Análise de dados correlacionados *unit*-Lindley baseada em equações de estimação

Danilo Vieira Silva

Dissertação apresentada ao
Instituto de Matemática e Estatística
da Universidade de São Paulo
para obtenção do título de
Mestre em Ciências

Programa: Estatística
Orientador: Prof. Dr. Gilberto Alvarenga Paula

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES/CNPq

São Paulo
Agosto de 2023

Análise de dados correlacionados *unit*-Lindley baseada em equações de estimação

Danilo Vieira Silva

Esta versão da dissertação contém
as correções e alterações sugeridas
pela Comissão Julgadora durante a
defesa da versão original do trabalho,
realizada em 17 de Agosto de 2023.

Uma cópia da versão original está
disponível no Instituto de Matemática e
Estatística da Universidade de São Paulo.

Comissão julgadora:

Prof. Dr. Gilberto Alvarenga Paula (orientador) – IME-USP
Profª. Drª. Cibele Maria Russo Novelli – ICMC-USP
Profª. Drª. Michelli Karinne Barros da Silva – UFCG

Agradecimentos

"Não vou smokar a Banana se não falar para smokar a Banana, vou guardar minha smoke."
— FalleN

Ao Prof. Gilberto pela orientação e confiança. Aos amigos e professores pela inspiração e discussões. À Profª. Hatice Akdur da Universidade de Gazi, Turquia, pela colaboração. À CCP-MAE do IME-USP pelo apoio na participação de eventos. Às Profª. Cibele e Profª. Michelli pelas correções e sugestões. Ao CNPq e CAPES pelo apoio financeiro.

Resumo

Danilo Vieira Silva. **Análise de dados correlacionados *unit*-Lindley baseada em equações de estimação**. Dissertação (Mestrado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2023.

Neste texto derivamos equações de estimação para modelar conjuntos de dados correlacionados em que as distribuições marginais seguem as distribuições uniparamétricas *unit*-Lindley com domínio no intervalo $(0, 1)$. Uma classe de modelos de regressão é proposta para modelar o parâmetro de posição e um processo iterativo reponderado é desenvolvido para a estimação dos coeficientes da regressão e da estrutura de correlação. Estudos de simulação são realizados para verificar as propriedades empíricas dos estimadores derivados e métodos de diagnóstico, como análise de resíduos e estudos de sensibilidade baseados na curvatura conformal local são apresentados. Como aplicação, analisamos a proporção de pessoas em domicílios com abastecimento inadequado de água e esgoto nas unidades federativas do Brasil com os procedimentos desenvolvidos no texto. Finalmente, estendemos as equações de estimação para modelar conjuntamente o parâmetro de posição e a estrutura de correlação. Um processo iterativo simultâneo é derivado e algumas propriedades assintóticas dos estimadores são apresentadas.

Palavras-chave: Distribuição *unit*-Lindley. Dados correlacionados. Métodos de diagnóstico. Equações de estimação.

Abstract

Danilo Vieira Silva. **Analysis of correlated unit-Lindley data based on estimating equations**. Thesis (Master's). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2023.

In this text we derive estimating equations for modeling correlated data in which the marginal distributions follow the one parameter unit-Lindley distributions with domain on the interval $(0, 1)$. A class of regressions models is proposed for modeling the location parameter and a reweighted iterative process is developed for the estimation of the regression coefficients and the correlation structure. Simulation studies are performed to assess the empirical properties of the derived estimators and diagnostic procedures, such as residual analysis and sensitivity studies based on conformal local influence are given. As illustration, we analyze the proportion of people in households with inadequate water supply and sewage within federation units of Brazil by the procedures developed in the text. Finally, we extend the estimating equations for modeling jointly the position parameter and the correlation structure. A simultaneous iterative process is derived and some asymptotic properties of the estimators are presented.

Keywords: Unit-Lindley distribution. Correlated data. Diagnostic procedures. Estimating equations.

Sumário

Introdução

Distribuições no intervalo unitário como as distribuições beta, simplex e Kumaraswamy são amplamente conhecidas na literatura Estatística (veja, por exemplo, Ferrari e Cribari-Neto, 2004; Barndorff-Nielsen e Jørgensen, 1991; Kumaraswamy, 1980) e a recente literatura inclui novas distribuições para modelar respostas com tal característica. Por exemplo, Mazuchelli et al. (2019) introduziram a distribuição *unit*-Lindley junto ao seu respectivo modelo de regressão como uma alternativa à distribuição beta usando uma transformação da variável de distribuição Lindley (Ghitany et al., 2008) e mais recentemente Altun et al. (2021) apresentaram a distribuição log-Bilal usando uma transformação da variável de distribuição Bilal (Abd-Elrahman, 2013), enquanto Queiroz e Ferrari (2023) propuseram a classe *power logit* com parâmetros de posição, dispersão e assimetria. Embora Grassia (1977) apresente a distribuição *unit*-gamma, a mesma não foi utilizada como uma distribuição para respostas no intervalo unitário contínuo até Mousa et al. (2016) desenvolverem o modelo de regressão. Para respostas do tipo proporções distribuídas de forma independente, modelos de regressão beta, simplex, *unit*-gamma, *unit*-Lindley, log-Bilal e *power logit* podem ser preferidos dependendo da adequabilidade ao conjunto de dados. Entretanto, planejamentos de medidas repetidas, ensaios clínicos longitudinais e planejamentos de amostras com *clusters* induzem a estruturas de dados multiníveis que não são apropriadas de serem analizados sob modelos de regressão com respostas independentes.

