., n$, a função de verossimilhança da distribuição conjunta entre $\boldsymbol{y}_1$ e $\boldsymbol{y}_2$, é dada por

$$\begin{aligned}
L(\boldsymbol{\theta}) = &\prod_{i=1}^{n} \{2\pi\sigma_{1i}^2[y_{1i}(1-y_{1i})]^3\}^{-1/2} \exp\left\{-\frac{1}{2\sigma_{1i}^2} d(y_{1i}; \mu_{1i})\right\} \times\\
&\{2\pi\sigma_{2i}^2[y_{2i}(1-y_{2i})]^3\}^{-1/2} \exp\left\{-\frac{1}{2\sigma_{2i}^2} d(y_{2i}; \mu_{2i})\right\} \times\\
&\{1 + \lambda[1 - 2F_1(y_{1i})][1 - 2F_2(y_{2i})]\},
\end{aligned} \tag{4.4}$$

em que $F_1(\cdot)$ e $F_2(\cdot)$ são funções de distribuições acumuladas de $\boldsymbol{y}_1$ e $\boldsymbol{y}_2$, e respectivamente $\boldsymbol{\theta} = (\underset{\sim}{\boldsymbol{\beta}}, \underset{\sim}{\boldsymbol{\gamma}}, \lambda)^\top$.

A expressão (4.4) um caso particular do modelo Simplex multivariado. O logaritmo da função de verossimilhança do modelo de regressão Simplex bivariada é

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2, \lambda),$$

onde

$$\begin{aligned}\ell_i(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2, \lambda) &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_{1i}^2) - \frac{3}{2}\log[y_{1i}(1-y_{1i})] - \frac{1}{2\sigma_{1i}^2}d(y_{1i};\mu_{1i}) - \\ &\quad \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_{2i}^2) - \frac{3}{2}\log[y_{2i}(1-y_{2i})] - \frac{1}{2\sigma_{2i}^2}d(y_{2i};\mu_{2i}) + \\ &\quad \log\{1+\lambda[1-2F_1(y_{1i})][1-2F_2(y_{2i})]\}, \end{aligned} \tag{4.5}$$

em que $g(\mu_{ij}) = \sum_{l=0}^{k_1} x_{il}\boldsymbol{\beta}_{lj}$, $h(\sigma_{ij}^2) = \sum_{l=0}^{k2} Z_{il}\gamma_{lj}$, $i = 1, ..., n$ e $j = 1, 2$.

### 4.2.3 MRSB via cópula Clayton

Seja $\boldsymbol{y}_1 \sim S^-(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$ e $\boldsymbol{y}_2 \sim S^-(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2)$ variáveis aleatórias com função de densidade dada em 2.2. Para $n$ observações independentes de variáveis aleatórias dependentes $(\boldsymbol{y}_{1i}, \boldsymbol{y}_{2i})$, $i = 1, 2, ..., n$, a função de verossimilhança da distribuição conjunta entre $\boldsymbol{y}_1$ e $\boldsymbol{y}_2$, é dada por

$$\begin{aligned}L(\boldsymbol{\theta}) &= \prod_{i=1}^{n}\{2\pi\sigma_{1i}^2[y_{1i}(1-y_{1i})]^3\}^{-1/2}\exp\left\{-\frac{1}{2\sigma_{1i}^2}d(y_{1i};\mu_{1i})\right\} \times \\ &\quad \{2\pi\sigma_{2i}^2[y_{2i}(1-y_{2i})]^3\}^{-1/2}\exp\left\{-\frac{1}{2\sigma_{2i}^2}d(y_{2i};\mu_{2i})\right\} \times \\ &\quad (1+\lambda)F_1(y_{1i})^{-1-\lambda}F_2(y_{2i})^{-1-\lambda}(F_1(y_{1i})^{-\lambda} + F_2(y_{2i})^{-\lambda})^{-2-1/\lambda}, \end{aligned} \tag{4.6}$$

em que $F_1(\cdot)$ e $F_2(\cdot)$ são funções de distribuições acumuladas de $\boldsymbol{y}_1$ e $\boldsymbol{y}_2$ e $\boldsymbol{\theta} = (\underset{\sim}{\boldsymbol{\beta}}, \underset{\sim}{\boldsymbol{\gamma}}, \lambda)^\top$.

A expressão (4.6) é um caso particular do modelo Simplex multivariado. O logaritmo da função de verossimilhança do modelo de regressão Simplex bivariada é

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2, \lambda),$$

onde

$$\begin{aligned}\ell_i(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2, \lambda) &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_{1i}^2) - \frac{3}{2}\log[y_{1i}(1-y_{1i})] - \frac{1}{2\sigma_{1i}^2}d(y_{1i};\mu_{1i}) - \\ &\quad \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_{2i}^2) - \frac{3}{2}\log[y_{2i}(1-y_{2i})] - \frac{1}{2\sigma_{2i}^2}d(y_{2i};\mu_{2i}) + \\ &\quad \log(\lambda+1) - (2+\lambda^{-1})\log(F_1(y_{1i})^{-\lambda} + F_2(y_{2i})^{-\lambda}) - \\ &\quad (\lambda+1)\log(F_1(y_{1i})F_2(y_{2i})) \end{aligned} \tag{4.7}$$

em que $g(\mu_{ij}) = \sum_{l=0}^{k_1} x_{il}\boldsymbol{\beta}_{lj}$, $h(\sigma_{ij}^2) = \sum_{l=0}^{k2} Z_{il}\gamma_{lj}$, $i = 1, ..., n$ e $j = 1, 2$.

### 4.2.4 MRSB via cópula Frank

Seja $\boldsymbol{y}_1 \sim S^-(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$ e $\boldsymbol{y}_2 \sim S^-(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2)$ variáveis aleatórias com função de densidade dada em 2.2. Para $n$ observações independentes de variáveis aleatórias dependentes $(y_{1i}, y_{2i})$, $i = 1, 2, ..., n$,

a função de verossimilhança da distribuição conjunta entre $\boldsymbol{y}_1$ e $\boldsymbol{y}_2$, é dada por

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n}\{2\pi\sigma_{1i}^2[y_{1i}(1-y_{1i})]^3\}^{-1/2}\exp\left\{-\frac{1}{2\sigma_{1i}^2}d(y_{1i};\mu_{1i})\right\}\times \\ &\quad \{2\pi\sigma_{2i}^2[y_{2i}(1-y_{2i})]^3\}^{-1/2}\exp\left\{-\frac{1}{2\sigma_{2i}^2}d(y_{2i};\mu_{2i})\right\}\times \\ &\quad \frac{\lambda\exp[\lambda(1+F_1(y_{1i})+F_2(y_{2i}))][-1+\exp(\lambda)]}{\{\exp(\lambda)-\exp[\lambda(1+F_1(y_{1i}))]-\exp[\lambda(1+F_2(y_{2i}))]+\exp[\lambda(F_1(y_{1i})+F_2(y_{2i}))]\}^2}, \end{aligned} \tag{4.8}$$

em que $F_1(\cdot)$ e $F_2(\cdot)$ são funções de distribuições acumuladas de $\boldsymbol{y}_1$ e $\boldsymbol{y}_2$ e $\boldsymbol{\theta} = (\underset{\sim}{\beta}, \underset{\sim}{\gamma}, \lambda)^\top$.

A expressão (4.8) é um caso particular do modelo simplex multivariado. O logaritmo da função de verossimilhança do modelo de regressão Simplex bivariada é

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n}\ell(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2, \lambda),$$

onde

$$\begin{aligned} \ell_i(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2, \lambda) &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_{1i}^2) - \frac{3}{2}\log[y_{1i}(1-y_{1i})] - \frac{1}{2\sigma_{1i}^2}d(y_{1i};\mu_{1i}) - \\ &\quad \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_{2i}^2) - \frac{3}{2}\log[y_{2i}(1-y_{2i})] - \frac{1}{2\sigma_{2i}^2}d(y_{2i};\mu_{2i}) + \\ &\quad \log(\lambda) + \log(1-e^{-\lambda}) - \lambda(F_1(y_{1i}) + F_2(y_{2i})) - \\ &\quad 2\log(e^{-\lambda}-1) + (e^{-\lambda F_1(y_{1i})}-1)(e^{-\lambda F_2(y_{2i})}-1). \end{aligned} \tag{4.9}$$

em que $g(\mu_{ij}) = \sum_{l=0}^{k_1} x_{il}\boldsymbol{\beta}_{lj}$, $h(\sigma_{ij}^2) = \sum_{l=0}^{k2} Z_{il}\gamma_{lj}$, $i = 1, ..., n$ e $j = 1, 2$.

Para as seções 4.2.2 - 4.2.4, derivando parcialmente o logaritmo das funções de verossimilhança (4.5, 4.7 e 4.9), com respeito a vetor de parâmetros, obtemos os vetores escores $U(\boldsymbol{\theta}) = (U_{\underset{\sim}{\beta}}, U_{\underset{\sim}{\gamma}}, U_\lambda)^\top$, matriz de informação de Fisher observada, $K(\boldsymbol{\theta}) = \partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$, que sob condições de regularidade, o estimador de máxima verossimilhança $\widehat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$, aproxima-se a uma distribuição Normal com média zero e matriz de variâncias e covariâncias $K^{-1}(\boldsymbol{\theta})$.

## 4.3 Estudo de simulação

Nesta seção conduziremos um estudo de simulação via Monte Carlo (MC), com o intuito de estudar o comportamento assintótico dos estimadores de máxima verossimilhança do modelo de regressão Simplex bivariado considerando as cópulas FGM, Clayton e Frank.

Números aleatórios são encontrados considerando a inversão da função de distribuição condicional de $U_2 = u_2$ dado $U_1 = u_1$, isto é:

$$\begin{aligned} C_{2|1}(u_2|u_1) &= P(U_2 \leq u_2 | U_1 = u_1) \\ &= \lim_{\Delta u_1 \to \infty} \frac{C(u_1+\Delta u_1, u_2) - C(u_1, u_2)}{\Delta u_1} = \frac{\partial C(u_1, u_2)}{\partial u_1}. \end{aligned}$$

- Cópula FGM

A distribuição condicional na cópula FGM resulta em

$$v = C_{2|1}(u_2) = \frac{\partial C(u_1, u_2)}{\partial u_1} = u_2[1 + \lambda(1 - u_2)(1 - u_1)].$$

A equação $v = C_{2|1}(u_2)$ tem uma única raiz $u_2$ para $0 < v < 1$, que é o nosso caso de interesse. Então, conforme descrito Johnson (1987), é pertinente utilizar o seguinte procedimento para simular o par $u_1, u_2$ da cópula FGM:

1. gera-se $u_1$ e $v$ independentes da distribuição $U(0, 1)$;
2. encontra-se $A = \lambda(2u_1 - 1) - 1$ e $B = [1 - \lambda(2u_1 - 1)]^2 + 4v\lambda(2u_1 - 1)$;
3. $u_2 = 2v/(\sqrt{B} - A)$;
4. após a obtenção dos pares $(u_1, u_2)$, mediante os passos supracitados, as variáveis $\boldsymbol{y}_1$ e $\boldsymbol{y}_2$ são encontradas por

$$\boldsymbol{y}_{1i} = F_2^{-1}(u_{1i}|\boldsymbol{\mu}_{1i}, \boldsymbol{\sigma}_{1i}^2) \quad \text{e} \quad \boldsymbol{y}_{2i} = F_1^{-1}(u_{2i}|\boldsymbol{\mu}_{2i}^2, \boldsymbol{\sigma}_{2i}^2)$$

onde $F_j^{-1}(\cdot)$ é a inversa da função de distribuição acumulada da distribuição Simplex univariada, para $j = 1, 2$.

- Cópula Clayton

A distribuição condicional na cópula Clayton resulta em

$$v = C_{2|1}(u_2) = \frac{\partial C(u_1, v_2)}{\partial u_1} = u_1^{-(1+\lambda)} \left(u_1^{-\lambda} + u_2^{-\lambda} - 1\right)^{-1-1/\lambda},$$

sendo que

$$u_2 = \left(1 - u_1^{-\lambda} + \left[v u_1^{1+\lambda}\right]^{-\frac{\lambda}{1+\lambda}}\right). \tag{4.10}$$

Considera-se o procedimento abaixo para gerar o par $(u_1, u_2)$ de observações desta cópula:

1. gerar $u_1$ e $v$ independente da distribuição $U(0, 1)$;
2. obter $u_2$ a partir da equação acima;
3. após a obtenção dos pares $(u_1, u_2)$, mediante os passo supracitados, as variáveis $\boldsymbol{y}_1$ e $\boldsymbol{y}_2$ será dada por

$$\boldsymbol{y}_{1i} = F_2^{-1}(u_{1i}|\boldsymbol{\mu}_{1i}, \boldsymbol{\sigma}_{1i}^2) \quad \text{e} \quad \boldsymbol{y}_{2i} = F_1^{-1}(u_{2i}|\boldsymbol{\mu}_{2i}, \boldsymbol{\sigma}_{2i}^2)$$

onde $F_j^{-1}(\cdot)$ é a inversa da função de distribuição acumulada da distribuição Simplex, para $j = 1, 2$.

- Cópula Frank

A distribuição condicional na cópula Frank resulta em

$$\begin{aligned} v &= C_{2|1}(u_2) = \frac{\partial C(u_1, u_2)}{\partial u_1} = \frac{e^{-\lambda u_1}(e^{-\lambda u_2} - 1)}{(e^{-\lambda} - 1)(e^{-\lambda u_1} - 1)(e^{-\lambda u_2} - 1)} \\ u_2 &= C_{2|1}^{-1}(v|u_1) = -\frac{1}{\lambda} \log\left(1 + \frac{v(e^{-\lambda} - 1)}{v + (1 - v)e^{-\lambda u_1}}\right), \end{aligned}$$

onde $C_{2|1}^{-1}$ é a função inversa de $C_{2|1}$.

Considere o procedimento abaixo para gerar o par $(u_1, u_2)$ de observações desta cópula:

1. gerar $u_1$ e $v$ independente da distribuição $U(0,1)$;
2. encontrar $u_2 = C^{-1}_{2|1}(v|u1)$;
3. após a obtenção dos pares $(u_1, u_2)$, mediante os passo supracitados, as variáveis $\boldsymbol{y}_1$ e $\boldsymbol{y}_2$ será dada por

$$\boldsymbol{y}_{1i} = F_2^{-1}(u_{1i}|\boldsymbol{\mu}_{1i}, \boldsymbol{\sigma}^2_{1i}) \quad \text{e} \quad \boldsymbol{y}_{2i} = F_1^{-1}(u_{2i}|\boldsymbol{\mu}_{2i}, \boldsymbol{\sigma}^2_{2i})$$

onde $F_j^{-1}(\cdot)$ é a inversa da função de distribuição acumulada da distribuição Simplex, para $j = 1, 2$.

Neste estudo de simulação de Monte Carlo, considera-se 1.000 réplicas de Monte Carlo (MC) para amostas de tamanhos {25, 50, 75 e 100}. Para cada amostra gerada são obtidas as estimativas de máxima verossimilhança dos parâmetros $\boldsymbol{\beta}$'s, $\boldsymbol{\gamma}$'s e $\lambda$ cujas performances foram calculadas por meio da média em cada réplica dos $\widehat{\beta}$'s, $\widehat{\gamma}$'s, $\widehat{\lambda}$, do viés, da raiz do erro quadrático médio (REQM) e Taxa de Cobertura (TC) de 95% de confiança.

Suponha $R$ o número de réplicas de Monte Carlo: a média, o viés e REQM são encontrados a partir de

$$\mathrm{E}(\widehat{\theta}_{\mathrm{t}}) = \frac{1}{\mathrm{R}}\sum_{\mathrm{i}=1}^{\mathrm{R}}\widehat{\theta}_{\mathrm{t}} \qquad \mathrm{Viés}(\widehat{\theta}_{\mathrm{t}}) = \frac{1}{\mathrm{R}}\sum_{\mathrm{i}=1}^{\mathrm{R}}(\widehat{\theta}_{\mathrm{ti}} - \widehat{\theta}_{\mathrm{t}}) \qquad \mathrm{EQM}(\widehat{\theta}_{\mathrm{t}}) = \sqrt{\frac{1}{\mathrm{R}}\sum_{\mathrm{i}=1}^{\mathrm{R}}(\widehat{\theta}_{\mathrm{ti}} - \widehat{\theta}_{\mathrm{t}})^2}$$

em que $\widehat{\theta}_t$ é a estimativa de $\theta$ obtida a partir da amostra gerada na réplica $t$ do processo de simulação de Monte Carlo (MC), para $t = 1, ..., 9$. A Taxa de Cobertura (TC) para o intervalos de confiança de 95% é estimada da seguinte forma

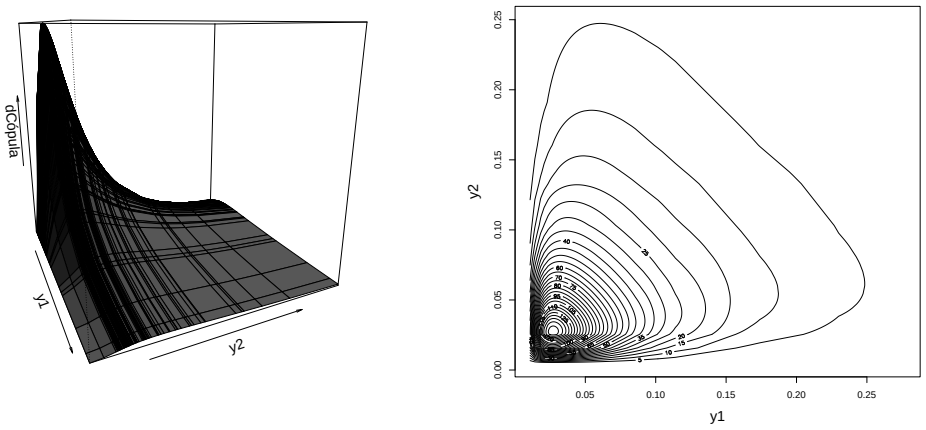
$$\mathrm{TC} = \frac{\widehat{\boldsymbol{\theta}}_t \in \mathrm{IC}[\boldsymbol{\theta}_t; 1-\alpha]}{R}$$

em que $\widehat{\boldsymbol{\theta}}_t$ é a $t$-ésima estimativa de $\boldsymbol{\theta}$, $\alpha$ é o nível de significância e $\mathrm{IC}[\boldsymbol{\theta}; 1-\alpha]$ é o intervalo de confiança para $\boldsymbol{\theta}$.

Consideramos 3 cenários com uma única covariável $x$ que segue uma distribuição uniforme no intervalo (0,1), sendo a escolha dos parâmetros de modo que os dados fiquem concentrados em torno de zero, meio e um, respectivamente, que a seguir são apresentados de forma detalhada.

### 4.3.1 Cenário 1

Neste cenário é consideramos o vetor de parâmetros reais $\boldsymbol{\theta} = (\beta_{01} = -3,5; \beta_{11} = 1,2; \beta_{02} = -3,5; \beta_{12} = 1,2; \gamma_{01} = -0,8; \gamma_{11} = 1,6; \gamma_{02} = -0,8; \gamma_{12} = 1,6; \lambda = 0,5)^\top$ que é concretizado pelo gráfico em 4.1 (a) corresponde ao gráfico de superfícies próximo a 0; uma representação das curvas de nível pode ser vista em 4.1 (b).




**(a)** **(b)**

**Figura 4.1:** *Gráficos de superfície e curvas de nível para uma amostra de tamanho 100 e* $\boldsymbol{\theta} = (\beta_{01} = -3,5; \beta_{11} = 1,2; \beta_{02} = -3,5; \beta_{12} = 1,2; \gamma_{01} = -0,8; \gamma_{11} = 1,6; \gamma_{02} = -0,8; \gamma_{12} = 1,6; \lambda = 0,5)^\top$.

Os resultados encontrados neste cenário são apresentados na Tabela 4.2, onde podemos concluir que:

- O viés dos estimadores de máxima verossimilhança dos $\boldsymbol{\beta}$'s aproxima-se de zero à medida que o tamanho da amostra aumenta; os estimadores de máxima verossimilhança de $\boldsymbol{\gamma}_{01}$ e $\boldsymbol{\gamma}_{02}$ apresentam viés considerável para amostras pequenas, entretanto diminui a medida que o tamanho da amostra aumenta; o estimador de máxima verossimilhança do $\lambda$ apresenta viés considerável para amostras pequenas nos MRSB via cópulas Clayton e Frank, entretanto diminui a medida que o tamanho da amostra aumenta;

- O REQM dos estimadores de máxima verossimilhança dos $\boldsymbol{\beta}_{01}$ e $\boldsymbol{\beta}_{11}$ aproxima-se de zero à medida que o tamanho da amostra aumenta, enquanto que os estimadores de máxima verossimilhança dos $\boldsymbol{\beta}_{11}$ e $\boldsymbol{\beta}_{12}$ apresentam REQM considerável para amostras pequenas, entretanto diminui a medida que o tamanho da amostra aumenta; os estimadores de máxima verossimilhança dos $\boldsymbol{\gamma}$'s e do $\lambda$ apresentam REQM considerável para amostas pequenas, entretanto diminui a medida que o tamanho da amostra aumenta;

- Os estimadores de máxima verossimilhança dos $\boldsymbol{\beta}$'s e $\boldsymbol{\gamma}$'s apresentam Taxa de Cobertura levemente inferior ao nível nominal de 95% para amostra pequena, mas a medida que a amostra aumenta aproxima-se ao nível nominal de 95%; o estimador de máxima verossimilhança do $\lambda$ apresenta Taxa de Cobertura superior ao coeficiente de confiança de 95% nos MRSB via cópula FGM e Clayton.

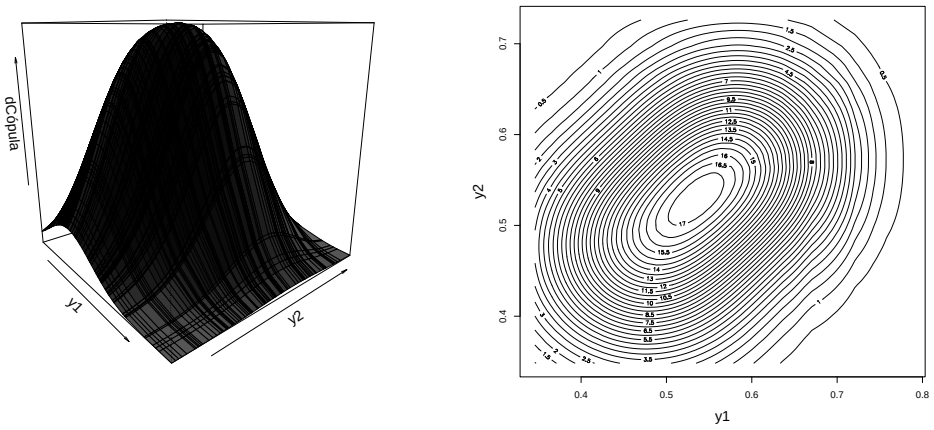
Pode-se concluir que assintoticamente os estimadores de máxima verossimilhança podem ser considerados estimadores com boas propriedades, isto é, são não viesados, eficientes e consistente.

**Tabela 4.2:** *Média, viés, REQM e Taxa de Cobertura de 95% de confiança. Cenário 1.*

| Cópula | $n$ | Medida | $\beta_{01}$ | $\beta_{11}$ | $\beta_{02}$ | $\beta_{12}$ | $\gamma_{01}$ | $\gamma_{11}$ | $\gamma_{02}$ | $\gamma_{12}$ | $\lambda$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FGM | 25 | Média | -3,50 | 1,20 | -3,50 | 1,20 | -0,99 | 1,62 | -1,00 | 1,60 | 0,46 |
| | | Viés | 0,00 | 0,00 | 0,00 | 0,00 | 0,19 | -0,02 | 0,20 | 0,00 | 0,04 |
| | | REQM | 0,05 | 0,13 | 0,05 | 0,13 | 0,53 | 0,99 | 0,53 | 0,99 | 0,43 |
| | | TC(95%) | 91,00 | 92,00 | 92,00 | 91,00 | 91,00 | 92,00 | 92,00 | 91,00 | 97,00 |
| | 50 | Média | -3,50 | 1,20 | -3,50 | 1,20 | -0,90 | 1,64 | -0,87 | 1,58 | 0,47 |
| | | Viés | 0,00 | 0,00 | 0,00 | 0,00 | 0,10 | -0,04 | 0,07 | 0,02 | 0,03 |
| | | REQM | 0,04 | 0,11 | 0,04 | 0,10 | 0,38 | 0,73 | 0,36 | 0,69 | 0,31 |
| | | TC(95%) | 92,00 | 94,00 | 94,00 | 94,00 | 93,00 | 94,00 | 94,00 | 95,00 | 96,00 |
| | 75 | Média | -3,50 | 1,20 | -3,50 | 1,20 | -0,86 | 1,61 | -0,87 | 1,62 | 0,49 |
| | | Viés | 0,00 | 0,00 | 0,00 | 0,00 | 0,06 | -0,01 | 0,07 | -0,02 | 0,01 |
| | | REQM | 0,03 | 0,07 | 0,03 | 0,07 | 0,26 | 0,47 | 0,28 | 0,48 | 0,26 |
| | | TC(95%) | 95,00 | 94,00 | 94,00 | 95,00 | 96,00 | 95,00 | 93,00 | 95,00 | 97,00 |
| | 100 | Média | -3,50 | 1,20 | -3,50 | 1,20 | -0,86 | 1,64 | -0,86 | 1,64 | 0,49 |
| | | Viés | 0,00 | 0,00 | 0,00 | 0,00 | 0,06 | -0,04 | 0,06 | -0,04 | 0,01 |
| | | REQM | 0,02 | 0,06 | 0,03 | 0,06 | 0,25 | 0,42 | 0,24 | 0,42 | 0,23 |
| | | TC(95%) | 95,00 | 94,00 | 94,00 | 96,00 | 94,00 | 94,00 | 94,00 | 94,00 | 96,00 |
| Clayton | 25 | Média | -3,50 | 1,19 | -3,50 | 1,20 | -0,96 | 1,55 | -0,93 | 1,52 | 0,63 |
| | | Viés | 0,00 | 0,01 | 0,00 | 0,00 | 0,16 | 0,05 | 0,13 | 0,08 | -0,13 |
| | | REQM | 0,05 | 0,14 | 0,06 | 0,15 | 0,53 | 0,94 | 0,51 | 0,94 | 0,36 |
| | | TC(95%) | 92,00 | 92,00 | 90,00 | 90,00 | 89,00 | 90,00 | 91,00 | 90,00 | 93,00 |
| | 50 | Média | -3,50 | 1,20 | -3,50 | 1,19 | -0,90 | 1,62 | -0,89 | 1,61 | 0,55 |
| | | Viés | 0,00 | 0,00 | 0,00 | 0,01 | 0,10 | -0,02 | 0,09 | -0,01 | -0,05 |
| | | REQM | 0,03 | 0,09 | 0,03 | 0,09 | 0,33 | 0,59 | 0,32 | 0,59 | 0,22 |
| | | TC(95%) | 93,00 | 93,00 | 94,00 | 94,00 | 93,00 | 93,00 | 94,00 | 94,00 | 94,00 |
| | 75 | Média | -3,50 | 1,19 | -3,50 | 1,20 | -0,85 | 1,59 | -0,85 | 1,61 | 0,53 |
| | | Viés | 0,00 | 0,01 | 0,00 | 0,00 | 0,05 | 0,01 | 0,05 | -0,01 | -0,03 |
| | | REQM | 0,03 | 0,07 | 0,03 | 0,08 | 0,25 | 0,43 | 0,25 | 0,45 | 0,17 |
| | | TC(95%) | 94,00 | 96,00 | 92,00 | 93,00 | 95,00 | 95,00 | 93,00 | 95,00 | 93,00 |
| | 100 | Média | -3,50 | 1,20 | -3,50 | 1,20 | -0,85 | 1,62 | -0,85 | 1,63 | 0,52 |
| | | Viés | 0,00 | 0,00 | 0,00 | 0,00 | 0,05 | -0,02 | 0,05 | -0,03 | -0,02 |
| | | REQM | 0,03 | 0,06 | 0,03 | 0,06 | 0,24 | 0,39 | 0,25 | 0,40 | 0,15 |
| | | TC(95%) | 94,00 | 95,00 | 93,00 | 94,00 | 94,00 | 95,00 | 92,00 | 93,00 | 95,00 |
| Frank | 25 | Média | -3,50 | 1,20 | -3,50 | 1,19 | -0,96 | 1,56 | -0,99 | 1,61 | 0,27 |
| | | Viés | 0,00 | 0,00 | 0,00 | 0,01 | 0,16 | 0,04 | 0,19 | -0,01 | 0,23 |
| | | REQM | 0,04 | 0,14 | 0,04 | 0,13 | 0,50 | 0,97 | 0,51 | 0,96 | 0,65 |
| | | TC(95%) | 90,00 | 91,00 | 90,00 | 93,00 | 90,00 | 91,00 | 90,00 | 91,00 | 100 |
| | 50 | Média | -3,50 | 1,20 | -3,50 | 1,20 | -0,99 | 1,77 | -0,94 | 1,71 | 0,36 |
| | | Viés | 0,00 | 0,00 | 0,00 | 0,00 | 0,19 | -0,17 | 0,14 | -0,11 | 0,14 |
| | | REQM | 0,04 | 0,09 | 0,04 | 0,09 | 0,45 | 0,71 | 0,44 | 0,68 | 0,54 |
| | | TC(95%) | 93,00 | 94,00 | 93,00 | 93,00 | 92,00 | 92,00 | 92,00 | 94,00 | 100 |
| | 75 | Média | -3,50 | 1,19 | -3,50 | 1,20 | -0,85 | 1,59 | -0,87 | 1,65 | 0,39 |
| | | Viés | 0,00 | 0,01 | 0,00 | 0,00 | 0,05 | 0,01 | 0,07 | -0,05 | 0,11 |
| | | REQM | 0,03 | 0,07 | 0,03 | 0,07 | 0,27 | 0,51 | 0,28 | 0,52 | 0,47 |
| | | TC(95%) | 94,00 | 95,00 | 93,00 | 93,00 | 94,00 | 95,00 | 93,00 | 93,00 | 98,00 |
| | 100 | Média | -3,50 | 1,20 | -3,50 | 1,20 | -0,86 | 1,63 | -0,86 | 1,63 | 0,43 |
| | | Viés | 0,00 | 0,00 | 0,00 | 0,00 | 0,06 | -0,03 | 0,06 | -0,03 | 0,07 |
| | | REQM | 0,02 | 0,06 | 0,02 | 0,06 | 0,25 | 0,38 | 0,26 | 0,40 | 0,41 |
| | | TC(95%) | 93,00 | 95,00 | 94,00 | 94,00 | 94,00 | 94,00 | 93,00 | 94,00 | 97,00 |

### 4.3.2 Cenário 2

Neste cenário é consideramos o vetor de parâmetros reais $\boldsymbol{\theta} = (\beta_{01} = -0,5; \beta_{11} = 1,2; \beta_{02} = -0,5; \beta_{12} = 1,2; \gamma_{01} = -1,5; \gamma_{11} = 1,3; \gamma_{02} = -1,5; \gamma_{12} = 1,3; \lambda = 0,5)^\top$ que é concretizado pelo gráfico em 4.2 (a) corresponde ao gráfico de superfícies próximo a 0,5; uma representação das curvas de nível pode ser vista em 4.2 (b).




**(a)** **(b)**

**Figura 4.2:** *Gráficos de superfície e curvas de nível para uma amostra de tamanho 100 e* $\boldsymbol{\theta} = (\beta_{01} = -0,5; \beta_{11} = 1,2; \beta_{02} = -0,5; \beta_{12} = 1,2; \gamma_{01} = -1,5; \gamma_{11} = 1,3; \gamma_{02} = -1,5; \gamma_{12} = 1,3; \lambda = 0,5)^\top$.

Os resultados encontrados neste cenário são apresentados na Tabela 4.3, onde podemos concluir que:

- O viés dos estimadores de máxima verossimilhança dos $\boldsymbol{\beta}$'s aproxima-se de zero à medida que o tamanho da amostra aumente; os estimadores de máxima verossimilhança dos $\boldsymbol{\gamma}$'s aproxima-se de zero à medida que o tamanho da amostra aumente para o MRSB via cópulas FGM e viés considerável para amostras pequenas nos MRSB via cópulas Clayton e Frank, entretanto diminui a medida que o tamanho da amostra aumenta; o estimador de máxima verossimilhança do $\lambda$ apresenta viés considerável para amostras pequenas, entretanto diminui a medida que o tamanho da amostra aumenta;

- Os estimadores de máxima verossimilhança de $\boldsymbol{\beta}_{11}$, $\boldsymbol{\beta}_{11}$, dos $\boldsymbol{\gamma}$'s e do $\lambda$ apresentam REQM considerável para amostras pequenas, entretanto diminui a medida que o tamanho da amostra aumenta;

- Os estimadores de máxima verossimilhança dos $\boldsymbol{\beta}$'s e $\boldsymbol{\gamma}$'s apresentam Taxa de Cobertura levemente inferior ao nível nominal de 95% para amostra pequena, mas a medida que a amostra aumenta aproxima-se ao nível nominal de 95%; o estimador de máxima verossimilhança do $\lambda$ apresenta Taxa de Cobertura superior ao coeficiente de confiança de 95%.

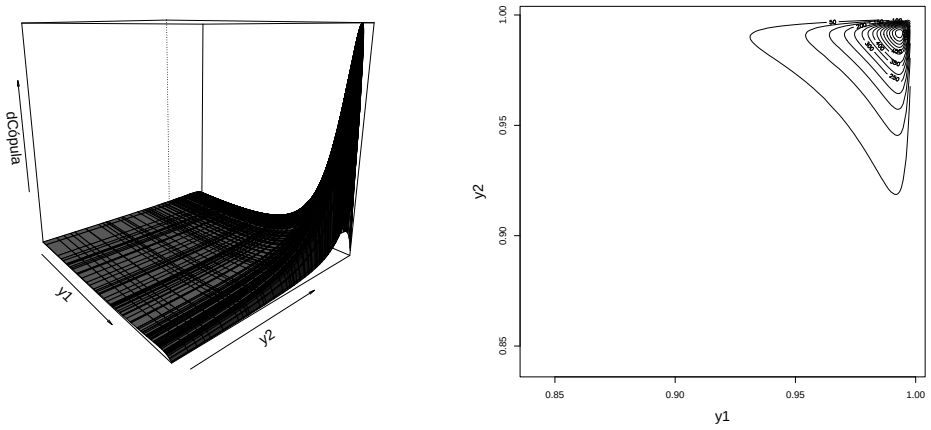
Pode-se concluir que assintoticamente os estimadores de máxima verossimilhança podem ser considerados estimadores com boas propriedades, isto é, são não viesados, eficientes e consistente.

**Tabela 4.3:** *Média, viés, REQM e Taxa de Cobertura de 95% de confiança. Cenário 2.*

| Cópula | $n$ | Medida | $\beta_{01}$ | $\beta_{11}$ | $\beta_{02}$ | $\beta_{12}$ | $\gamma_{01}$ | $\gamma_{11}$ | $\gamma_{02}$ | $\gamma_{12}$ | $\lambda$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FGM | 25 | Média | -0,50 | 1,22 | -0,50 | 1,22 | -1,62 | 1,34 | -1,62 | 1,30 | 0,40 |
| | | Viés | 0,00 | -0,02 | 0,00 | -0,02 | 0,12 | -0,04 | 0,12 | 0,00 | 0,10 |
| | | REQM | 0,11 | 0,23 | 0,11 | 0,23 | 0,63 | 1,16 | 0,62 | 1,17 | 0,47 |
| | | TC(95%) | 92,00 | 93,00 | 93,00 | 93,00 | 93,00 | 93,00 | 94,00 | 93,00 | 97,00 |
| | 50 | Média | -0,51 | 1,20 | -0,51 | 1,19 | -1,56 | 1,32 | -1,58 | 1,34 | 0,48 |
| | | Viés | 0,01 | 0,00 | 0,01 | 0,01 | 0,06 | -0,02 | 0,08 | -0,04 | 0,02 |
| | | REQM | 0,07 | 0,15 | 0,07 | 0,16 | 0,38 | 0,76 | 0,38 | 0,73 | 0,35 |
| | | TC(95%) | 92,00 | 93,00 | 94,00 | 94,00 | 94,00 | 93,00 | 94,00 | 94,00 | 96,00 |
| | 75 | Média | -0,50 | 1,20 | -0,50 | 1,20 | -1,56 | 1,31 | -1,56 | 1,32 | 0,49 |
| | | Viés | 0,00 | 0,00 | 0,00 | 0,00 | 0,06 | -0,01 | 0,06 | -0,02 | 0,01 |
| | | REQM | 0,05 | 0,11 | 0,05 | 0,11 | 0,29 | 0,53 | 0,29 | 0,54 | 0,27 |
| | | TC(95%) | 94,00 | 93,00 | 94,00 | 94,00 | 94,00 | 95,00 | 92,00 | 94,00 | 97,00 |
| | 100 | Média | -0,50 | 1,20 | -0,50 | 1,20 | -1,58 | 1,34 | -1,55 | 1,31 | 0,50 |
| | | Viés | 0,00 | 0,00 | 0,00 | 0,00 | 0,08 | -0,04 | 0,05 | -0,01 | 0,00 |
| | | REQM | 0,04 | 0,08 | 0,04 | 0,08 | 0,25 | 0,40 | 0,24 | 0,40 | 0,23 |
| | | TC(95%) | 94,00 | 95,00 | 94,00 | 94,00 | 94,00 | 95,00 | 95,00 | 94,00 | 95,00 |
| Clayton | 25 | Média | -0,51 | 1,20 | -0,50 | 1,19 | -1,74 | 1,52 | -1,70 | 1,43 | 0,74 |
| | | Viés | 0,01 | 0,00 | 0,00 | 0,01 | 0,24 | -0,22 | 0,20 | -0,13 | -0,24 |
| | | REQM | 0,09 | 0,18 | 0,09 | 0,18 | 0,57 | 0,98 | 0,55 | 0,99 | 0,37 |
| | | TC(95%) | 90,00 | 92,00 | 91,00 | 92,00 | 90,00 | 92,00 | 91,00 | 92,00 | 97,00 |
| | 50 | Média | -0,50 | 1,21 | -0,50 | 1,20 | -1,54 | 1,28 | -1,52 | 1,23 | 0,60 |
| | | Viés | 0,00 | -0,01 | 0,00 | 0,00 | 0,04 | 0,02 | 0,02 | 0,07 | -0,10 |
| | | REQM | 0,06 | 0,13 | 0,06 | 0,12 | 0,33 | 0,60 | 0,32 | 0,60 | 0,22 |
| | | TC(95%) | 94,00 | 94,00 | 95,00 | 94,00 | 93,00 | 93,00 | 93,00 | 93,00 | 96,00 |
| | 75 | Média | -0,50 | 1,20 | -0,50 | 1,20 | -1,53 | 1,29 | -1,54 | 1,32 | 0,58 |
| | | Viés | 0,00 | 0,00 | 0,00 | 0,00 | 0,03 | 0,01 | 0,04 | -0,02 | -0,08 |
| | | REQM | 0,05 | 0,09 | 0,05 | 0,09 | 0,25 | 0,44 | 0,27 | 0,46 | 0,18 |
| | | TC(95%) | 94,00 | 95,00 | 94,00 | 95,00 | 94,00 | 94,00 | 93,00 | 94,00 | 96,00 |
| | 100 | Média | -0,50 | 1,20 | -0,50 | 1,20 | -1,52 | 1,30 | -1,51 | 1,29 | 0,55 |
| | | Viés | 0,00 | 0,00 | 0,00 | 0,00 | 0,02 | 0,00 | 0,01 | 0,01 | -0,05 |
| | | REQM | 0,04 | 0,08 | 0,04 | 0,08 | 0,22 | 0,37 | 0,22 | 0,39 | 0,15 |
| | | TC(95%) | 93,00 | 94,00 | 95,00 | 94,00 | 94,00 | 95,00 | 94,00 | 94,00 | 96,00 |
| Frank | 25 | Média | -0,50 | 1,20 | -0,49 | 1,19 | -1,75 | 1,40 | -1,80 | 1,50 | 0,29 |
| | | Viés | 0,00 | 0,00 | -0,01 | 0,01 | 0,25 | -0,10 | 0,30 | -0,20 | 0,21 |
| | | REQM | 0,09 | 0,16 | 0,09 | 0,16 | 0,56 | 0,84 | 0,61 | 0,88 | 0,65 |
| | | TC(95%) | 89,00 | 91,00 | 90,00 | 92,00 | 91,00 | 93,00 | 89,00 | 93,00 | 100 |
| | 50 | Média | -0,50 | 1,20 | -0,50 | 1,20 | -1,55 | 1,23 | -1,57 | 1,29 | 0,38 |
| | | Viés | 0,00 | 0,00 | 0,00 | 0,00 | 0,05 | 0,07 | 0,07 | 0,01 | 0,12 |
| | | REQM | 0,05 | 0,13 | 0,05 | 0,13 | 0,29 | 0,64 | 0,30 | 0,63 | 0,53 |
| | | TC(95%) | 94,00 | 95,00 | 93,00 | 93,00 | 94,00 | 94,00 | 93,00 | 94,00 | 100 |
| | 75 | Média | -0,50 | 1,21 | -0,50 | 1,21 | -1,59 | 1,35 | -1,56 | 1,33 | 0,41 |
| | | Viés | 0,00 | -0,01 | 0,00 | -0,01 | 0,09 | -0,05 | 0,06 | -0,03 | 0,09 |
| | | REQM | 0,05 | 0,10 | 0,05 | 0,10 | 0,28 | 0,47 | 0,27 | 0,46 | 0,45 |
| | | TC(95%) | 94,00 | 94,00 | 93,00 | 94,00 | 94,00 | 94,00 | 93,00 | 94,00 | 97,00 |
| | 100 | Média | -0,50 | 1,20 | -0,50 | 1,20 | -1,55 | 1,33 | -1,54 | 1,31 | 0,44 |
| | | Viés | 0,00 | 0,00 | 0,00 | 0,00 | 0,05 | -0,03 | 0,04 | -0,01 | 0,06 |
| | | REQM | 0,04 | 0,08 | 0,04 | 0,08 | 0,23 | 0,41 | 0,23 | 0,39 | 0,41 |
| | | TC(95%) | 95,00 | 94,00 | 94,00 | 96,00 | 93,00 | 93,00 | 93,00 | 94,00 | 98,00 |

### 4.3.3 Cenário 3

Neste cenário consideramos o vetor de parâmetros reais $\boldsymbol{\theta} = (\beta_{01} = 2,5; \beta_{11} = 1,2; \beta_{02} = 2,5; \beta_{12} = 1,2; \gamma_{01} = 0,8; \gamma_{11} = 1,6; \gamma_{02} = 0,8; \gamma_{12} = 1,6; \lambda = 0,5)^\top$ que é concretizado pelo gráfico em 4.3 (a) corresponde ao gráfico de superfícies próximo a 1; uma representação das curvas de nível pode ser vista em 4.3 (b).




**(a)** **(b)**

**Figura 4.3:** *Gráficos de superfície e curvas de nível para uma amostra de tamanho 100 e* $\boldsymbol{\theta} = (\beta_{01} = 2,5; \beta_{11} = 1,2; \beta_{02} = 2,5; \beta_{12} = 1,2; \gamma_{01} = 0,8; \gamma_{11} = 1,6; \gamma_{02} = 0,8; \gamma_{12} = 1,6; \lambda = 0,5)^\top$.