Um exemplo para respostas do tipo proporções longitudinais trata-se da porcentagem de gás restante no olho oriundas de um estudo de oftalmologia, em que foi analizado usando modelos simplex com efeitos aleatórios (Qiu et al., 2008). Respostas multiníveis do tipo proporções, como a proporção de locais dentais doentes de um estudo odontológico e o índice de qualidade da água foram analizados, respectivamente, usando modelos beta com efeito aleatório e modelos quase-beta longitudinais por Galvis et al. (2014) e Petterle et al. (2019). Recentemente, Akdur (2021) desenvolveu modelos *unit*-Lindley com efeito aleatório.

O foco deste texto é propor uma abordagem alternativa para modelar dados *unit*-Lindley desbalanceados correlacionados (unidades experimentais não têm o mesmo número de observações) baseada em equações de estimação similarmente aos trabalhos desenvolvidos por Artes e Jørgensen (2000), Venezuela et al. (2011) e Tsuyuguchi et al. (2020), contudo com contribuições à análise de resíduos. Assim, partindo da classe proposta por Crowder (1987), derivamos uma classe ótima de equações de estimação para modelar dados correlacionados em que as distribuições marginais são assumidas *unit*-Lindley no intervalo $(0, 1)$. As equações de estimação e as propriedades assintóticas dos estimadores obtidos são baseados na teoria desenvolvida por Godambe (1997), com a suposição de que as

correlações entre as unidades experimentais seguem a mesma estrutura das equações de estimação generalizadas (GEE) propostas por Liang e Zeger (1986). Um processo iterativo reponderado é desenvolvido para a estimação dos coeficientes da regressão e as propriedades assintóticas e empíricas dos estimadores derivados são discutidas. Métodos de diagnóstico são propostos assim como é realizada uma aplicação a um conjunto de dados real para ilustração e uma possível extensão para estruturas de correlação gerais em que as correlações são modeladas.

O texto está organizado como segue. No Capítulo 1 uma breve revisão sobre a distribuição *unit*-Lindley no intervalo $(0, 1)$ é realizada. Uma classe de modelos de regressão é proposta no Capítulo 2 para modelar taxas e proporções correlacionadas e uma classe ótima de funções de estimação é derivada, bem como um processo iterativo para a estimação dos coeficientes da regressão com algumas discussões das suas propriedades assintóticas. Também são propostos procedimentos de diagnóstico, tais como análise de resíduos baseada no resíduo quantílico marginal e estudos de sensibilidade baseados em influência local conformal. Estudos de simulação para avaliar as propriedades empíricas dos estimadores são realizados no Capítulo 3 e também é apresentada uma aplicação com um conjunto de dados reais para ilustrar a metodologia desenvolvida no texto. A classe de modelos de regressão proposta é estendida para estruturas de correlação gerais em que as correlações são modeladas, uma classe de funções de estimação conjunta é derivada e um processo iterativo simultâneo para estimação dos coeficientes da regressão dos dois componentes é desenvolvido no Capítulo 4, como também apresentadas propriedades assintóticas dos estimadores propostos. O Capítulo 5 trata das considerações finais, e alguns resultados técnicos e códigos em R são apresentados nos Apêndices A e B para melhor organização do texto.

Capítulo 1

Distribuição *unit*-Lindley

A distribuição uniparamétrica Lindley foi proposta por Lindley (1958) e é amplamente usada em análise de sobrevivência na modelagem do tempo em estudos de mortalidade. Denotando por z a variável aleatória com tal distribuição parametrizada por θ, cuja função densidade acumulada pode ser escrita como

$$F(z;\theta) = 1 - \left(1 + \frac{z\theta}{1+\theta}\right)\exp\left(-z\theta\right),$$

com $z > 0$ e $\theta > 0$. Usando a transformação $y = z/(1+z)$, Mazucheli et al. (2019) propõem a distribuição *unit*-Lindley com suporte no intervalo unitário.