Os resultados encontrados neste cenário são apresentados na Tabela 4.4, onde podemos concluir que:

- Os estimadores de máxima verossimilhança dos $\boldsymbol{\beta}$'s aproxima-se de zero à medida que o tamanho da amostra aumenta; os estimadores de máxima verossimilhança de $\boldsymbol{\gamma}_{01}$ e $\boldsymbol{\gamma}_{02}$ apresentam viés considerável para amostras pequenas, entretanto diminui a medida que o tamanho da amostra aumenta; o estimador de máxima verossimilhança do $\lambda$ apresenta viés nos MRSB via cópulas Clayton e Frank considerável para amostras pequenas, entretanto diminui a medida que o tamanho da amostra aumenta;

- Os estimadores de máxima verossimilhança de $\boldsymbol{\beta}_{11}$ e $\boldsymbol{\beta}_{11}$, dos $\boldsymbol{\gamma}$'s e do $\lambda$ apresenta REQM considerável para amostras pequenas, entretanto diminui a medida que o tamanho da amostra aumenta;

- Os estimadores de máxima verossimilhança dos $\boldsymbol{\beta}$'s e $\boldsymbol{\gamma}$'s apresentam Taxa de Cobertura levemente inferior ao nível nominal de 95% para amostras pequenas, mas a medida que a amostra aumenta aproxima-se ao nível nominal o valor da Taxa de 95%; o estimador de máxima verossimilhança do $\lambda$ apresenta Taxa de Cobertura superior ao coeficiente de confiança de 95% nos MRSB via cópulas FGM e Clayton.

Pode-se concluir que assintoticamente os estimadores de máxima verossimilhança podem ser considerados estimadores com boas propriedades, isto é, são não viesados, eficientes e consistente.

**Tabela 4.4:** *Média, viés, REQM e Taxa de Cobertura de 95% de confiança. Cenário 3.*

| Cópula | $n$ | Medida | $\beta_{01}$ | $\beta_{11}$ | $\beta_{02}$ | $\beta_{12}$ | $\gamma_{01}$ | $\gamma_{11}$ | $\gamma_{02}$ | $\gamma_{12}$ | $\lambda$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FGM | 25 | Média | 2,51 | 1,19 | 2,51 | 1,20 | 0,60 | 1,62 | 0,62 | 1,62 | 0,43 |
| | | Viés | -0,01 | 0,01 | -0,01 | 0,00 | 0,20 | -0,02 | 0,18 | -0,02 | 0,07 |
| | | REQM | 0,11 | 0,21 | 0,11 | 0,21 | 0,44 | 0,73 | 0,43 | 0,73 | 0,44 |
| | | TC(95%) | 90,00 | 93,00 | 92,00 | 93,00 | 90,00 | 93,00 | 92,00 | 92,00 | 96,00 |
| | 50 | Média | 2,51 | 1,19 | 2,50 | 1,19 | 0,64 | 1,72 | 0,68 | 1,66 | 0,47 |
| | | Viés | -0,01 | 0,01 | 0,00 | 0,01 | 0,16 | -0,12 | 0,12 | -0,06 | 0,03 |
| | | REQM | 0,11 | 0,19 | 0,11 | 0,19 | 0,42 | 0,66 | 0,41 | 0,68 | 0,31 |
| | | TC(95%) | 92,00 | 94,00 | 93,00 | 94,00 | 92,00 | 94,00 | 93,00 | 93,00 | 97,00 |
| | 75 | Média | 2,51 | 1,19 | 2,51 | 1,19 | 0,75 | 1,60 | 0,73 | 1,62 | 0,50 |
| | | Viés | -0,01 | 0,01 | -0,01 | 0,01 | 0,05 | 0,00 | 0,07 | -0,02 | 0,00 |
| | | REQM | 0,07 | 0,14 | 0,07 | 0,14 | 0,27 | 0,50 | 0,26 | 0,48 | 0,26 |
| | | TC(95%) | 94,00 | 95,00 | 94,00 | 94,00 | 94,00 | 93,00 | 93,00 | 94,00 | 97,00 |
| | 100 | Média | 2,50 | 1,21 | 2,51 | 1,19 | 0,76 | 1,60 | 0,74 | 1,63 | 0,50 |
| | | Viés | 0,00 | -0,01 | -0,01 | 0,01 | 0,04 | 0,00 | 0,06 | -0,03 | 0,00 |
| | | REQM | 0,07 | 0,12 | 0,07 | 0,12 | 0,26 | 0,41 | 0,25 | 0,41 | 0,23 |
| | | TC(95%) | 95,00 | 95,00 | 94,00 | 93,00 | 94,00 | 95,00 | 96,00 | 95,00 | 96,00 |
| Clayton | 25 | Média | 2,51 | 1,19 | 2,51 | 1,19 | 0,53 | 1,81 | 0,50 | 1,83 | 0,64 |
| | | Viés | -0,01 | 0,01 | -0,01 | 0,01 | 0,27 | -0,21 | 0,30 | -0,23 | -0,14 |
| | | REQM | 0,16 | 0,27 | 0,15 | 0,28 | 0,63 | 1,00 | 0,65 | 1,02 | 0,36 |
| | | TC(95%) | 87,00 | 90,00 | 87,00 | 89,00 | 90,00 | 91,00 | 89,00 | 92,00 | 93,00 |
| | 50 | Média | 2,51 | 1,20 | 2,50 | 1,20 | 0,69 | 1,68 | 0,69 | 1,65 | 0,56 |
| | | Viés | -0,01 | 0,00 | 0,00 | 0,00 | 0,11 | -0,08 | 0,11 | -0,05 | -0,06 |
| | | REQM | 0,09 | 0,17 | 0,09 | 0,16 | 0,35 | 0,55 | 0,36 | 0,57 | 0,22 |
| | | TC(95%) | 92,00 | 93,00 | 93,00 | 94,00 | 92,00 | 93,00 | 93,00 | 94,00 | 94,00 |
| | 75 | Média | 2,50 | 1,20 | 2,50 | 1,20 | 0,74 | 1,62 | 0,72 | 1,62 | 0,53 |
| | | Viés | 0,00 | 0,00 | 0,00 | 0,00 | 0,06 | -0,02 | 0,08 | -0,02 | -0,03 |
| | | REQM | 0,08 | 0,13 | 0,08 | 0,13 | 0,26 | 0,43 | 0,28 | 0,44 | 0,18 |
| | | TC(95%) | 93,00 | 93,00 | 93,00 | 92,00 | 95,00 | 95,00 | 93,00 | 94,00 | 93,00 |
| | 100 | Média | 2,50 | 1,20 | 2,50 | 1,20 | 0,76 | 1,58 | 0,77 | 1,58 | 0,52 |
| | | Viés | 0,00 | 0,00 | 0,00 | 0,00 | 0,04 | 0,02 | 0,03 | 0,02 | -0,02 |
| | | REQM | 0,06 | 0,13 | 0,06 | 0,12 | 0,22 | 0,40 | 0,21 | 0,40 | 0,15 |
| | | TC(95%) | 95,00 | 93,00 | 95,00 | 95,00 | 95,00 | 93,00 | 95,00 | 95,00 | 93,00 |
| Frank | 25 | Média | 2,50 | 1,22 | 2,50 | 1,21 | 0,56 | 1,67 | 0,61 | 1,63 | 0,22 |
| | | Viés | 0,00 | -0,02 | 0,00 | -0,01 | 0,24 | -0,07 | 0,19 | -0,03 | 0,28 |
| | | REQM | 0,13 | 0,25 | 0,14 | 0,25 | 0,54 | 0,92 | 0,53 | 0,92 | 0,66 |
| | | TC(95%) | 90,00 | 91,00 | 90,00 | 92,00 | 91,00 | 93,00 | 92,00 | 92,00 | 100 |
| | 50 | Média | 2,50 | 1,21 | 2,51 | 1,19 | 0,71 | 1,60 | 0,74 | 1,55 | 0,37 |
| | | Viés | 0,00 | -0,01 | -0,01 | 0,01 | 0,09 | 0,00 | 0,06 | 0,05 | 0,13 |
| | | REQM | 0,10 | 0,19 | 0,10 | 0,19 | 0,36 | 0,63 | 0,36 | 0,65 | 0,52 |
| | | TC(95%) | 93,00 | 94,00 | 92,00 | 93,00 | 93,00 | 94,00 | 93,00 | 93,00 | 100 |
| | 75 | Média | 2,50 | 1,19 | 2,50 | 1,19 | 0,73 | 1,62 | 0,74 | 1,60 | 0,40 |
| | | Viés | 0,00 | 0,01 | 0,00 | 0,01 | 0,07 | -0,02 | 0,06 | 0,00 | 0,10 |
| | | REQM | 0,09 | 0,17 | 0,09 | 0,16 | 0,33 | 0,56 | 0,31 | 0,53 | 0,47 |
| | | TC(95%) | 93,00 | 94,00 | 95,00 | 94,00 | 92,00 | 93,00 | 95,00 | 95,00 | 97,00 |
| | 100 | Média | 2,50 | 1,20 | 2,50 | 1,20 | 0,75 | 1,62 | 0,77 | 1,58 | 0,42 |
| | | Viés | 0,00 | 0,00 | 0,00 | 0,00 | 0,05 | -0,02 | 0,03 | 0,02 | 0,08 |
| | | REQM | 0,07 | 0,12 | 0,06 | 0,12 | 0,23 | 0,40 | 0,24 | 0,40 | 0,40 |
| | | TC(95%) | 94,00 | 95,00 | 96,00 | 95,00 | 95,00 | 95,00 | 94,00 | 93,00 | 97,00 |

## 4.4 Aplicações

Nesta seção iremos apresentar duas aplicações, como descrita a seguir:

### 4.4.1 Aplicação I

Nesta Seção temos por interesse estudar a relação entre as variáveis $\boldsymbol{y}_1$: proporção de pobres, $\boldsymbol{y}_2$: taxa de mortalidade infantil, com respeito a $x$: Índice de Desenvolvimento Humano (IDH) por município. Conforme informado na Seção 3.2, os dados podem ser acessados facilmente em (Municipal, 2023).

A seguir, na Tabela 4.5 apresenta-se as estimativas, erro padrão e valor-$p$ dos parâmetros para os modelos de regressão bivariado Simplex e Beta via cópulas FGM, Clayton e Frank, cujas estimativas são significativas ao nível de 1% de significância. Como descrito na Seção 3.1, as estimativas associadas ao parâmetro da média podem ser interpretados como a razão de chance, isto é, $\exp(0,01 \times \widehat{\beta}_{11}) = \exp(-11,205) \simeq 0,894$; $\exp(0,01 \times \widehat{\beta}_{11}) = \exp(-11,030) \simeq 0,895$; $\exp(0,01 \times \widehat{\beta}_{11}) = \exp(-11,229) \simeq 0,894$, ou seja, com o aumento de 1% no IDH por município ($x$), a chance de redução da proporção de pobres é de aproximadamente 10,60%, 10,44% e 10,62%, enquanto que com o aumento de 1% no IDH por município ($x$), isto é, $\exp(0,01 \times \widehat{\beta}_{12}) = \exp(-9,573) \simeq 0,909$; $\exp(0,01 \times \widehat{\beta}_{12}) = \exp(-9,093) \simeq 0,913$; $\exp(0,01 \times \widehat{\beta}_{12}) = \exp(-9,392) \simeq 0,910$, ou seja, a chance de redução da taxa de mortalidade infantil é de aproximadamente 9,13%, 8,69% e 8,96% com base no modelo de regressão Simplex bivariado pelas cópulas FGM, Clayton e Frank, respectivamente. Já pelo modelo de regressão Beta bivariado, as estimativas associadas ao parâmetro da média podem ser interpretados como a razão de chance, isto é, $\exp(0,01 \times \widehat{\beta}_{11}) = \exp(-11,664) \simeq 0,889$; $\exp(0,01 \times \widehat{\beta}_{11}) = \exp(-11,601) \simeq 0,890$; $\exp(0,01 \times \widehat{\beta}_{11}) = \exp(-11,738) \simeq 0,889$, ou seja, com o aumento de 1% no IDH por município, a chance de redução da proporção de pobre aumente em aproximadamente 11,00%, 10,95% e 11,07%, enquanto que com o aumento de 1% no IDH por município ($x$), isto é, $\exp(0,01 \times \widehat{\beta}_{12}) = \exp(-9,627) \simeq 0,908$; $\exp(0,01 \times \widehat{\beta}_{12}) = \exp(-9,333) \simeq 0,911$; $\exp(0,01 \times \widehat{\beta}_{12}) = \exp(-9,519) \simeq 0,909$, ou seja, a chance de redução da taxa de mortalidade infantil aumenta em aproximadamente 9,18%, 8,91% e 9,08% pelas cópulas FGM, Clayton e Frank, respectivamente. Em vista uma das cópulas, podemos notar que os modelos proposto através da cópula FGM ajustou melhor os dados, tanto para o modelo de regressão Simplex bivariado quanto para o modelo de regressão Beta bivariado, uma vez que, apresentou valores mínimos de AIC e BIC. Sendo melhor representado pela distribuição Simplex quando levado em consideração as cópulas FGM e Frank, e melhor representado pela distribuição Beta quando levado em consideração a cópula Clayton.

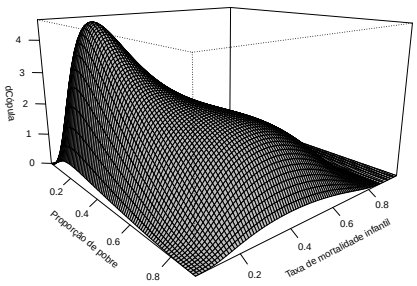
Com relação a proporção de pobres e a taxa de mortalidade infantil por dispersão, espera-se que o parâmetro de dispersão da proporção de pobres e da taxa de mortalidade infantil diminui tanto pelo modelo de regressão Simplex quanto pelo modelo de regressão Beta, tendo em vistas as estimativas negativas dos parâmetros $\boldsymbol{\gamma}_{01}$ e $\boldsymbol{\gamma}_{12}$, conforme apresentado na Tabela 4.5. O parâmetro $\tau$ reflete o grau de associação entre as variáveis dependentes, no caso $\boldsymbol{y}_1$ e $\boldsymbol{y}_2$, dependendo unicamente da cópula utilizada e obtido a partir da relação com o parâmetro $\lambda$, conforme apresentado na Tabela 4.1. Com isso, vemos que pelas cópulas FGM e Frank ambos os modelos transmitem uma concordância ou associação negativa e fraca, enquanto que pela cópula Clayton ambos os modelos transmitem uma concordância positiva e relativamente forte.

Nas Figuras 4.4 e 4.5 é apresentado o comportamento gráfico de superfície da densidade e curvas de nível com os valores de $\boldsymbol{y}_1$ e $\boldsymbol{y}_2$ plotados para as cópulas FGM, Clayton e Frank, um contraponto as estatística quanto a qualidade do ajuste dos modelos. Podemos notar que a distribuição Simplex nas Figuas 5.6a e 5.6e caracterizam o aspecto bimodal correspondente a variável proporção de pobre $\boldsymbol{y}_1$, sendo um pouco mais evidente em 5.6a, enquanto que a distribuição Beta não apresenta tal aspecto bimodal. Em síntese, podemos notar que os modelos ajustam bem os dados, tendo em vista que os dados, em sua grande maioria, estão circunscrito pelas curvas de nível.

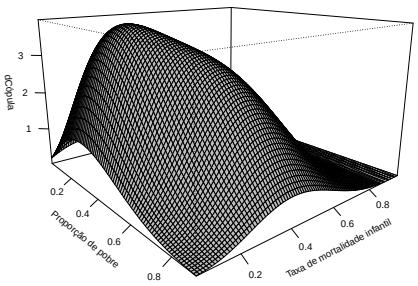
**Tabela 4.5:** *Estimativas, erro padrão e valor-p do modelo de regressão bivariado das distribuições Simplex e Beta.*

| Modelo | Distribuição | Cópula FGM | | | | Cópula de Clayton | | | Cópula de Frank | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Parâmetro | Estimativa | SE | $p$-valor | Estimativa | SE | $p$-valor | Estimativa | SE | $p$-valor |
| $y_1, y_2 \sim IDHM$ | Simplex | $\beta_{01}$ | 7,641 | 0,040 | 0,000 | 7,477 | 0,053 | 0,000 | 7,656 | 0,041 | 0,000 |
| | | $\beta_{11}$ | −11,205 | 0,060 | 0,000 | −11,030 | 0,071 | 0,000 | −11,229 | 0,061 | 0,000 |
| | | $\beta_{02}$ | 5,961 | 0,041 | 0,000 | 5,572 | 0,073 | 0,000 | 5,826 | 0,043 | 0,000 |
| | | $\beta_{12}$ | −9,573 | 0,056 | 0,000 | −9,093 | 0,091 | 0,000 | −9,392 | 0,058 | 0,000 |
| | | $\gamma_{01}$ | −1,413 | 0,079 | 0,000 | −1,402 | 0,082 | 0,000 | −1,330 | 0,081 | 0,000 |
| | | $\gamma_{11}$ | 1,801 | 0,113 | 0,000 | 1,946 | 0,125 | 0,000 | 1,694 | 0,114 | 0,000 |
| | | $\gamma_{02}$ | 0,044 | 0,067 | 0,000 | 0,298 | 0,072 | 0,000 | 0,158 | 0,068 | 0,000 |
| | | $\gamma_{12}$ | −0,269 | 0,095 | 0,000 | −0,479 | 0,097 | 0,000 | −0,420 | 0,096 | 0,000 |
| | | $\lambda$ | −0,359 | 0,044 | 0,000 | 0,670 | 0,126 | 0,000 | 0,800 | 0,092 | 0,000 |
| $y_1, y_2 \sim IDHM$ | Beta | $\beta_{01}$ | 8,011 | 0,049 | 0,000 | 7,938 | 0,053 | 0,000 | 8,062 | 0,051 | 0,000 |
| | | $\beta_{11}$ | −11,664 | 0,072 | 0,000 | −11,601 | 0,076 | 0,000 | −11,738 | 0,073 | 0,000 |
| | | $\beta_{02}$ | 5,988 | 0,043 | 0,000 | 5,758 | 0,050 | 0,000 | 5,915 | 0,044 | 0,000 |
| | | $\beta_{12}$ | −9,627 | 0,060 | 0,000 | −9,333 | 0,068 | 0,000 | −9,519 | 0,062 | 0,000 |
| | | $\gamma_{01}$ | −2,396 | 0,107 | 0,000 | −2,136 | 0,108 | 0,000 | −2,249 | 0,108 | 0,000 |
| | | $\gamma_{11}$ | 1,017 | 0,153 | 0,000 | 0,767 | 0,153 | 0,000 | 0,817 | 0,154 | 0,000 |
| | | $\gamma_{02}$ | 0,317 | 0,088 | 0,000 | 0,649 | 0,093 | 0,000 | 0,409 | 0,090 | 0,000 |
| | | $\gamma_{12}$ | −2,926 | 0,124 | 0,000 | −3,282 | 0,128 | 0,000 | −3,098 | 0,126 | 0,000 |
| | | $\lambda$ | −0,359 | 0,042 | 0,000 | 0,500 | 0,049 | 0,000 | 0,790 | 0,085 | 0,000 |

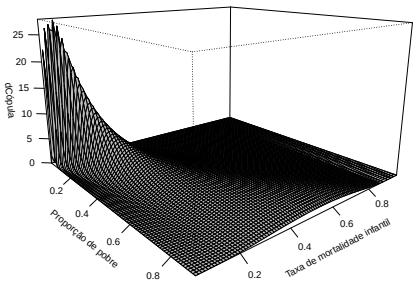
Simplex: [(Cópula FGM: AIC = 27604.4; BIC = 27663.9; $\tau$ =-0.0798), (Cópula de Clayton: AIC = 7941.5; BIC = 8001.0; $\tau$ = 0.6339), (Cópula de Frank: AIC = 27604.7; BIC = 27664.2; $\tau$ =-0.0774)];Beta: [(Cópula FGM: AIC = 28489.1; BIC = 28548.6; $\tau$ =-0.0782), (Cópula de Clayton: AIC = 7587.7; BIC = 7647.2; $\tau$ = 0.6142), (Cópula de Frank: AIC = 28486.4; BIC = 28545.9; $\tau$ =-0.0766)].



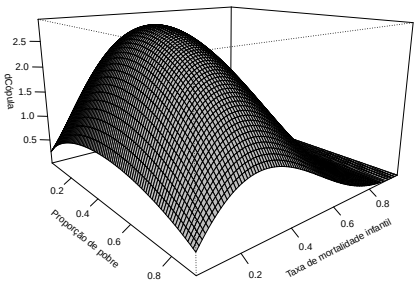

**(a)**
*FGM-Simplex:* $y_1, y_2 \sim IDHM$



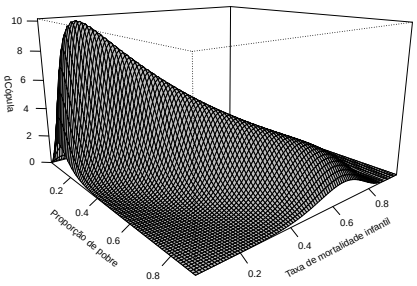

**(b)**
*FGM-Beta:* $y_1, y_2 \sim IDHM$



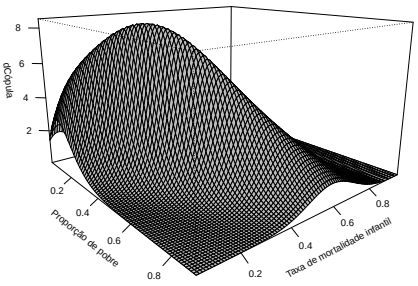

**(c)**
*Clayton-Simplex:* $y_1, y_2 \sim IDHM$




**(d)**
*Clayton-Beta:* $y_1, y_2 \sim IDHM$

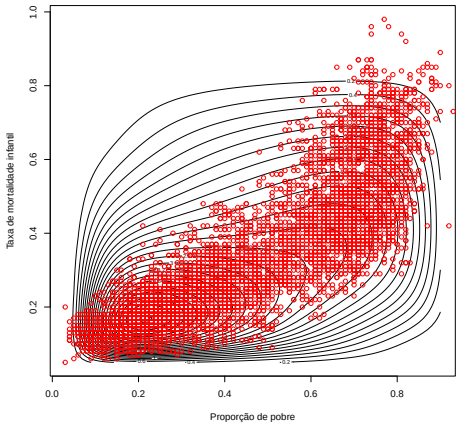



**(e)**
*Frank-Simplex:* $y_1, y_2 \sim IDHM$

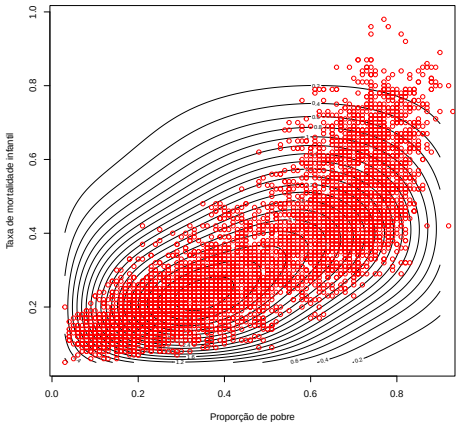



**(f)**
*Frank-Beta:* $y_1, y_2 \sim IDHM$

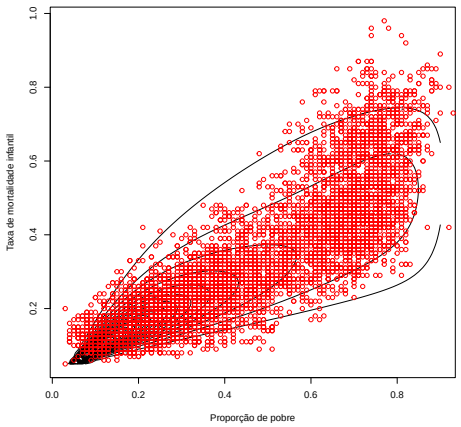
**Figura 4.4:** *Gráficos de superfícies das variáveis proporção de pobre e taxa de mortalidade infantil.*



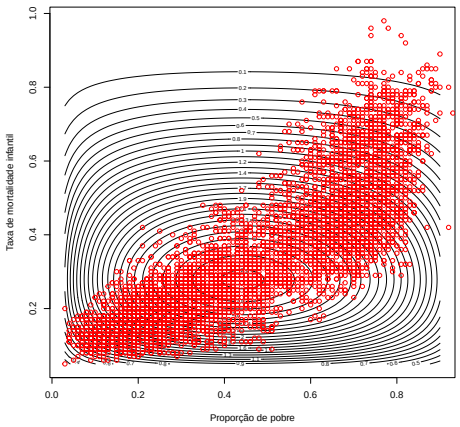

**(a)**
*FGM-Simplex:* $y_1, y_2 \sim IDHM$



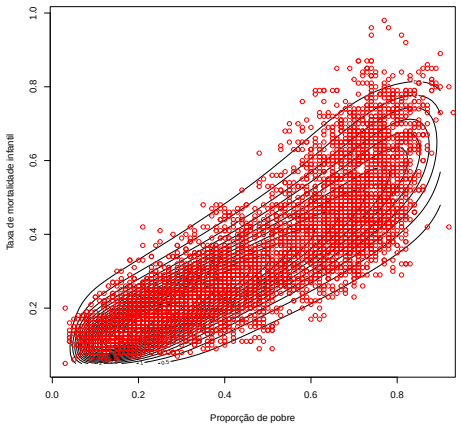

**(b)**
*FGM-Beta:* $y_1, y_2 \sim IDHM$



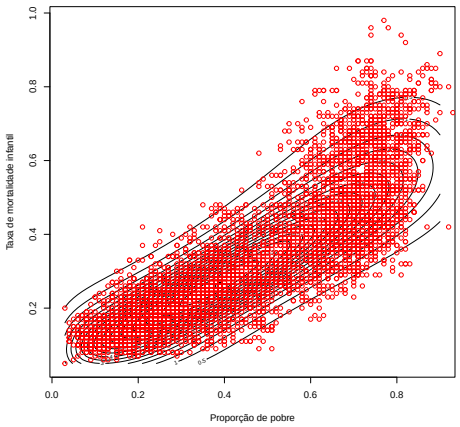

**(c)**
*Clayton-Simplex:* $y_1, y_2 \sim IDHM$




**(d)**
*Clayton-Beta:* $y_1, y_2 \sim IDHM$




**(e)**
*Frank-Simplex:* $y_1, y_2 \sim IDHM$




**(f)**
*Frank-Beta:* $y_1, y_2 \sim IDHM$

**Figura 4.5:** *Gráficos de curvas de nível das variáveis proporção de pobre e taxa de mortalidade infantil.*

### 4.4.2 Aplicação II

Nesta Seção temos por interesse estudar a relação entre as variáveis $\boldsymbol{y}_1$: Índice de Desenvolvimento Humano (IDH) por município[1], $\boldsymbol{y}_2$: Índice de Vulnerabilidade Social (IVS) por município[2] com respeito a $x$ = população economicamente dependente/população economicamente ativa (Razão de Dependência (RD), como variável independente)[3] dos 102 municípios do estado de Alagoas-Brasil, considera-se os dados registrados nos censos de 2000 e 2010, como sendo independentes e facilmente encontrados em (Ipea, 2023).

Na Tabela 4.6, apresenta-se algumas medidas descritivas para as variáveis em estudo, onde pode-se observar, que os municípios de Alagoas apresentam em média um IDH e IVS por município igual a 48% e 58%, respectivamente. A Razão de Dependência média é 68%, indicando que uma considerável população do estado de Alagoas, em idade produtiva, ampara uma grande número de dependentes. Observa-se também, a presença de assimetria negativa (à esquerda) para as variáveis IDH e IVS por município, assim como uma assimetria positiva (à direita) na variável RD.

**Tabela 4.6:** *Medidas descritivas $y_1$: Índice de Desenvolvimento Humano (IDH) municipal, $y_2$: Índice de Vulnerabilidade Social (IVS) e $x$: Razão Dependência (RD).*

| Variáveis | Max. | Min. | 1Q. | 3Q. | Med. | Md. | DP | Assi. | Curt. |
|---|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{y}_1$ | 0,72 | 0,28 | 0,38 | 0,56 | 0,48 | 0,50 | 0,09 | −0,06 | −1,25 |
| $\boldsymbol{y}_2$ | 0,82 | 0,34 | 0,52 | 0,64 | 0,58 | 0,58 | 0,09 | −0,09 | −0,33 |
| $x$ | 0,99 | 0,44 | 0,60 | 0,76 | 0,68 | 0,67 | 0,10 | 0,35 | −0,25 |

A seguir, na Tabela 4.7 apresenta-se as estimativas, erro padrão e valor-$p$ dos parâmetros para os modelos de regressão bivariado Simplex e Beta via cópulas FGM, Clayton e Frank, cujas estimativas são significativas ao nível de 1% de significância. Como descrito na Seção 3.1, as estimativas associadas ao parâmetro da média podem ser interpretados como a razão de chance, isto é, $\exp(-0,01 \times \widehat{\beta}_{11}) = \exp(3,64) \simeq 1,037$; $\exp(-0,01 \times \widehat{\beta}_{11}) = \exp(3,59) \simeq 1,0365$; $\exp(-0,01 \times \widehat{\beta}_{11}) = \exp(3,61) \simeq 1,0367$, ou seja, com a redução de 1% da Razão Dependência ($x$), a chance de redução do IDH por município diminui em 3,70%, 3,65% e 3,67%, enquanto que a redução de 1% da Razão Dependência ($x$), isto é, $\exp(-0,01 \times \widehat{\beta}_{12}) = \exp(-3,05) \simeq 0,9699$; $\exp(-0,01 \times \widehat{\beta}_{12}) = \exp(-3,34) \simeq ,0327$; $\exp(-0,01 \times \widehat{\beta}_{12}) = \exp(-2,94) \simeq 0,9712$, ou seja, a chance de aumentar o índice de vulnerabilidade social diminui em aproximadamente 3,05%, 3,27% e 2,89% com base no modelo de regressão Simplex bivariado pelas cópulas FGM, Clayton e Frank, respectivamente. Já pelo modelo de regressão Beta bivariado, as estimativas associadas ao parâmetro da média podem ser interpretados como a razão de chance, isto é, $\exp(-0,01 \times \widehat{\beta}_{11}) = \exp(3,49) \simeq 1,0355$; $\exp(-0,01 \times \widehat{\beta}_{11}) = \exp(3,57) \simeq 1,0364$; $\exp(-0,01 \times \widehat{\beta}_{11}) = \exp(3,63) \simeq 1,0369$, ou seja, com a redução de 1% da Razão Dependência ($x$), a chance de redução do IDH por municípios diminui em 3,55%, 3,64% e 3,69%, enquanto que a redução de 1% da Razão Dependência ($x$), isto é, $\exp(-0,01 \times \widehat{\beta}_{12}) = \exp(-2,922) \simeq 0,9712$; $\exp(-0,01 \times \widehat{\beta}_{12}) = \exp(-3,412) \simeq ,9664$; $\exp(-0,01 \times \widehat{\beta}_{12}) = \exp(-2,970) \simeq 0,9707$, ou seja, a chance de aumento do IVS por município diminui em 2,88%, 3,36% e 2,93% pelas cópulas FGM, Clayton e Frank, respectivamente. Em vista uma das cópulas, podemos notar que os modelos proposto através da cópula FGM ajustou melhor os dados, tanto para modelo de regressão Simplex bivariado quanto para o modelo de regressão Beta bivariado, uma vez que, apresentou valores mínimos de AIC e BIC.

[1]O Índice de Desenvolvimento Humano (IDH) municipal é uma medida composta de indicadores de três dimensões do desenvolvimento humano: longevidade, educação e renda. O índice varia entre 0 a 1; quanto mais próximo a 1, maior o desenvolvimento humano. Em suma, reflete o desenvolvimento econômico e a qualidade de vida.

[2]O Índice de Vulnerabilidade Social (IVS) é um índice que reflete situações de ausência ou insuficiência de alguns "ativos" (emprego, moradia, capital humano, capital social, entre outros) que, a princípio, deveriam estar à disposição de todos os brasileiros. O IVS varia também entre 0 e 1; quanto mais próximo de 1, maior é a vulnerabilidade social do município (Costa e Marguti, 2015).

[3]A Razão Dependência pressupõe que jovens e idosos de uma população são dependentes economicamente dos demais. Nesse sentido, é um indicador do contingente que é suportado pela população potencialmente produtiva (Vasconcelos Filho, 2016)

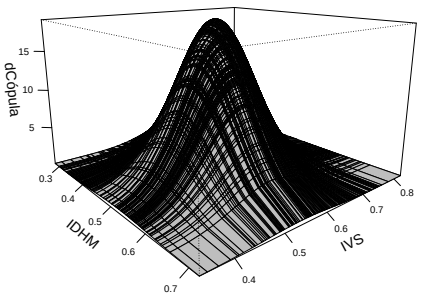
Com relação ao índice de desenvolvimento humano municipal e o índice de vulnerabilidade social por dispersão, espera-se que o parâmetro de dispersão da proporção de pobres e da taxa de mortalidade infantil diminui tanto pelo modelo de regressão Simplex quanto pelo modelo de regressão Beta, tendo em vistas as estimativas negativas dos parâmetros $\boldsymbol{\gamma}_{01}$ e $\boldsymbol{\gamma}_{02}$, conforme apresentado na Tabela 4.7. O parâmetro $\tau$ reflete o grau de associação entre as variáveis dependentes, no caso $\boldsymbol{y}_1$ e $\boldsymbol{y}_2$, dependendo unicamente da cópula utilizada e obtido a partir da relação com o parâmetro $\lambda$, conforme apresentado na Tabela 4.1. Com isso, vemos que pelas cópulas FGM e Frank ambos os modelos transmitem uma concordância ou associação negativa e relativamente fraca, enquanto que pela cópula Clayton ambos os modelos transmitem uma concordância positiva e relativamente fraca.

Nas Figuras 4.6 e 4.7 é apresentado os gráficos de superfície da densidade e curvas de nível dos modelos ajustados ao conjunto de dados. Os gráficos mostram que as variáveis IDH e IVS para o estado de Alagoas são inversamente proporcionais, isto é, quanto maior for o IDH municipal menor será o IVS municipal (o contrário também é valido). Na Figura 4.7, pode-se notar que o modelo ajusta bem os dados, tendo em vista que os dados estão circunscrito pelas curvas de nível, muito embora tenhamos observações em destaque fora das curvas, a saber: observações #63, #75 e #147 que correspondente aos municípios: Inhapi, Joaquim Gomes e Roteiro, respectivamente (municípios economicamente menos desenvolvidos com péssimos IDH e IVS municipal); observações #4, #184, #188 e #202 que correspondente aos municípios: Arapiraca, Barra de São Miguel, Maceió e Satuba, respectivamente (municípios economicamente mais desenvolvidos com bom IDH e IVS municipal).

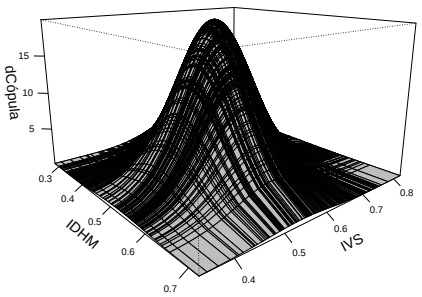
**Tabela 4.7:** *Estimativas, erro padrão e valor-p do modelo de regressão bivariado das distribuições Simplex e Beta.*

| Modelo | Distribuição | Cópula FGM | | | | Cópula Clayton | | | Cópula Frank | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Parâmetro | Estimativa | SE | $p$-valor | Estimativa | SE | $p$-valor | Estimativa | SE | $p$-valor |
| $y_1, y_2 \sim razdep$ | Simplex | $\beta_{01}$ | 2,396 | 0,079 | 0,000 | 2,423 | 0,095 | 0,000 | 2,374 | 0,082 | 0,000 |
| | | $\beta_{11}$ | −3,638 | 0,124 | 0,000 | −3,578 | 0,148 | 0,000 | −3,606 | 0,127 | 0,000 |
| | | $\beta_{02}$ | −1,757 | 0,118 | 0,000 | −1,847 | 0,162 | 0,000 | −1,680 | 0,122 | 0,000 |
| | | $\beta_{12}$ | 3,053 | 0,179 | 0,000 | 3,336 | 0,247 | 0,000 | 2,936 | 0,182 | 0,000 |
| | | $\gamma_{01}$ | −2,890 | 0,383 | 0,000 | −2,946 | 0,410 | 0,000 | −2,715 | 0,387 | 0,000 |
| | | $\gamma_{11}$ | 2,735 | 0,555 | 0,000 | 2,791 | 0,593 | 0,000 | 2,519 | 0,555 | 0,000 |
| | | $\gamma_{02}$ | −2,059 | 0,337 | 0,000 | −2,399 | 0,378 | 0,000 | −1,874 | 0,342 | 0,000 |
| | | $\gamma_{12}$ | 1,995 | 0,486 | 0,000 | 2,492 | 0,552 | 0,000 | 1,749 | 0,490 | 0,000 |
| | | $\lambda$ | −0,735 | 0,247 | 0,003 | 0,408 | 0,022 | 0,000 | −3,466 | 0,477 | 0,000 |
| $y_1, y_2 \sim razdep$ | Beta | $\beta_{01}$ | 2,287 | 0,079 | 0,000 | 2,420 | 0,101 | 0,000 | 2,390 | 0,087 | 0,000 |
| | | $\beta_{11}$ | −3,493 | 0,126 | 0,000 | −3,574 | 0,157 | 0,000 | −3,630 | 0,134 | 0,000 |
| | | $\beta_{02}$ | −1,678 | 0,113 | 0,000 | −1,893 | 0,166 | 0,000 | −1,700 | 0,127 | 0,000 |
| | | $\beta_{12}$ | 2,922 | 0,172 | 0,000 | 3,412 | 0,255 | 0,000 | 2,970 | 0,191 | 0,000 |
| | | $\gamma_{01}$ | −4,819 | 0,438 | 0,000 | −4,407 | 0,463 | 0,000 | −4,070 | 0,442 | 0,000 |
| | | $\gamma_{11}$ | 3,587 | 0,640 | 0,000 | 2,984 | 0,675 | 0,000 | 2,550 | 0,638 | 0,000 |
| | | $\gamma_{02}$ | −3,165 | 0,385 | 0,000 | −3,426 | 0,390 | 0,000 | −2,950 | 0,375 | 0,000 |
| | | $\gamma_{12}$ | 1,653 | 0,559 | 0,003 | 2,000 | 0,567 | 0,001 | 1,370 | 0,538 | 0,012 |
| | | $\lambda$ | −0,735 | 0,211 | 0,003 | 0,409 | 0,022 | 0,000 | −3,410 | 0,474 | 0,000 |

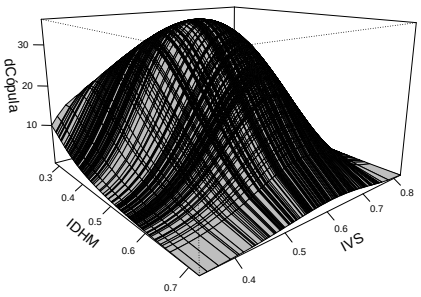
Simplex: [(Cópula FGM: AIC = 1345,7; BIC = 1423,4; $\tau = -0{,}1633$), (Cópula de Clayton: AIC = 1640,1; BIC = 1717,8; $\tau = 0{,}1696$), (Cópula de Frank: AIC = 1364,4; BIC = 1442,1; $\tau = -0{,}3439$)];Beta: [(Cópula FGM: AIC = 1345,1; BIC = 1422,9; $\tau = -0{,}1633$), (Cópula de Clayton: AIC = 1640,5; BIC = 1718,2; $\tau = 0{,}1697$), (Cópula de Frank: AIC = 1362,9; BIC = 1440,7; $\tau = -0{,}2948$)].



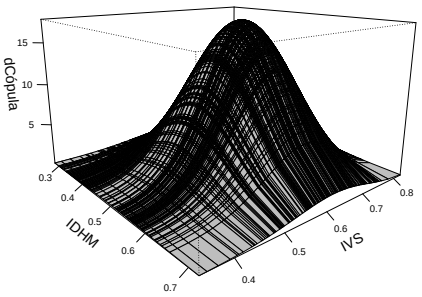

**(a)**
*FGM-Simplex:* $y_1, y_2 \sim razdep$



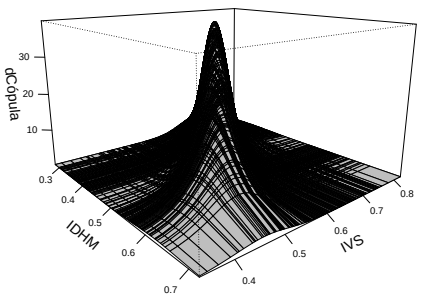

**(b)**
*FGM-Beta:* $y_1, y_2 \sim razdep$



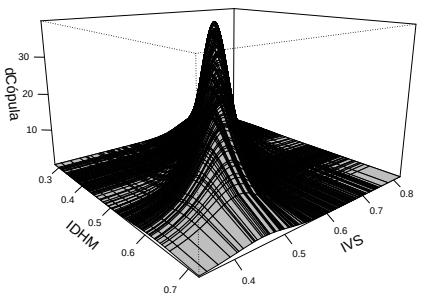

**(c)**
*Clayton-Simplex:* $y_1, y_2 \sim razdep$




**(d)**
*Clayton-Beta:* $y_1, y_2 \sim razdep$

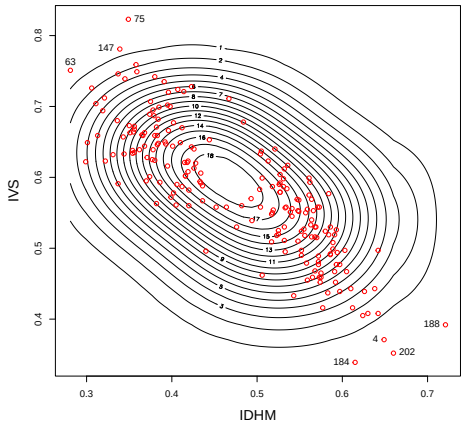



**(e)**
*Frank-Simplex:* $y_1, y_2 \sim razdep$

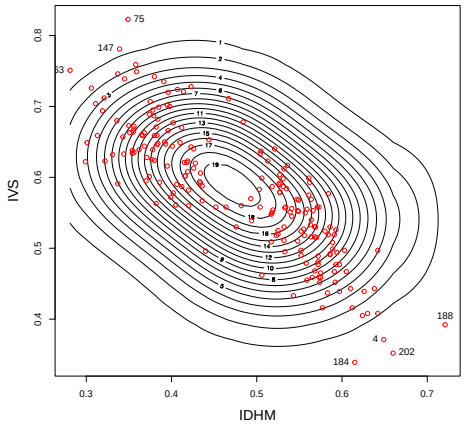



**(f)**
*Frank-Beta:* $y_1, y_2 \sim razdep$

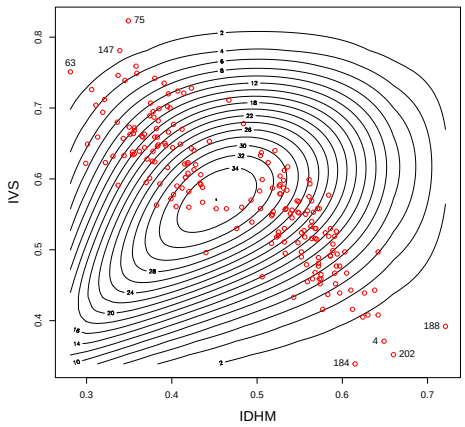
**Figura 4.6:** *Gráficos de superfícies para as variáveis índice de desenvolvimento humano e índice de vulnerabilidade social.*



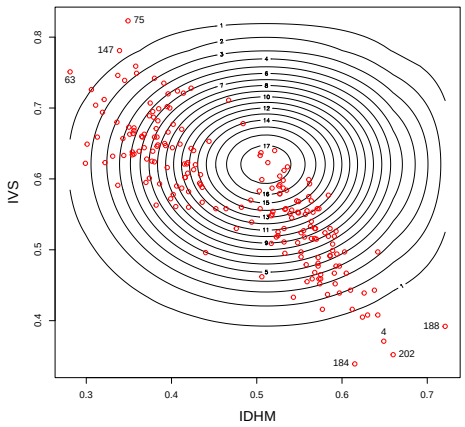

**(a)**
*FGM-Simplex:* $y_1, y_2 \sim razdep$



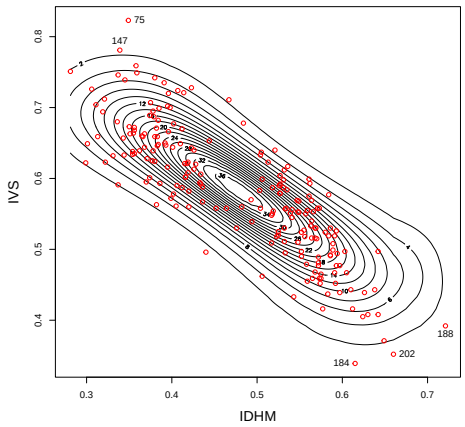

**(b)**
*FGM-Beta:* $y_1, y_2 \sim razdep$



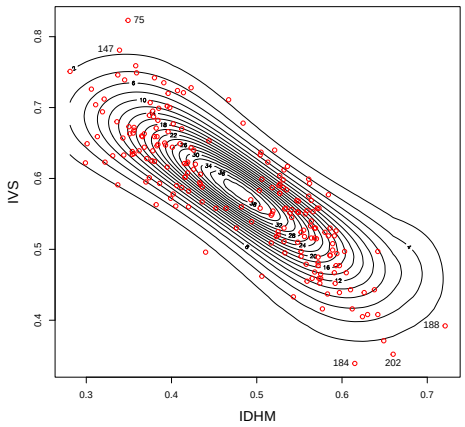

**(c)**
*Clayton-Simplex:* $y_1, y_2 \sim razdep$




**(d)**
*Clayton-Beta:* $y_1, y_2 \sim razdep$




**(e)**
*Frank-Simplex:* $y_1, y_2 \sim razdep$




**(f)**
*Frank-Beta:* $y_1, y_2 \sim razdep$

**Figura 4.7:** *Gráficos de curvas de nível para as variáveis índice de desenvolvimento humano e índice de vulnerabilidade social.*

# Capítulo 5

# Análise de Dagnóstico para o MRSB

Neste capítulo é apresentado uma etapa importante no ajuste do modelo de regressão, a análise de diagnóstico, que permite verificar possíveis afastamentos das suposições feitas para o modelo auxiliando na identificação de observações extremas com alguma interferência desproporcional nos resultados do ajuste.

## 5.1 Análise de resíduos

Na análise de diagnóstico, verificamos possíveis problemas nas suposições feitas para o modelo, como por exemplo, problemas na escolha da função de ligação, problemas na escolha das covariáveis, problemas de não-linearidade e até mesmo problemas com observações discrepantes que interferem nas estimativas dos parâmetros. Uma das técnicas de diagnóstico mais utilizada é a análise de resíduos. Os resíduos desempenham um papel importante na verificação da adequação do modelo e na identificação de "*outliers*" entre os valores ajustados a partir do modelo e os valores observados Silva (2019). Na literatura resultados de diagnóstico, para classe de modelos que contemplam o intervalo unitário (0,1), podem ser encontrados nos trabalhos de Ferrari e Cribari-Neto (2004), Espinheira *et al.* (2008), Lemonte e Bazán (2016), entre outros.

Neste trabalho de dissertação iremos utilizar os resíduos quantílicos como ferramenta de análise.

### 5.1.1 Resíduos quantílico

Apresentado por Dunn e Smyth (1996), os resíduos quantílicos aleatorizados pode ser definido como

$$r_{ji}^{q} = \Phi^{-1}\{F(y_{ji}; \mu_{ji}, \sigma_{ji}^2)\},\ i = 1, ..., n \text{ e } j = 1, 2, ..., p.$$

em que $\Phi^{-1}(\cdot)$ é a inversa da função de distribuição acumulada da distribuição normal padrão e $F(\cdot)$ é a função de distribuição acumulada do modelo de regressão Simplex bivariado.

Dado uma função de distribuição $F(y_{ji}; \mu_{ji}, \sigma_{ji}^2)$. Se $F(\cdot)$ é contínua, então $F(y_{ji}; \mu_{ji}, \sigma_{ji}^2)$ é uniformemente distribuída no intervalo unitário. Seja $Y$ uma variável aleatória contínua, o resíduo quantílico é definido como uma transformação da função de distribuição acumulada de $Y$, denotada aqui por $F(y)$. Ou seja, se $U = F(y)$, temos que $U$ é uma variável aleatória uniformemente distribuida em (0,1). Então, se $U = F(y)$, então $P(U \leq k) = P(F(Y) \leq k) = P(Y \leq F^{-1}(k)) = F(F^{-1}(k)) = k$, para todo $0 < k < 1$. Portanto, $U \sim U(0, 1)$ e, em conformidade com sua definição, quando o modelo for especificado corretamente, estes resíduos devem ser normalmente distribuídos em caso de adequabilidade da predição ou ajuste (Rigby e Stasinopoulos, 2005). Logo, a utilização destes resíduos é vantajosa por ele possui distribuição assintótica conhecida (normal padrão) independente da distribuição da variável resposta, sob o modelo postulado. Além disso, as distribuições marginais são conhecidas e os estimadores de máxima verossimilhança possui boas propriedades, como comprovado pelo estudo de simulação na Seção 4.3, tendo portanto aplicação direta em modelos de regressão e sua distribuição é fácil de ser checada por procedimentos usuais de teste de normalidade e avaliação gráficos.

Dentre as diversas etapas da análise de diagnóstico de um modelo, há o interesse em se avaliar a veracidade da hipótese referente à distribuição de probabilidade assumida para a variável resposta dadas as covariáveis. Assim, a construção de bandas de confiança, a qual denomina-se envelope, é implementado em um gráfico de probabilidade por meio de simulação construído a partir dos resíduos gerados pelo modelo ajustado, é sugerido (Paula, 2023).

Atkinson (1981), por sua vez, sugere a utilização de uma espécie de banda para a flutuação dos ponto, a qual denominou envelope simulado e que tem sido utilizado por diversos pesquisadores. Este procedimento consiste em determinar os limites do envelope realizando os seguinte passo:

1. Gerar $n$ observações $y_{i,1}$ considerando que o modelo ajustado é verdadeiro e guarda esses valores em $Y = (y_{1,1}, ..., y_{n,1})$;

2. Ajustamos $Y$ usando a matriz de variáveis preditoras $X$ e obtemos os resíduos $r_{i,1}$, para $i = 1, 2, ..., n$;

3. Repetir os passos 1) e 2) $m$ vezes. Logo, teremos $r_{i,j}$, para $i = 1, 2, ..., n$ e $j = 1, 2, ..., m$;

4. Colocamos cada grupo de $n$ resíduos em ordem crescente, obtemos assim $r_{[i],j}$, para $i = 1, 2, ..., n$ e $j = 1, 2, ..., m$;

5. Obtemos os limites $r_{[i],1} = min_j(r_{[i],j})$ e $r_{[i],m} = max_j(r_{[i],j})$. Assim, os limites correspondentes a $r_{[i]}$ serão dados por $r_{[i],[1]}$ e $r_{[i],[m]}$.

O objetivo desse procedimento de simulação é fornecer amostras de resíduos com mesma estrutura de covariância do modelo ajustado (Atkinson, 1981). Além disso, a sugestão de Atkinson é usar $m = 19$, de modo que a probabilidade do maior resíduo em um envelope particular exceder o limite superior seja em torno de 1/20. Analogamente, se considerarmos tal ocorrência para o limite inferior, podemos supor que a probabilidade do maior resíduo ser inferior a esse limite também seja por volta de 1/20, resultando em uma probabilidade aproximada de 0,1 do maior resíduo não estar entre as bandas do envelope (Fernandes, 2019).

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

Para a obtenção do gráfico de probabilidade normal, considere $r_i$, para $i = 1, 2, ..., n$ um resíduo qualquer. Colocamos no gráfico os pontos $(E(Z_{[i]}), r_i)$, em que $[i]$ representa o $i$-ésimo menor valor e $E(Z_{[i]})$ se refere aos valores esperados das estatísticas de ordem da normal padrão. Temos que

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

em que $\Phi$ é a função de distribuição acumulada da normal padrão.

Já para o gráfico meio normal probabilístico, tomamos o valor absoluto de $r_i$, com $i = 1, 2, ..., n$. Dispomos no gráfico os pontos $(E(|Z_{[i]}|), |r_i|)$, em que $[i]$ representa o $i$-ésimo menor valor e $E(|Z_{[i]}|)$ agora se refere aos valores esperados de $|Z_{[i]}|$, que obtemos fazendo

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

sendo que $\Phi$ também a função de distribuição acumulada da normal padrão (Fernandes, 2019).

## 5.2 Análise de influência global

Tal procedimento consiste numa medida de sensibilidade sob pequenas perturbações no modelo de acordo com a exclusão de caso (Cook, 1977). A exclusão de caso é uma abordagem comum para estudar o efeito de eliminar a $i$-ésima observação do conjunto de dados. A exclusão da $i$-ésima observação para o modelo 4.2 é dada por

$$\begin{cases} g(\mu_{rj}) = \sum_{l=0}^{k_1} X_{rl}\beta_{lj} \\ h(\sigma^2_{rj}) = \sum_{l=0}^{k2} Z_{rl}\gamma_{lj} \end{cases}, \quad r = 1, 2, ..., n; \quad r \neq i. \tag{5.1}$$

O subíndice "$(i)$" indica o conjunto de dados original com a $i$-ésima observação excluída. Para o modelo 5.1, a função do logaritmo da função de verossimilhança é denotada por $\ell_i(\theta)$. Seja $\widehat{\theta}_{(i)}$ o estimador de máxima verossimilhança de $\theta$, obtido da maximização de $\ell_i(\theta)$. Para avaliar a influência da $i$-ésima observação $\widehat{\theta}_{(i)}$, comparamos a diferença entre $\widehat{\theta}_{(i)}$ e $\widehat{\theta}$. Se a exclusão de uma observação influenciar seriamente uma estimativa, deve-se prestar mais atenção a essa observação específica. Portanto, se $\widehat{\theta}_{(i)}$ está longe de $\widehat{\theta}$, então este caso é considerado uma observação influente. Uma medida inicial de influência global baseada no teste de Wald é definida como a norma padronizada de $\widehat{\theta}_{(i)} - \widehat{\theta}$ (que é conhecida como a distância de Cook generalizada):

$$GD_i(\theta) = (\widehat{\theta}_{(i)} - \widehat{\theta})^\top [-\ddot{\ell}(\theta)](\widehat{\theta}_{(i)} - \widehat{\theta}). \tag{5.2}$$

Outra alternativa é avaliar os valores $GD_i(\beta)$ e $GD_i(\gamma)$, que revelam o impacto da $i$-ésima observação na estimativa de $\beta$ e $\gamma$, respectivamente. Outra medida popular da diferença entre $\widehat{\theta}_{(i)}$ e $\widehat{\theta}$ é o seguinte deslocamento de probabilidade:

$$LD_i(\theta) = 2\{\ell(\widehat{\theta}) - \ell(\widehat{\theta}_{(i)})\}. \tag{5.3}$$

Além disso, também podemos calcular $\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{j(i)}}$ $(j = 1, 2, ..., k)$, para calcular a diferença entre $\widehat{\boldsymbol{\beta}}$ e $\widehat{\boldsymbol{\beta}}_{\boldsymbol{(i)}}$, e $\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}_{\boldsymbol{j(i)}}$ $(j = 1, 2, ..., h)$, para calcular a diferença entre $\widehat{\gamma}$ e $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{(i)}}$. Outras medidas de influência global também são possíveis. Pode-se observar o comportamento de uma estatística de teste sob um esquema de exclusão de casos; tais estatísticas podem ser, por exemplo, o teste de Wald para variáveis explicativas (Rocha e Simas, 2011).

Para evitar o emprego da estimativa de modelo direto para todas as observações, podemos usar a seguinte aproximação de uma etapa para reduzir o número de modelos a serem ajustados:

$$\widehat{\theta}_{(i)} \approx \widehat{\theta} + \ddot{\ell}(\theta)^{-1} \left.\frac{\partial \ell(i)\theta}{\partial \theta}\right|_{\theta=\widehat{\theta}}.$$

Segundo Thomas e Cook (1989), $\partial \ell(i)\theta / \partial \theta$ é avaliado em $\theta = \widehat{\theta}$.

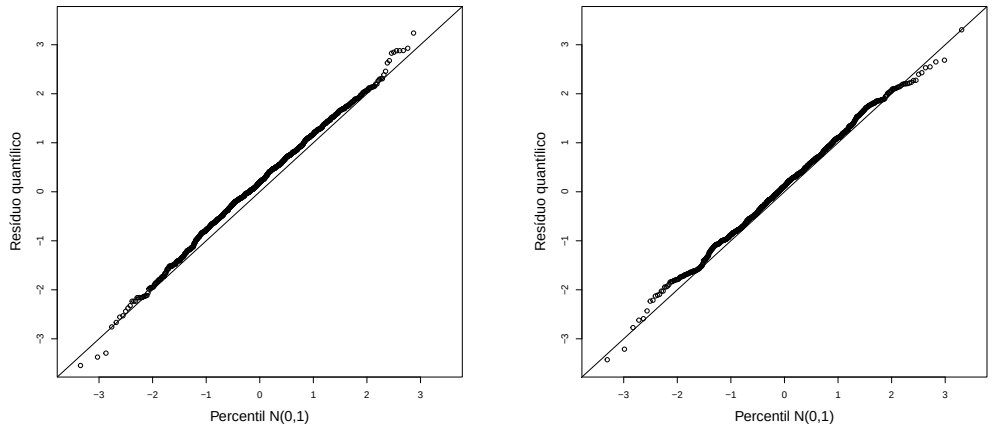
## 5.3 Aplicações

Nesta Seção apresenta-se a análise dos resíduos referente aos conjuntos de dados descritos em 4.4.1 e 4.4.2.

### 5.3.1 Aplicação I

Na Figura 5.1 é apresentado os gráficos dos resíduos quantílicos correspondente aos ajustes dos modelos de regressão bivariado Simplex e Beta via cópulas FGM, Clayton e Frank. Podemos concluir que não há evidências fortes contra a suposição de normalidade dos resíduos, ou seja, os modelos ajustam bem os dados. Podemos notar um melhor destaque para o modelo de regressão Simplex bivariado via cópula FGM pela Figura 5.1a comparado ao modelo de regressão Beta

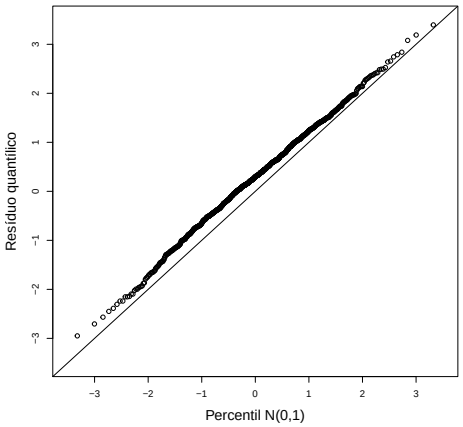
bivariado na Figura 5.1b e um melhor destaque para o modelo de regressão Beta bivariado via cópula Frank na Figura 5.1f comparado ao modelo de regressão Simplex bivariado na Figura 5.1e.



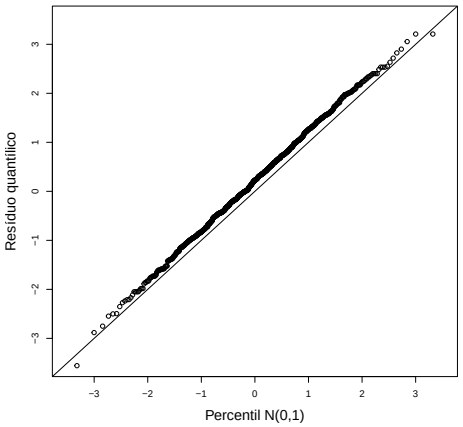

**(a)**
*Simplex-FGM:* $y_1, y_2 \sim IDHM$



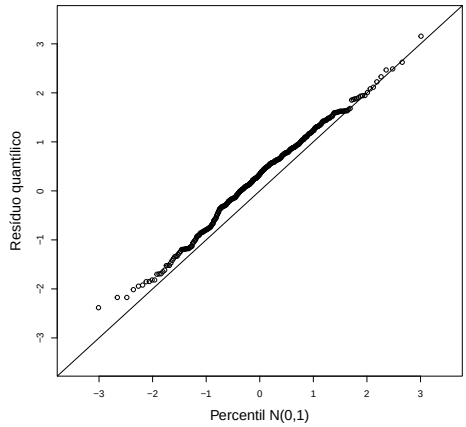

**(b)**
*Beta-FGM:* $y_1, y_2 \sim IDHM$



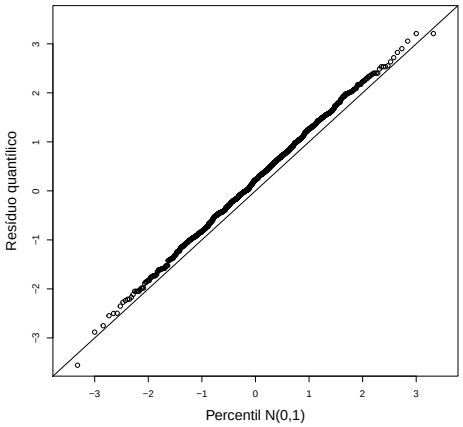

**(c)**
*Simplex-Clayton:* $y_1, y_2 \sim IDHM$




**(d)**
*Beta-Clayton:* $y_1, y_2 \sim IDHM$




**(e)**
*Simplex-Frank:* $y_1, y_2 \sim IDHM$



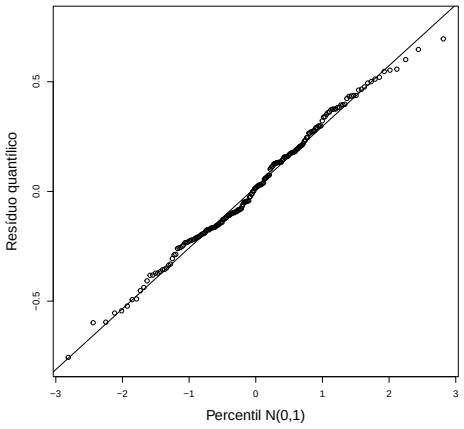

**(f)**
*Beta-Frank:* $y_1, y_2 \sim IDHM$

**Figura 5.1:** *Gráficos dos resíduos quantílicos - QQ-plot.*

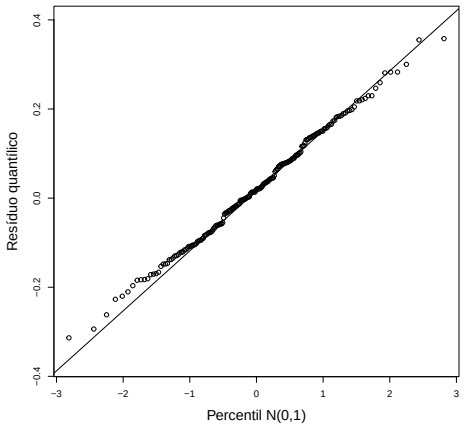
### 5.3.2 Aplicação II

Na Figura 5.2 é apresentado os gráficos dos resíduos quantílicos correspondente aos ajustes dos modelos de regressão bivariado das distribuição Simplex e Beta via cópulas FGM, Clayton e Frank. Podemos concluir que não há evidências fortes contra a suposição de normalidade dos resíduos, ou seja, todos os modelos ajustam bem os dados. O modelo de regressão Simplex bivariado transmite um melhor ajuste aos dados, tendo em vista que os resíduos apresentam-se linearmente mais próximo a reta comparado ao modelo de regressão Beta bivariado. Na Figura 5.3 nota-se que os resíduos positivos e negativos estão distribuídos sem algum padrão sistemático e que sua variabilidade é razoavelmente uniforme ao longo dos diferentes valores da variável explicativa, sugerindo que relativamente à suposição de homocedasticidade, com isso os modelos são adequado.

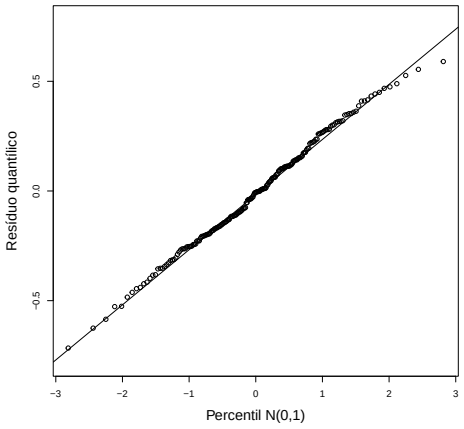
Em complemento a análise de diagnóstico, temos os gráficos de envelope simulado apresentado na Figura 5.4, onde evidenciamos a suposição de normalidade dos resíduos, apesar de termos pontos a partir de um dado momento sobrepondo a bando de confiança superior para os modelos de regressão bivariado Simplex e Beta via cópula Clayton. Os modelos de regressão bivariado Simplex e Beta via cópula FGM tendem a ajustar melhor os dados.



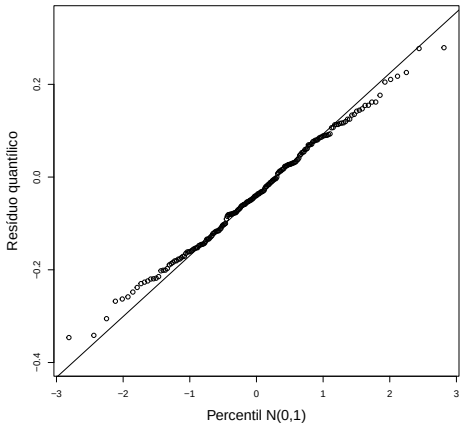

**(a)**
*FGM-Simplex:* $y_1, y_2 \sim razdep$



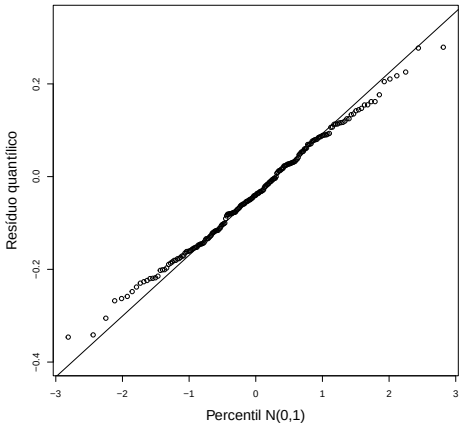

**(b)**
*FGM-Beta:* $y_1, y_2 \sim razdep$



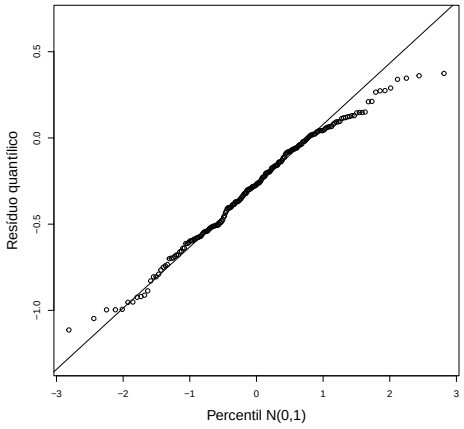

**(c)**
*Clayton-Simplex:* $y_1, y_2 \sim razdep$




**(d)**
*Clayton-Beta:* $y_1, y_2 \sim razdep$

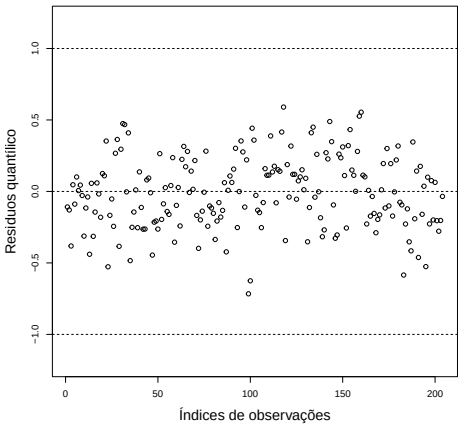



**(e)**
*Frank-Simplex:* $y_1, y_2 \sim razdep$

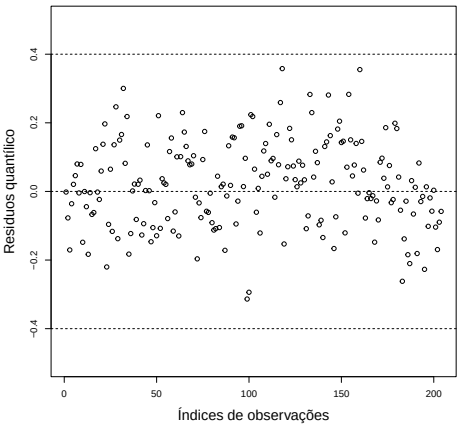



**(f)**
*Frank-Beta:* $y_1, y_2 \sim razdep$

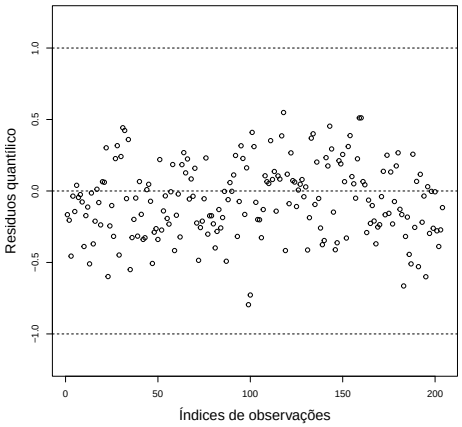
**Figura 5.2:** *Gráficos dos resíduos quantílicos - QQ-plot.*



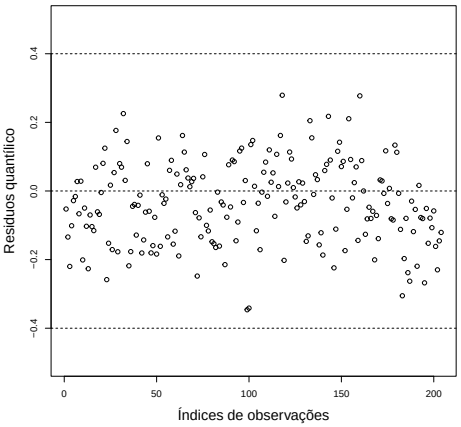

**(a)**
*FGM-Simplex:* $y_1, y_2 \sim razdep$



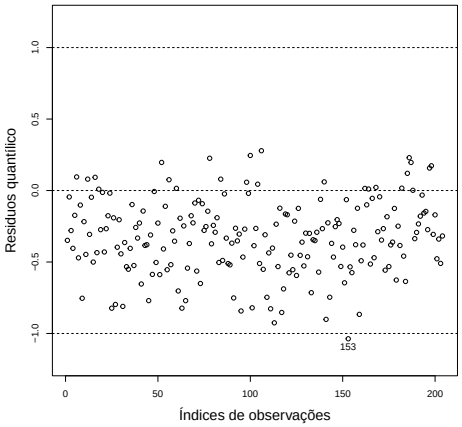

**(b)**
*FGM-Beta:* $y_1, y_2 \sim razdep$



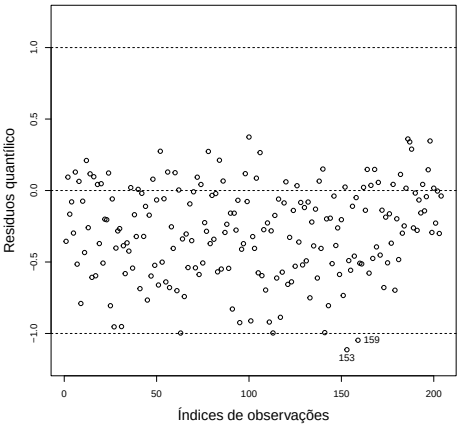

**(c)**
*Clayton-Simplex:* $y_1, y_2 \sim razdep$




**(d)**
*Clayton-Beta:* $y_1, y_2 \sim razdep$

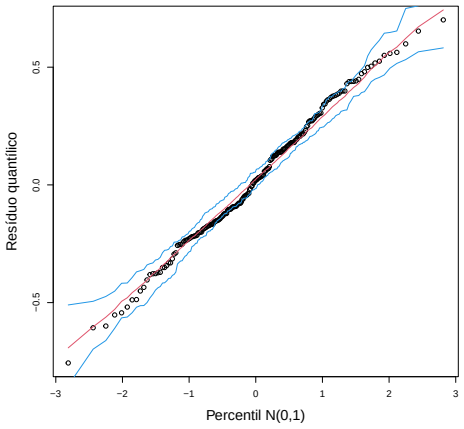



**(e)**
*Frank-Simplex:* $y_1, y_2 \sim razdep$

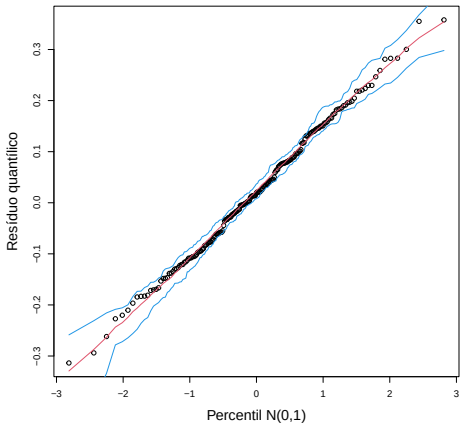



**(f)**
*Frank-Beta:* $y_1, y_2 \sim razdep$

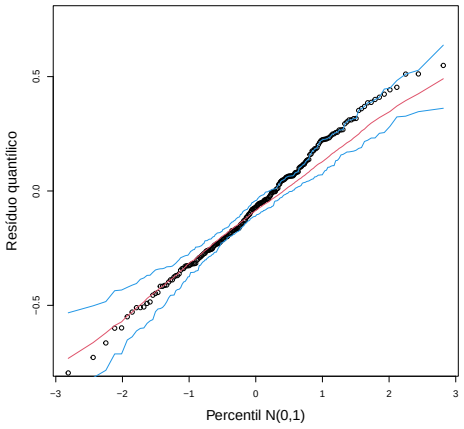
**Figura 5.3:** *Gráficos dos resíduos quantílicos x índices de observação.*



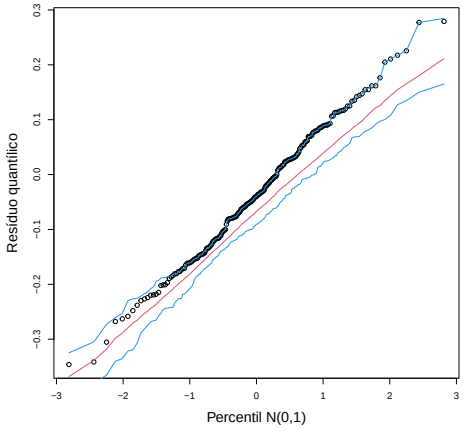

**(a)**
*FGM-Simplex:* $y_1, y_2 \sim razdep$



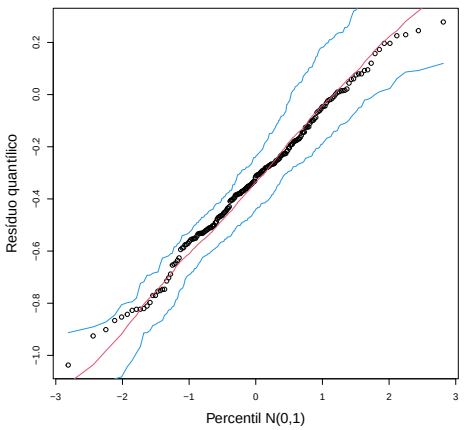

**(b)**
*FGM-Beta:* $y_1, y_2 \sim razdep$



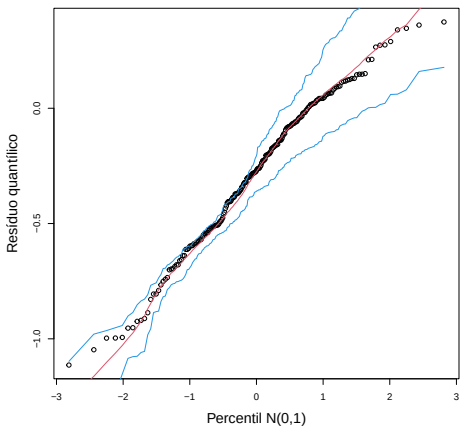

**(c)**
*Clayton-Simplex:* $y_1, y_2 \sim razdep$




**(d)**
*Clayton-Beta:* $y_1, y_2 \sim razdep$




**(e)**
*Frank-Simplex:* $y_1, y_2 \sim razdep$



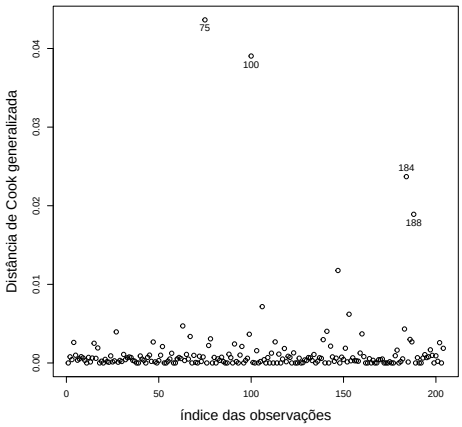

**(f)**
*Frank-Beta:* $y_1, y_2 \sim razdep$

**Figura 5.4:** *Gráficos de envelope simulado.*

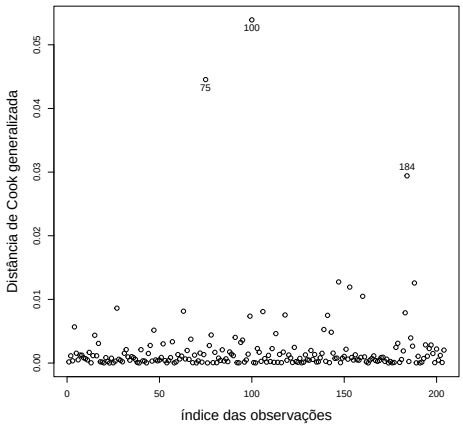
As Figuras 5.5 e 5.6 apresentam os gráficos das medidas de influência global, distancia de Cook generalizada $GD_i(\boldsymbol{\theta})$ e afastamento da verossimilhança $LD_i(\boldsymbol{\theta})$, contra o índice das observações. Estes gráficos mostram que algumas observações apresentam características distintas das demais observações, ou seja, indicam ser possíveis observações influentes, a saber: #75 (município Joaquim Gomes; IDH= $0,34$, IVS= $0,82$ e $x = 0,80$), #100 (município Minador do Negrão; IDH= $0,56$, IVS= $0,53$ e RD= $0,52$) e #184 (município Barra de São Miguel, IDH= $0,61$, IVS= $0,34$ e RD= $0,51$). Observa-se que os municípios Joaquim Gomes e Barra de São Miguel apresentam maior e menor IVS, respectivamente. Para avaliar a influência que tais observações têm sobre as estimativas dos parâmetros de regressão e dispersão, vamos utilizar a seguinte medida

$$\text{Variação Percentual} = \frac{\widehat{\theta}^* - \widehat{\theta}}{\widehat{\theta}} \times 100\%,$$

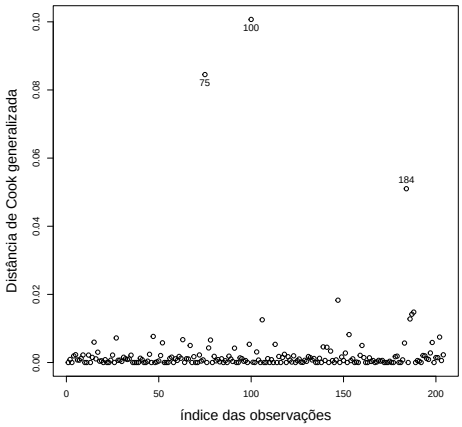
em que $\widehat{\theta}^*$ é a estimativa do parâmetro $\theta$ sem as observações distintas das demais e $\widehat{\theta}$ é a estimativa de $\theta$ com todas as observações no modelo. Essa medida verifica, descritivamente, o quanto variam as estimativas dos parâmetros na ausência das observações com comportamento distintos das demais.



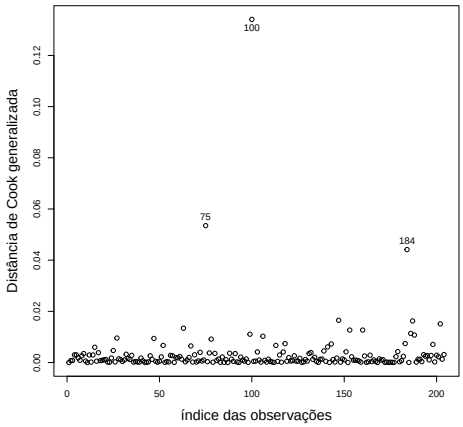

**(a)**
*FGM-Simplex:* $y_1, y_2 \sim razdep$



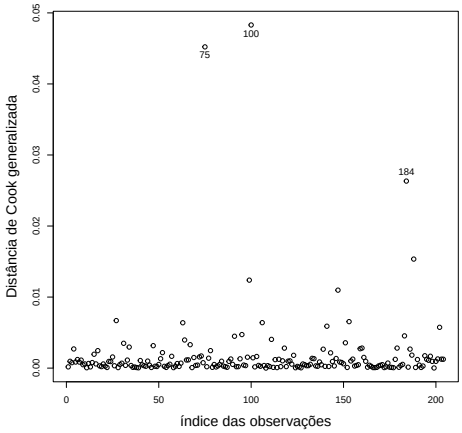

**(b)**
*FGM-Beta:* $y_1, y_2 \sim razdep$



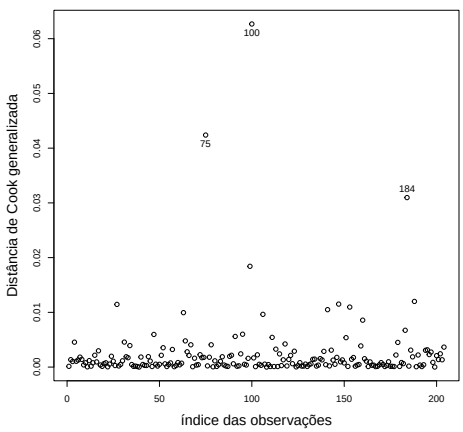

**(c)**
*Clayton-Simplex:* $y_1, y_2 \sim razdep$




**(d)**
*Clayton-Beta:* $y_1, y_2 \sim razdep$

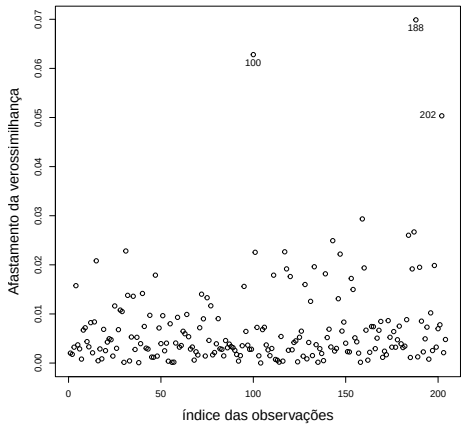



**(e)**
*Frank-Simplex:* $y_1, y_2 \sim razdep$

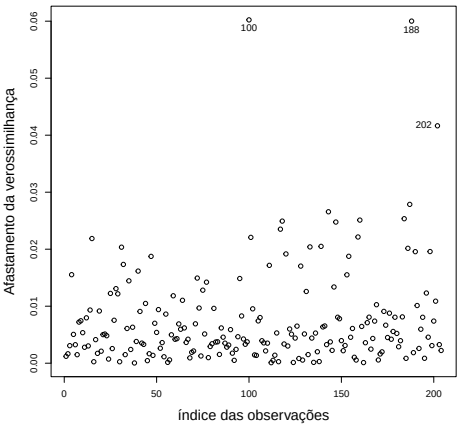



**(f)**
*Frank-Beta:* $y_1, y_2 \sim razdep$

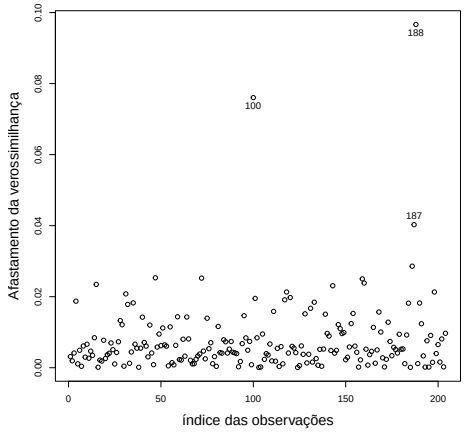
**Figura 5.5:** *Gráficos distância de Cook generalizada.*



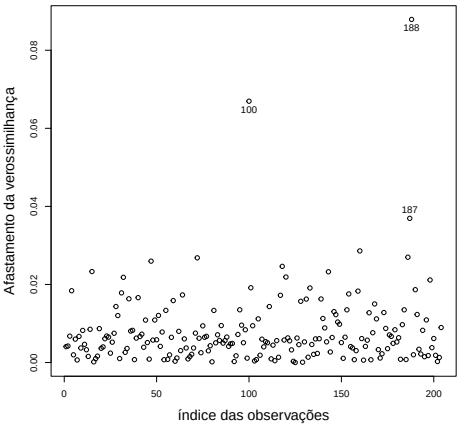

**(a)**
*FGM-Simplex:* $y_1, y_2 \sim razdep$



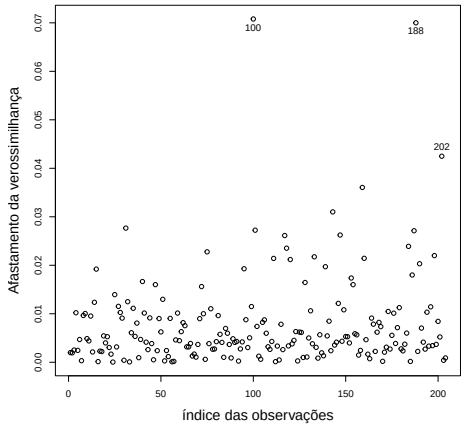

**(b)**
*FGM-Beta:* $y_1, y_2 \sim razdep$



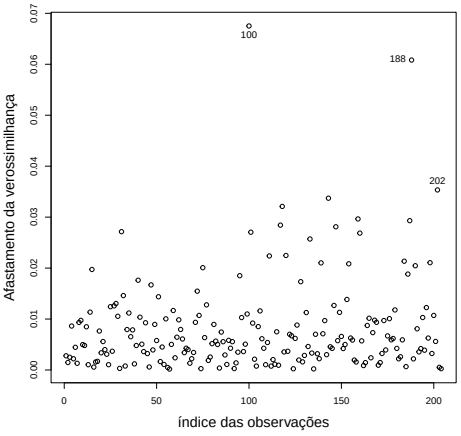

**(c)**
*Clayton-Simplex:* $y_1, y_2 \sim razdep$




**(d)**
*Clayton-Beta:* $y_1, y_2 \sim razdep$




**(e)**
*Frank-Simplex:* $y_1, y_2 \sim razdep$




**(f)**
*Frank-Beta:* $y_1, y_2 \sim razdep$

**Figura 5.6:** *Gráficos afastamento da verossimilhança.*

Ajustamos novamente os modelo de regressão bivariado Simplex e Beta bivariado, excluindo as observações #75, #100 e #184 uma a uma e posteriormente todas. Nas Tabelas 5.1, 5.2 e 5.3 são apresentadas as variações percentuais dos parâmetros desses ajustes após a exclusão destas observações para avaliar a influência nas estimativas dos parâmetros no ajuste do modelo de regressão bivariado Simplex e Beta. Como temos $n = 204$ observações, isto implica que cada observação deve influenciar em 0,49% nas estimativas dos parâmetros. Pela Tabela 5.1 podemos notar que não há grandes alterações nas estimativas dos parâmetros $\boldsymbol{\beta}_{01}$ e $\boldsymbol{\beta}_{11}$ sem as observações #75 e #184; $\boldsymbol{\gamma}_{01}$ e $\boldsymbol{\gamma}_{11}$ sem a observação #184, não havendo também grandes alteração em $\boldsymbol{\gamma}_{01}$ sem a obserção #75 e $\lambda$ sem a observação #75 em ambos os modelos de regressão bivariado Simplex e Beta via cópula FGM.

Pela Tabela 5.2, podemos notar que não há grandes alterações nas estimativas dos parâmetros $\boldsymbol{\beta}_{01}$, $\boldsymbol{\beta}_{11}$ com a exclusão das observações #75 e #184 nos modelos de regressão bivariado Simplex e Beta; $\boldsymbol{\beta}_{02}$ e $\boldsymbol{\beta}_{12}$ com a exclusão da observação #184 e para o $\lambda$ com a exclusões das observações #75, #100, 84 e com a exclusão de todas três observações supracitadas, para os modelos de regressão bivariado Simplex e Beta via cópula Clayton.

Na Tabela 5.3 notamos que não há grandes alterações nas estimativas dos parâmetros $\boldsymbol{\beta}_{01}$, $\boldsymbol{\beta}_{11}$ na exclusão das observação #184 no modelo de regressão bivariado Simplex via cópula Frank e na exclusão das #75 e #184 no modelo de regressão Beta bivariado via cópula Frank e dos parâmetrso $\boldsymbol{\gamma}_{01}$, $\boldsymbol{\gamma}_{11}$ na exclusão das das observações #75 para o modelo de regressão Simplex bivariado via cópula Frank.

De modo geral, mesmo após a exclusão das observações supracitadas e re-estimação dos parâmetros para os modelos de regressão bivariado Simplex e Beta via cópulas FGM, Clayton e Frank, as conclusões inferenciais não mudaram.

**Tabela 5.1:** *Variação percentual das estimativas dos parâmetros do modelo de regressão Simplex e Beta bivariado via cópula FGM.*

| Parâmetro | Modelo Simplex | | | | Modelo Beta | | | |
|---|---|---|---|---|---|---|---|---|
| | Sem obs #75 | Sem obs #100 | Sem obs #184 | Sem as obs #75, #100 e #184 | Sem obs #75 | Sem obs #100 | Sem obs #184 | Sem as obs #75, #100 e #184 |
| $\beta_{01}$ | -0,23% | 1,45% | 0,43% | 1,76% | -0,23% | 1,35% | 0,42% | 1,63% |
| $\beta_{11}$ | -0,24% | 1,39% | 0,42% | 1,65% | -0,24% | 1,28% | 0,40% | 1,52% |
| $\beta_{02}$ | -3,00% | 2,49% | -0,89% | -1,36% | -2,76% | 2,32% | -1,24% | -1,46% |
| $\beta_{12}$ | -2,75% | 2,05% | -0,60% | -1,26% | -2,53% | 1,92% | -0,93% | -1,35% |
| $\gamma_{01}$ | 0,44% | 5,67% | -0,09% | 6,40% | 0,32% | 4,05% | -0,16% | 4,53% |
| $\gamma_{11}$ | 0,78% | 8,30% | -0,10% | 9,55% | 0,84% | 8,62% | -0,31% | 9,81% |
| $\gamma_{02}$ | -8,62% | 4,41% | 7,80% | 3,36% | -5,26% | 2,77% | 4,77% | 2,51% |
| $\gamma_{12}$ | -14,37% | 6,39% | 10,53% | 2,25% | -16,48% | 7,51% | 12,11% | 3,88% |
| $\lambda$ | -0,22% | -2,52% | 1,44% | -0,72% | -0,22% | -2,52% | 1,44% | -0,72% |

**Tabela 5.2:** *Variação percentual das estimativas dos parâmetros do modelo de regressão Simplex e Beta bivariado via cópula Clayton.*

| Parâmetro | Modelo Simplex | | | | Modelo Beta | | | |
|---|---|---|---|---|---|---|---|---|
| | Sem obs #75 | Sem obs #100 | Sem obs #184 | Sem as obs #75, #100 e #184 | Sem obs #75 | Sem obs #100 | Sem obs #184 | Sem as obs #75, #100 e #184 |
| $\beta_{01}$ | 0,30% | 1,60% | -0,35% | 1,49% | 0,19% | 1,38% | -0,28% | 1,36% |
| $\beta_{11}$ | 0,31% | 1,63% | -0,39% | 1,49% | 0,20% | 1,40% | -0,31% | 1,36% |
| $\beta_{02}$ | -4,80% | 1,65% | -0,14% | -2,48% | -3,68% | 1,70% | -0,59% | -2,66% |
| $\beta_{12}$ | -4,26% | 1,29% | 0,00% | -2,28% | -3,31% | 1,35% | -0,40% | -2,43% |
| $\gamma_{01}$ | 1,51% | 8,36% | -2,18% | 6,78% | 0,65% | 5,67% | -1,78% | 4,50% |
| $\gamma_{11}$ | 2,48% | 12,27% | -3,17% | 10,18% | 1,57% | 11,70% | -3,65% | 9,50% |
| $\gamma_{02}$ | -8,63% | 4,10% | 6,98% | 3,86% | -4,42% | 3,02% | 4,57% | 2,90% |
| $\gamma_{12}$ | -13,22% | 5,41% | 8,77% | 3,00% | -12,14% | 7,10% | 10,21% | 4,56% |
| $\lambda$ | -0,15% | 0,17% | 0,02% | 0,02% | -0,15% | 0,15% | 0,00% | 0,00% |

**Tabela 5.3:** *Variação percentual das estimativas dos parâmetros do modelo de regressão Simplex e Beta bivariado via cópula Frank.*

| Parâmetro | Modelo Simplex | | | | Modelo Beta | | | |
|---|---|---|---|---|---|---|---|---|
| | Sem obs #75 | Sem obs #100 | Sem obs #184 | Sem as obs #75, #100 e #184 | Sem obs #75 | Sem obs #100 | Sem obs #184 | Sem as obs #75, #100 e #184 |
| $\beta_{01}$ | -0,52% | 1,50% | 0,50% | 1,62% | -0,41% | 1,43% | 0,46% | 1,52% |
| $\beta_{11}$ | -0,54% | 1,41% | 0,49% | 1,50% | -0,42% | 1,35% | 0,45% | 1,40% |
| $\beta_{02}$ | -3,10% | 2,63% | -0,55% | -0,83% | -2,75% | 2,58% | -0,76% | -0,89% |
| $\beta_{12}$ | -2,79% | 2,14% | -0,34% | -0,83% | -2,48% | 2,11% | -0,54% | -0,88% |
| $\gamma_{01}$ | -0,27% | 5,29% | 0,75% | 6,39% | 0,27% | 3,79% | 0,45% | 4,51% |
| $\gamma_{11}$ | -0,33% | 7,76% | 1,19% | 9,59% | 0,77% | 8,29% | 1,08% | 10,13% |
| $\gamma_{02}$ | -9,69% | 4,31% | 8,61% | 3,05% | -5,51% | 2,76% | 5,29% | 2,35% |
| $\gamma_{12}$ | -16,70% | 6,40% | 12,18% | 1,69% | -19,04% | 8,20% | 15,10% | 3,75% |
| $\lambda$ | 0,86% | -2,93% | 1,45% | -0,42% | 1,14% | -2,56% | 1,94% | 0,33% |

# Capítulo 6

# Conclusões e Perspectivas Futuras

## 6.1 Considerações Finais

Nesta dissertação, damos uma introdução a Teoria dos Modelos de Dispersão; apresentamos a distribuição Simplex e suas propriedades; modelo de regressão Simplex univariado, função escore e matriz de informação de Fisher. Por conseguinte, adentramos no escopo maior desta dissertação definindo função cópula, medida de dependência, distribuição conjunta, modelo de regressão Simplex multivariado via função cópula com destaque nos modelos de regressão Simplex bivariado via função cópula Farlie-Gumbel-Morgenstern (FGM), Clayton e Frank. Apresentamos também como medida de diagnóstico os resíduos quantílicos aleatorizados, gráfico de envelope simulado e análise de influência global: distância de Cook generalizada e afastamento da verossimilhança.

Apresentamos um estudo de simulação via Monte Carlo (MC) para avaliar o comportamento assintóticas dos estimadores de máxima verossimilhança para o modelo de regressão Simplex bivariado considerando 3 cenários, cuja propriedade de consistência dos estimadores são confirmadas.

Duas aplicações a dados reais são apresentadas para o modelo de regressão Simplex bivariado em paralelo ao modelo de regressão Beta bivariado. Na primeira aplicação o modelo de regressão bivariado Simplex e Beta via cópula FGM apresentaram um melhor ajuste aos dados, com um destaque para o MRSB tendo em vista a análise dos resíduos e a característica gráfica da distribuição conjunta dos dados captada pela distribuição Simplex e não identificada pela distribuição Beta. Enquanto que quando considerado a cópula Frank o modelo de regressão Beta bivariado apresentou um melhor destaque.

Analogamente, na segunda aplicação o modelo de regressão bivariado Simplex e Beta via cópula FGM apresentaram um melhor ajuste aos dados. Contudo, pela análise dos resíduos o MRSB torna-se o mais adequado.

Em síntese, o presente trabalho concentrou-se na construção de modelos de regressão Simplex multivariado via função cópulas, um assunto inédito na classe dos modelos de regressão limitados no intervalo unitário aberto (0,1), sendo sua eficiência e importância apresentadas por meio de aplicações e estudo de simulação.

## 6.2 Sugestões para Pesquisas Futuras

- Como possíveis trabalhos futuros propormos:
  - Dada a complexidade do modelo de regressão Simplex multivariado para mais de duas variáveis respostas, utilizar o método de verossimilhança composta (pareada), condicionando o modelo de regressão Simplex multivariado a soma de casos particulares de modelos de regressão Simplex bivariado;
  - Investigar a distribuição empírica dos resíduos via simulação de Monte Carlo;
  - Avaliar um modelo de regressão Simplex multivariado com erro de medida nas variáveis;
  - Introduzir pertubações e realizar estudo de influência local;

– Otimizar os códigos com implementação na linguagem R;
– etc.

# Referências Bibliográficas


Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**(6), 716–723. 9

Anjos, U., Ferreira, F., Kolev, N., and Mendes, B. (2004). Modelando dependências via cópulas. *São Paulo: Associação Brasileira de Estatística*. 21, 22, 23

Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika*, **68**(1), 13–20. 50

Barndorff-Nielsen, O. E. and Jørgensen, B. (1991). Some parametric models on the simplex. *Journal of multivariate analysis*, **39**(1), 106–116. 3, 5

Burnham, K. P. and Anderson, D. R. (2004). Model selection and multimodel inference. *A practical information-theoretic approach*, **2**. 9

Carrasco, J. M. and Reid, N. (2021). Simplex regression models with measurement error. *Communications in Statistics-Simulation and Computation*, **50**(11), 3420–3435. 1

Cepeda-Cuervo, E., Achcar, J. A., and Lopera, L. G. (2014). Bivariate beta regression models: joint modeling of the mean, dispersion and association parameters. *Journal of Applied statistics*, **41**(3), 677–687. 1

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, **19**(1), 15–18. 51

Cordeiro, G. M., Labouriau, R., and Botter, D. A. (2021). An introduction to Bent Jorgensen ideas. *Brazilian journal of Probability and Statistics*, **35**(1), 2–20. 3

Costa, M. A. and Marguti, B. O. E. (2015). Atlas da vulnerabilidade social nos municípios brasileiros. 43

Cox, D. R. and Snell, E. J. (1989). *Analysis of binary data*, volume 32. CRC press. 9

de Oliveira, M. S. (2004). *Um Modelo de Regressão Beta: teoria e aplicações*. Ph.D. thesis, Instituto de Matemática e Estatística da Universidade de São Paulo, 12/04/2004. 14

Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and graphical statistics*, **5**(3), 236–244. 49

Espinheira, P. L. and de Oliveira Silva, A. (2019). Residual and influence analysis to a general class of simplex regression. *Test*, pages 1–30. 1

Espinheira, P. L., Ferrari, S. L., and Cribari-Neto, F. (2008). On beta regression residuals. *Journal of Applied Statistics*, **35**(4), 407–419. 1, 49

Fernandes, V. V. (2019). Contribuições sobre o envelope simulado na análise de diagnóstico em modelos de regressão. 50

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics*, **31**(7), 799–815. 1, 49

Gutiérrez, J. M. and Hernández, F. (2020). maxlogl: A general computational procedure for maximum likelihood estimation in r. 2

Henningsen, A. and Toomet, O. (2011). maxlik: A package for maximum likelihood estimation in r. *Computational Statistics*, **26**(3), 443–458. 2

Hoeffding, W. (1940). Masstabinvariante korrelationstheorie. *Schriften des Mathematischen Instituts und Instituts fur Angewandte Mathematik der Universitat Berlin*, **5**, 181–233. 21

Hofert, M., Kojadinovic, I., Maechler, M., Yan, J., Maechler, M. M., and Suggests, M. (2014). Package copula. *http://ie.archive.ubuntu.com/disk1/disk1/cran.r-project.org/web/packages/copula/copula.pdf*. 2

Ipea (2023). Índice de vulnerabilidade social. http://ivs.ipea.gov.br/index.php/pt/. acessado em 07/05/2023. 43

Johnson, M. E. (1987). *Multivariate statistical simulation: A guide to selecting and generating continuous multivariate distributions*, volume 192. John Wiley & Sons. 31

Jorgensen, B. (1997). *The theory of dispersion models*. CRC Press. 1, 3, 4, 5

Koochemeshkian, P., Manouchehri, N., and Bouguila, N. (2020). Bivariate beta regression model and its medical applications. In *2020 International Symposium on Networks, Computers and Communications (ISNCC)*, pages 1–5. IEEE. 1

Lemonte, A. J. and Bazán, J. L. (2016). New class of johnson distributions and its associated regression model for rates and proportions. *Biometrical Journal*, **58**(4), 727–746. 49

Liu, P., Yuen, K. C., Wu, L.-C., Tian, G.-L., and Li, T. (2020). Zero-one-inflated simplex regression models for the analysis of continuous proportion data. *Statistics and Its Interface*, **13**(2), 193–208. 1

Miyashiro, E. S. (2008). *Modelos de regressão beta e simplex para análise de proporções*. Ph.D. thesis, Universidade de São Paulo. 1

Moura, A. R. *et al.* (2021). Critérios de seleção de modelos: um estudo comparativo. 9

Municipal, A. d. D. H. (2023). Atlas do desenvolvimento humano no brasil 2013. *Ranking. Disponível em: http://www.atlasbrasil.org.br/2023/pt/ranking/. Acesso em*. 9, 17, 39

Nadarajah, S., Afuecheta, E., and Chan, S. (2017). A compendium of copulas. *Statistica*, **77**(4), 279–328. 21

Nagelkerke, N. J. *et al.* (1991). A note on a general definition of the coefficient of determination. *Biometrika*, **78**(3), 691–692. 9

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, **135**(3), 370–384. 5

Nelsen, R. B. (2006). An introduction to copulas, 2nd-edition. 21

Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, **9**(4), 325–346. 1

Paula, G. A. (2023). *Modelos de regressão: com apoio computacional*. IME-USP São Paulo. 50

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 2

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**(3), 507–554. 49

Rocha, A. V. and Simas, A. B. (2011). Influence diagnostics in a general class of beta regression models. *Test*, **20**(1), 95–119. 1, 51

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464. 9

Silva, A. d. O. (2015). *Regressão simplex não linear: inferência e diagnóstico*. Master's thesis, Universidade Federal de Pernambuco. 1

Silva, F. C. d. (2016). *Teste de diagnóstico baseado em influência local aplicado ao modelo de regressão simplex*. Master's thesis, Universidade Federal de Pernambuco. 6

Silva, L. C. M. d. (2019). Diagnóstico em modelos de regressão simplex. 1, 49

Simas, A. B., Barreto-Souza, W., and Rocha, A. V. (2010). Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis*, **54**(2), 348–366. 1

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, **8**, 229–231. 21

Song, P. X.-K. and Tan, M. (2000). Marginal models for longitudinal continuous proportional data. *Biometrics*, **56**(2), 496–502. 6

Souza, D. (2011). *Regressao beta multivariada com aplicaçoes em pequenas áreas*. Ph.D. thesis, Tese de doutorado do programa de pós-graduaçao em estatıstica. 22, 23

Souza, D. F. and Moura, F. A. (2016). Multivariate beta regression with application in small area estimation. *Journal of Official Statistics*, **32**(3), 747–768. 1

Stasinopoulos, D. M. and Rigby, R. A. (2008). Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, **23**, 1–46. 2

Thomas, W. and Cook, R. D. (1989). Assessing influence on regression coefficients in generalized linear models. *Biometrika*, **76**(4), 741–749. 51

Vasconcellos, K. L. and Cribari-Neto, F. (2005). Improved maximum likelihood estimation in a new class of beta regression models. *Brazilian Journal of Probability and Statistics*, pages 13–31. 1

Vasconcelos Filho, R. E. d. (2016). Uma análise dos indicadores demográficos do estado de alagoas. 43

Veall, M. R. and Zimmermann, K. F. (1996). Pseudo-r2 measures for some common limited dependent variable models. *Journal of Economic surveys*, **10**(3), 241–259. 9

Zeileis, A., Cribari-Neto, F., Gruen, B., Kosmidis, I., Simas, A. B., Rocha, A. V., and Zeileis, M. A. (2016). Package betareg. *R package*, **3**(2). 2

Zhang, P., Qiu, Z., and Shi, C. (2016). simplexreg: An r package for regression analysis of proportional data using the simplex distribution. *Journal of Statistical Software*, **71**, 1–21. 2

# Apêndice

## Códigos na Linguagem de Programação R

### Estimação dos parâmetros

```
#############################################################################
# Cópula FGM
#-----------------------------
dDSB_FGM <- function(y1,y2,x,b01,b11,b02,b12,g01,g11,g02,g12,lambda){
  mu1 <- exp(b01+b11*x)/(1+exp(b01+b11*x))
  sigma1 <- exp(g01+g11*x)

  mu2 <- exp(b02+b12*x)/(1+exp(b02+b12*x))
  sigma2 <- exp(g02+g12*x)

  # Densidade da variável y1
  dsim1 <- dSIMPLEX(y1,mu1,sigma1)

  # Densidade da variável y2
  dsim2 <- dSIMPLEX(y2,mu2,sigma2)

  # Cópula
  u <- pSIMPLEX(y1,mu1,sigma1)
  v <- pSIMPLEX(y2,mu2,sigma2)
  cop_fgm <- 1+(lambda*(1-2*u)*(1-2*v))
  dsim_f <- dsim1*dsim2*cop_fgm
  return(dsim_f)
}
#-----------------------------
lvero_FGM <- function(theta,y1,y2,x){
  b01 <- theta[1]
  b11 <- theta[2]
  b02 <- theta[3]
  b12 <- theta[4]
  g01 <- theta[5]
  g11 <- theta[6]
  g02 <- theta[7]
  g12 <- theta[8]
  lambda <- theta[9]
  func <- dDSB_FGM(y1,y2,x,b01,b11,b02,b12,g01,g11,g02,g12,lambda)
  lfunc <- -sum(log(func))
  #print(lfunc)
  return(lfunc)
```

```
}
#------------------------------
mle.DBS_FGM <- function(y1,y2,x){
  op <- optim(par=chute,lvero_FGM,method="L-BFGS-B",y1=y1,y2=y2,x=x,
    hessian=T,control=list(maxit=200),lower=c(rep(-Inf,8),-1),
        upper=c(rep(Inf,8),1))

  fisher.information <- solve(op$hessian)
  standard.deviance <- sqrt(diag(fisher.information))
  t.simplex <- op$par/standard.deviance
  pvalue <- round(2*(1-pt(abs(t.simplex), length(x)-length(op$par))),3)
  results <- cbind(op$par,standard.deviance,t.simplex,p-value)
  colnames(results) <- c("parâmetros", "desvio-padrão", "estatística t",
                                                          "p-valor")
  rownames(results) <- c("b01","b11","b02","b12","g01","g11","g02","g12",
                                                            "theta")
  return(print(results,digits=3))
}
##############################################################################

##############################################################################
# Cópula Clayton
#------------------------------
dDSB_CLAY <- function(y1,y2,x,b01,b11,b02,b12,g01,g11,g02,g12,lambda){
  mu1 <- exp(b01+b11*x)/(1+exp(b01+b11*x))
  sigma1 <- exp(g01+g11*x)

  mu2 <- exp(b02+b12*x)/(1+exp(b02+b12*x))
  sigma2 <- exp(g02+g12*x)

  # Densidade da variável y1
  dsim1 <- dSIMPLEX(y1,mu1,sigma1)

  # Densidade da variável y2
  dsim2 <- dSIMPLEX(y2,mu2,sigma2)

  # Cópula
  u <- pSIMPLEX(y1,mu1,sigma1)
  v <- pSIMPLEX(y2,mu2,sigma2)
  cop_CLAY <- (1+lambda)*u^(-1-lambda)*v^(-1-lambda)*(u^(-lambda)+
                                  v^(-lambda)-1)^(-2*lambda-1/lambda)
  dsim_f <- dsim1*dsim2*cop_CLAY
  return(dsim_f)
}
#------------------------------
lvero_CLAY <- function(theta,y1,y2,x){
  b01 <- theta[1]
  b11 <- theta[2]
  b02 <- theta[3]
  b12 <- theta[4]
  g01 <- theta[5]
  g11 <- theta[6]
  g02 <- theta[7]
```

```
  g12 <- theta[8]
  lambda <- theta[9]
  func <- dDSB_CLAY(y1,y2,x,b01,b11,b02,b12,g01,g11,g02,g12,lambda)
  lfunc <- -sum(log(func))
  #print(lfunc)
  return(lfunc)
}
#------------------------------
mle.DBS_CLAY <- function(y1,y2,x){
  op <- optim(par=chute,lvero_CLAY,method="L-BFGS-B",y1=y1,y2=y2,x=x,
           hessian=T,control=list(maxit=200),lower=c(rep(-Inf,8),0),
                                              upper=c(rep(Inf,8),Inf))

  fisher.information <- solve(op$hessian)
  standard.deviance <- sqrt(diag(fisher.information))
  t.simplex <- op$par/standard.deviance
  pvalue <- round(2*(1-pt(abs(t.simplex), length(x)-length(op$par))),3)
  results <- cbind(op$par,standard.deviance,t.simplex,pvalue)
  colnames(results) <- c("parâmetros", "desvio-padrão", "estatística t",
                                                          "p-valor")
  rownames(results) <- c("b01","b11","b02","b12","g01","g11","g02","g12",
                                                            "theta")
  return(print(results, digits = 3))
}
##############################################################################

##############################################################################
# Cópula Frank
#------------------------------
dDSB_FRANK <- function(y1,y2,x,b01,b11,b02,b12,g01,g11,g02,g12,lambda){
  mu1 <- exp(b01+b11*x)/(1+exp(b01+b11*x))
  sigma1 <- exp(g01+g11*x)

  mu2 <- exp(b02+b12*x)/(1+exp(b02+b12*x))
  sigma2 <- exp(g02+g12*x)

  # Densidade da variável y1
  dsim1 <- dSIMPLEX(y1, mu1, sigma1)

  # Densidade da variável y2
  dsim2 <- dSIMPLEX(y2, mu2, sigma2)

  # Cópula
  u <- pSIMPLEX(y1, mu1, sigma1)
  v <- pSIMPLEX(y2, mu2, sigma2)
  a <- lambda*exp(lambda*(1+u+v))*(-1+exp(lambda))
  b <- (exp(lambda)-exp(lambda*(1+u))-exp(lambda*(1+v))+exp(lambda*(u+v)))^2
  cop_FRANK <- a / b
  dsim_f <- dsim1*dsim2*cop_FRANK
  return(dsim_f)
}
#------------------------------
lvero_FRANK <- function(theta,y1,y2,x){
```

```
  b01 <- theta[1]
  b11 <- theta[2]
  b02 <- theta[3]
  b12 <- theta[4]
  g01 <- theta[5]
  g11 <- theta[6]
  g02 <- theta[7]
  g12 <- theta[8]
  lambda <- theta[9]
  func <- dDSB_FRANK(y1,y2,x,b01,b11,b02,b12,g01,g11,g02,g12,lambda)
  lfunc <- -sum(log(func))
  #print(lfunc)
  return(lfunc)
}
#------------------------------
mle.DBS_FRANK<- function(y1,y2,x){
  op <- optim(par=chute,lvero_FRANK,method="L-BFGS-B",y1=y1,y2=y2,x=x,
  hessian=T,control=list(maxit=200),lower=c(rep(-Inf,8),-Inf),
    upper=c(rep(Inf,8),Inf))

  fisher.information <- solve(op$hessian)
  standard.deviance <- sqrt(diag(fisher.information))
  t.simplex <- op$par/standard.deviance
  pvalue <- round(2*(1-pt(abs(t.simplex),length(x)-length(op$par))),3)
  results <- cbind(op$par,standard.deviance,t.simplex,pvalue)
  colnames(results) <- c("parâmetros", "desvio-padrão", "estatística t",
                                                          "p-valor")
  rownames(results) <- c("b01","b11","b02","b12","g01","g11","g02","g12",
                                                            "theta")
  return(print(results, digits = 3))
}
##############################################################################
```

### Resíduos Quantílico

```
##############################################################################
# Cópula FGM
#------------------------------
r.DBS_FGM <- function(z1,z2){
  beta01 <- theta[1]
  beta02 <- theta[2]
  gama01 <- theta[3]
  gama02 <- theta[4]
  beta11 <- theta[5]
  beta12 <- theta[6]
  gama11 <- theta[7]
  gama12 <- theta[8]
  lambda <- theta[9]

  eta1 <- beta01+beta02*x
  mu1 <- exp(eta1)/(1+exp(eta1))

  rho1 <- gama01+gama02*x
```

```
  sigma1 <- exp(rho1)

  eta2 <- beta11+beta12*x
  mu2 <- exp(eta2)/(1+exp(eta2))

  rho2 <- gama11+gama12*x
  sigma2 <- exp(rho2)

  f1 <- dSIMPLEX(z1,mu1,sigma1)
  f2 <- dSIMPLEX(z2,mu2,sigma2)

  F1 <- pSIMPLEX(z1,mu1,sigma1)
  F2 <- pSIMPLEX(z2,mu2,sigma2)

  r <- f1*f2*(1+lambda*(1-2*F1)*(1-2*F2))

  return(r)
}
#-----------------------------
cIntegral_FGM <- function(y1,y2){
  aux1 <- integral2(r.DBS_FGM,0,y1,0,y2)
  return(aux1$Q)
}
#-----------------------------
aux <- numeric()
n <- length(y1)
x.tem <- variável explicativa (x)
i <- 1
while(i <= n){
  theta <- theta
  yy1 <- y1[j]
  yy2 <- y2[j]
  x <- x.tem[j]
  ff <- cIntegral_FGM(y1=yy1,y2=yy2)
  fff <- qnorm(ff)
  aux[j] <- fff
  x <- x.tem
  i <- i + 1
}
#-----------------------------
plot(aux,main="",xlab = "Índices de observações",
                                ylab="Residuos quantílico",cex.lab=1.5)
abline(-2,0,lty=2);abline(0,0,lty=2);abline(2,0,lty=2);cutI=-2;cutS=2
qqnorm(aux1,main="",xlab="Percentil N(0,1)",ylab="Resíduo quantílico")
##########################################################################

##########################################################################
# Resíduo Cópula Clayton
#-----------------------------
r.DBS_CLAY <- function(z1,z2){
  beta01 <- theta[1]
  beta02 <- theta[2]
  gama01 <- theta[3]
```

```
  gama02 <- theta[4]
  beta11 <- theta[5]
  beta12 <- theta[6]
  gama11 <- theta[7]
  gama12 <- theta[8]
  lambda <- theta[9]

  eta1 <- beta01+beta02*x
  mu1 <- exp(eta1)/(1+exp(eta1))

  rho1 <- gama01+gama02*x
  sigma1 <- exp(rho1)

  eta2 <- beta11+beta12*x
  mu2 <- exp(eta2)/(1+exp(eta2))

  rho2 <- gama11+gama12*x
  sigma2 <- exp(rho2)

  f1 <- dSIMPLEX(z1,mu1,sigma1)
  f2 <- dSIMPLEX(z2,mu2,sigma2)

  F1 <- pSIMPLEX(z1,mu1,sigma1)
  F2 <- pSIMPLEX(z2,mu2,sigma2)

  cop_CLAY <- (1+lambda)*F1^(-1-lambda)*F2^(-1-lambda)*(F1^(-lambda)+
                                  F2^(-lambda)-1)^(-2*lambda-1/lambda)
  r <- f1*f2*cop_CLAY

  return(r)
}
#-----------------------------
cIntegral_CLAY <- function(y1,y2){
  aux1 <- integral2(r.DBS_CLAY,0,y1,0,y2)
  return(aux1$Q)
}
#-----------------------------
aux <- numeric()
n <- length(y1)
x.tem <- variável explicativa (x)
i <- 1
while(i <= n){
  theta <- theta
  yy1 <- y1[j]
  yy2 <- y2[j]
  x <- x.tem[j]
  ff <- cIntegral_CLAY(y1=yy1,y2=yy2)
  fff <- qnorm(ff)
  aux[j] <- fff
  x <- x.tem
  i <- i + 1
}
```

```
#-----------------------------
plot(aux,main="",xlab = "Índices de observações",
                                ylab="Residuos quantílico",cex.lab=1.5)
abline(-2,0,lty=2);abline(0,0,lty=2);abline(2,0,lty=2);cutI=-2;cutS=2
qqnorm(aux1,main="",xlab="Percentil N(0,1)",ylab="Resíduo quantílico")
##############################################################################

##############################################################################
# Resíduo Cópula Frank
#-----------------------------
r.DBS_FRANK <- function(z1,z2){
  beta01 <- theta[1]
  beta02 <- theta[2]
  gama01 <- theta[3]
  gama02 <- theta[4]
  beta11 <- theta[5]
  beta12 <- theta[6]
  gama11 <- theta[7]
  gama12 <- theta[8]
  lambda <- theta[9]

  eta1 <- beta01+beta02*x
  mu1 <- exp(eta1)/(1+exp(eta1))

  rho1 <- gama01+gama02*x
  sigma1 <- exp(rho1)

  eta2 <- beta11+beta12*x
  mu2 <- exp(eta2)/(1+exp(eta2))

  rho2 <- gama11+gama12*x
  sigma2 <- exp(rho2)

  f1 <- dSIMPLEX(z1,mu1,sigma1)
  f2 <- dSIMPLEX(z2,mu2,sigma2)

  F1 <- pSIMPLEX(z1,mu1,sigma1)
  F2 <- pSIMPLEX(z2,mu2,sigma2)

  a <- lambda*exp(lambda*(1+F1+F2))*(-1+exp(lambda))
  b <- (exp(lambda)-exp(lambda*(1+F1))-exp(lambda*(1+F2))
                                  +exp(lambda*(F1+F2)))^2
  cop_FRANK <- a / b

  r <- f1*f2*cop_FRANK)

  return(r)
}
#-----------------------------
cIntegral_FRANK <- function(y1,y2){
  aux1 <- integral2(r.DBS_FRANK,0,y1,0,y2)
  return(aux1$Q)
}
```

```
#-----------------------------
aux <- numeric()
n <- length(y1)
x.tem <- variável explicativa (x)
i <- 1
while(i <= n){
  theta <- theta
  yy1 <- y1[j]
  yy2 <- y2[j]
  x <- x.tem[j]
  ff <- cIntegral_FRANK(y1=yy1,y2=yy2)
  fff <- qnorm(ff)
  aux[j] <- fff
  x <- x.tem
  i <- i + 1
}
#-----------------------------
plot(aux,main="",xlab = "Índices de observações",
                                ylab="Residuos quantílico",cex.lab=1.5)
abline(-2,0,lty=2);abline(0,0,lty=2);abline(2,0,lty=2);cutI=-2;cutS=2
qqnorm(aux1,main="",xlab="Percentil N(0,1)",ylab="Resíduo quantílico")
##########################################################################
```

### Distância de Cook-generalizada e Afastamento da verossimilhança

```
##########################################################################
# Cópula FGM
#-----------------------------
dDSB_FGM <- function(y1,y2,x,b01,b11,b02,b12,g01,g11,g02,g12,lambda){
  mu1 <- exp(b01+b11*x)/(1+exp(b01+b11*x))
  sigma1 <- exp(g01+g11*x)

  mu2 <- exp(b02+b12*x)/(1+exp(b02+b12*x))
  sigma2 <- exp(g02+g12*x)

  # Densidade da variável y1
  dsim1 <- dSIMPLEX(y1,mu1,sigma1)

  # Densidade da variável y2
  dsim2 <- dSIMPLEX(y2,mu2,sigma2)

  # Cópula
  u <- pSIMPLEX(y1,mu1,sigma1)
  v <- pSIMPLEX(y2,mu2,sigma2)
  cop_fgm <- 1+(lambda*(1-2*u)*(1-2*v))
  dsim_f <- dsim1*dsim2*cop_fgm
  return(dsim_f)
}
#-----------------------------
lvero_FGM <- function(theta,y1,y2,x){
  b01 <- theta[1]
  b11 <- theta[2]
  b02 <- theta[3]
```

```
  b12 <- theta[4]
  g01 <- theta[5]
  g11 <- theta[6]
  g02 <- theta[7]
  g12 <- theta[8]
  lambda <- theta[9]
  func <- dDSB_FGM(y1,y2,x,b01,b11,b02,b12,g01,g11,g02,g12,lambda)
  lfunc <- -sum(log(func))
  #print(lfunc)
  return(lfunc)
}
#------------------------------
op <- optim(par=chute,lvero_FGM,method="L-BFGS-B",y1=y1,y2=y2,x=x,
        hessian=T,control=list(maxit=200),lower=c(rep(-Inf,8),-1),
                                            upper=c(rep(Inf,8),1))

d_cook <- c()
afast_vs <- c()
i <- 1
while (i<=204) {
  y1 <- variável dependente 1 [-i]
  y2 <- variável dependente 2 [-i]
  x <- (t_razdep/100)[-i]

op2 <- optim(par=chute,logLikFun_sfgm2,method="L-BFGS-B",y1=y1,
                      y2=y2,x=x,hessian=T,control=list(maxit=200),
                    lower=c(rep(-Inf,8),-1),upper=c(rep(Inf,8),1))

  fshs1 <- -solve(op$hessian)
  mts1 <- as.vector(op2$par) - as.vector(op$par)

  afast_vs[i] <- (2*(op$par)-(op2$par))
  d_cook[i] <- t(mts1)  %*% fshs1  %*% (mts1)

  i <- i + 1
  print(i)
}
plot(abs(d_cook), ylab="Distância de Cook generalizada",
                              xlab="índice das observações",cex.lab=1.5)
identify(abs(d_cook))

plot(abs(afast_vs), ylab="Afastamento da verossimilhança",
                              xlab="índice das observações",cex.lab=1.5)
identify(abs(afast_vs))
##############################################################################
```

### Gráfico de envelope simulado

```
##############################################################################
#-----------------------------------------------------------------------------
# Cópula FGM
#------------------------------
# library(simplexreg)
```

```
R = 100
n = 204
theta.fixo <-c(b01="",b11="",b02="",b12="",g01="",g11="",g02="",g12="",
                                                                lambda="")
y1 <- numeric(length(n))
y2 <- numeric(length(n))
x <- variável explicativa (x)
#mats <- matrix(NA, nrow = R, ncol = 9) #
mats_res_FGM <- matrix(NA, nrow = n, ncol = R) #
j = 1
while (j <= R){
  u1 <- runif(n)
  v <- runif(n)
  A <- (theta.fixo[9]*(2*u1-1)-1)
  B <- (1-(theta.fixo[9]*(2*u1-1)))^2+(4*v*theta.fixo[9]*(2*u1-1))
  u2 <- (2*v/(sqrt(B)-A))

  mu1 <- exp(theta.fixo[1]+theta.fixo[2]*x)/
                                  (1+exp(theta.fixo[1]+theta.fixo[2]*x))
  mu2 <- exp(theta.fixo[3]+theta.fixo[4]*x)/
                                  (1+exp(theta.fixo[3]+theta.fixo[4]*x))

  sigma1 <- exp(theta.fixo[5]+theta.fixo[6]*x)
  sigma2 <- exp(theta.fixo[7]+theta.fixo[8]*x)

  y1 <- qSIMPLEX(u1, mu1, sigma1)
  y2 <- qSIMPLEX(u2, mu2, sigma2)


  chute <- theta.fixo

  op_s <- try(optim(par=chute,logLikFun_sfgm2, method = "L-BFGS-B",
  hessian=T,y1=y1,y2=y2,x=x,lower=c(rep(-Inf,8),-1),
  upper=c(rep(Inf,8),1),control=list(maxit=300)),silent=T)

  theta_ <- c(op_s$par[1],op_s$par[2],op_s$par[3],op_s$par[4],op_s$par[5],
                          op_s$par[6],op_s$par[7],op_s$par[8],op_s$par[9])

  if(!class(op_s)=="try-error"){
    resi <- numeric()
    n <- length(y1)
    x.tem <- x
    for(i in 1:n){
      theta <- theta_
      yy1 <- y1[i]
      yy2 <- y2[i]
      x <- x.tem[i]
      ff <- try(cIntegral_fgm(y1=yy1,y2=yy2), silent = T)
      if(!class(ff)=="try-error" && ff !="NaN"){
        resi[i] <- qnorm(ff)
      }else{
        resi[i] <- rq.FGM[i]
```

```
        }
      x <- x.tem
    }
    mats_res_FGM[,j] <- resi
    j <- j + 1
    print(j)
  }
}
#---------------------------
res <- cbind(mats_res_FGM,rq.FGM)
res1 <- apply(res,2,sort)
rq_min1 <- apply(res1,1,min)
rq_mean1 <- apply(res1,1,mean)
rq_max1 <- apply(res1,1,max)

a <- qqnorm(rq_min1,axes=F,xlab="",ylab="",type="l",main="")
b <- qqnorm(rq_mean1,axes=F,xlab="",ylab="",type="l",main="")
c <- qqnorm(rq_max1,axes=F,xlab="",ylab="",type="l",main="")

qqnorm(rq.FGM,main="",xlab="Percentil N(0,1)",ylab="Resíduo quantílico")
points(a$x,a$y,type="l",col=4)
points(b$x,b$y,type="l",col=2)
points(c$x,c$y,type="l",col=4)
#-----------------------------------------------------------------------------
###############################################################################
#-----------------------------------------------------------------------------
# Cópula Clayton
#------------------------------
R = 100
n = 204
theta.fixo <-c(b01="",b11="",b02="",b12="",g01="",g11="",g02="",g12="",
                                                              lambda="")
y1 <- numeric(length(n))
y2 <- numeric(length(n))
x <- variável explicativa (x)
mats_res_Clay <- matrix(NA, nrow = n, ncol = R) #
j = 1
while (j <= R){
  u1 <- runif(n)
  v <- runif(n)
  u2 <- ((1 - (u1^-theta.fixo[9])) +
    (v*(u1^(1+theta.fixo[9])))^-(theta.fixo[9]/
                                 (1+theta.fixo[9])))^(-1/theta.fixo[9])

  mu1 <- exp(theta.fixo[1]+theta.fixo[2]*x)/
                                 (1+exp(theta.fixo[1]+theta.fixo[2]*x))
  mu2 <- exp(theta.fixo[3]+theta.fixo[4]*x)/
                                 (1+exp(theta.fixo[3]+theta.fixo[4]*x))

  sigma1 <- exp(theta.fixo[5]+theta.fixo[6]*x)
  sigma2 <- exp(theta.fixo[7]+theta.fixo[8]*x)
```

```
  y1 <- qSIMPLEX(u1, mu1, sigma1)
  y2 <- qSIMPLEX(u2, mu2, sigma2)

  chute <- theta.fixo

  op_s <- try(optim(par=chute,logLikFun_sclay2, method = "L-BFGS-B",
          hessian=T,y1=y1,y2=y2,x=x,lower=c(rep(-Inf,8),0),
          upper=c(rep(Inf,8),Inf),control=list(maxit=300)),silent=T)

  theta_ <- c(op_s$par[1],op_s$par[2],op_s$par[3],op_s$par[4],op_s$par[5],
                          op_s$par[6],op_s$par[7],op_s$par[8],op_s$par[9])

  if(!class(op_s)=="try-error"){
    resi <- numeric()
    n <- length(y1)
    x.tem <- x
    for(i in 1:n){
      theta <- theta_
      yy1 <- y1[i]
      yy2 <- y2[i]
      x <- x.tem[i]
      ff <- try(cIntegral_clay(y1=yy1,y2=yy2), silent = T)
      if(!class(ff)=="try-error" && ff !="NaN"){
        resi[i] <- qnorm(ff)
      }else{
        resi[i] <- aux2[i]
      }
      x <- x.tem
    }
    mats_res_Clay[,j] <- resi
    j <- j + 1
    print(j)
  }
}
#-------------------------------------------------------------------------
res <- cbind(mats_res_Clay[,c(1:5,7:25)],aux2)
res1 <- apply(res,2,sort)
rq_min1 <- apply(res1,1,min)
rq_mean1 <- apply(res1,1,mean)
rq_max1 <- apply(res1,1,max)

a <- qqnorm(rq_min1,axes=F,xlab="",ylab="",type="l",main="")
b <- qqnorm(rq_mean1,axes=F,xlab="",ylab="",type="l",main="")
c <- qqnorm(rq_max1,axes=F,xlab="",ylab="",type="l",main="")

qqnorm(aux2,main="",xlab="Percentil N(0,1)",ylab="Resíduo quantílico")
points(a$x,a$y,type="l",col=4)

points(b$x,b$y,type="l",col=2)
points(c$x,c$y,type="l",col=4)
#-------------------------------------------------------------------------
#########################################################################
```

```
#-------------------------------------------------------------------------
# Cópula Frank
#-----------------------------
R = 100
n = 204
theta.fixo <-c(b01="",b11="",b02="",b12="",g01="",g11="",g02="",g12="",
                                                              lambda="")
y1 <- numeric(length(n))
y2 <- numeric(length(n))
x <- variável explicativa (x)
mats_res_Frank <- matrix(NA, nrow = n, ncol = R) #
j = 1
while (j <= R){
  u1 <- runif(n)
  v <- runif(n)
  u2 <- -(1/theta.fixo[9])*log(1+(v*(exp(1)^-theta.fixo[9]-1)/
                                  (v+(1-v)*exp(1)^(-theta.fixo[9]*u1))))

  mu1 <- exp(theta.fixo[1]+theta.fixo[2]*x)/
                                  (1+exp(theta.fixo[1]+theta.fixo[2]*x))
  mu2 <- exp(theta.fixo[3]+theta.fixo[4]*x)/
                                  (1+exp(theta.fixo[3]+theta.fixo[4]*x))

  sigma1 <- exp(theta.fixo[5]+theta.fixo[6]*x)
  sigma2 <- exp(theta.fixo[7]+theta.fixo[8]*x)

  y1 <- qSIMPLEX(u1, mu1, sigma1)
  y2 <- qSIMPLEX(u2, mu2, sigma2)

  chute <- theta.fixo

  op_s <- try(optim(par=chute,logLikFun_sfrank2, method = "L-BFGS-B",
                hessian=T,y1=y1,y2=y2,x=x,lower=c(rep(-Inf,8),-Inf),
          upper=c(rep(Inf,8),Inf),control=list(maxit=300)),silent=T)

  theta_ <- c(op_s$par[1],op_s$par[2],op_s$par[3],op_s$par[4],op_s$par[5],
                          op_s$par[6],op_s$par[7],op_s$par[8],op_s$par[9])

  if(!class(op_s)=="try-error"){
    resi <- numeric()
    n <- length(y1)
    x.tem <- x
    for(i in 1:n){
      theta <- theta_
      yy1 <- y1[i]
      yy2 <- y2[i]
      x <- x.tem[i]
      ff <- try(cIntegral(y1=yy1,y2=yy2), silent = T)
      if(!class(ff)=="try-error" && ff !="NaN"){
        resi[i] <- qnorm(ff)
      }else{
        resi[i] <- aux3[i]
```

```
      }
      x <- x.tem
    }
    mats_res_Frank[,j] <- resi
    j <- j + 1
    print(j)
  }
}
#-------------------------------------------------------------------------
res <- mats_res_Frank
res1 <- apply(res,2,sort)
rq_min1 <- apply(res1,1,min)
rq_mean1 <- apply(res1,1,mean)
rq_max1 <- apply(res1,1,max)

a <- qqnorm(rq_min1,axes=F,xlab="",ylab="",type="l",main="")
b <- qqnorm(rq_mean1,axes=F,xlab="",ylab="",type="l",main="")
c <- qqnorm(rq_max1,axes=F,xlab="",ylab="",type="l",main="")

qqnorm(aux3,main="",xlab="Percentil N(0,1)",ylab="Resíduo quantílico")
points(a$x,a$y,type="l",col=4)
points(b$x,b$y,type="l",col=2)
points(c$x,c$y,type="l",col=4)
#-------------------------------------------------------------------------
```