Denotando por y a variável aleatória com distribuição *unit*-Lindley parametrizada por μ para ajustar taxas e proporções, cujas funções densidade de probabilidade e acumulada podem ser escritas, respectivamente, como

$$f(y;\mu) = \frac{(1-\mu)^2}{\mu(1-y)^3}\exp\left\{-\frac{y(1-\mu)}{\mu(1-y)}\right\},$$

e

$$F(y;\mu) = 1 - \left(\frac{1-y\mu}{1-y}\right)\exp\left\{-\frac{(1-\mu)y}{(1-y)\mu}\right\},$$

com $0 < y < 1$, $0 < \mu < 1$ e $\mathrm{E}(y) = \mu$. As funções densidade acima podem ser reparametrizadas com $\theta = (1-\mu)/\mu$, $\theta > 0$. A função de variância pode ser expressa como

$$\mathrm{Var}(y) = \frac{(1-\mu)^2}{\mu}\left[\mathrm{E}_1\left(\frac{1-\mu}{\mu}\right)\exp\left(\frac{1-\mu}{\mu}\right) - \mu\right],$$

em que $\mathrm{E}_n(x) = \int_1^\infty t^{-n}\exp(-xt)dt$ denota a função exponencial integral, para $n \in \{0, 1, \ldots\}$ e $x \in \mathbb{R}$, com $\mathrm{E}_1(x) = \int_1^\infty t^{-1}\exp(-xt)dt$. Também, o p-ésimo quantil assume a forma

$$F^{-1}(p;\mu) = \frac{\frac{1}{\mu} + \mathrm{W}_{-1}\left\{\frac{p-1}{\mu\exp(\mu^{-1})}\right\}}{1 + \mathrm{W}_{-1}\left\{\frac{p-1}{\mu\exp(\mu^{-1})}\right\}},$$

com $0 < p < 1$ e $W_{-1}(a)$ denota a ramo negativo da função Lambert-W, para $a \in [-e^{-1}, 0)$. Propriedades adicionais da distribuição *unit*-Lindley podem ser encontradas em Mazucheli et al. (2019). Vamos denotar ao longo do texto $y \sim \text{UL}(\mu)$ para a variável aleatória *unit*-Lindley. A Figura 1.1 apresenta para y as formas das funções densidade de probabilidade e variância, podemos notar que a densidade decai mais rapidamente à direita do que à esquerda e observar a assimetria da função de variância que tem máximo de aproximadamente 0.048 quando $\mu = 0.53$.

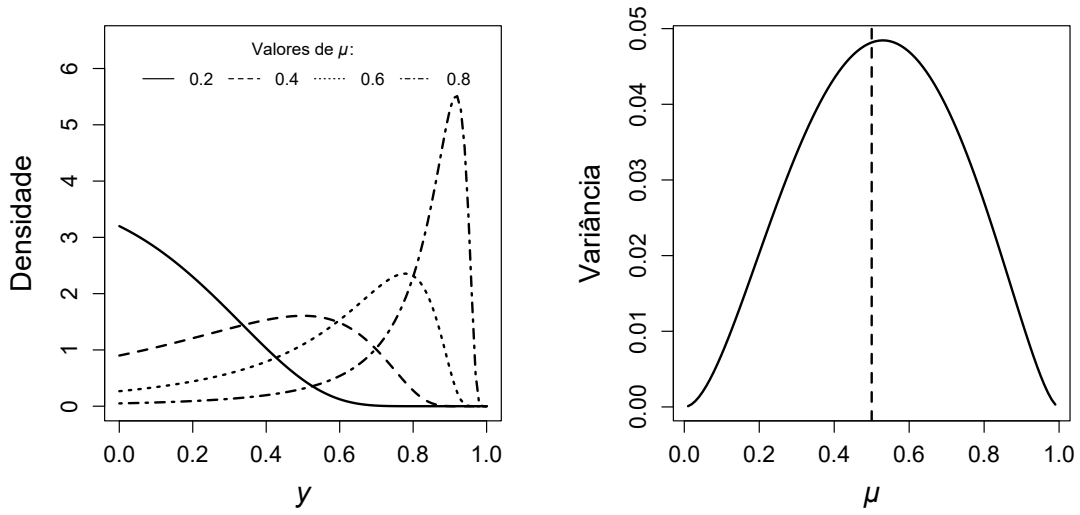

Figura 1.1: *Formas da distribuição UL. O painel esquerdo mostra a função densidade de probabilidade para alguns valores de μ. O painel direito mostra o comportamento da função de variância.*

Na sequência vamos apresentar alguns resultados importantes relacionados com a função escore do parâmetro μ, necessários para escrever as equações de estimação para modelar dados correlacionados com distribuição marginal UL.

O logaritmo da função densidade de probabilidade tem a forma

$$\text{L}(\mu) = 2\log\left(\frac{1-\mu}{\mu}\right) - \frac{y(1-\mu)}{\mu(1-y)} - 3\log(1-y),$$

e consequentemente a função escore de μ é dada por

$$u = \frac{d\text{L}(\mu)}{d\mu} = \frac{z}{\mu^2} - \frac{1+\mu}{\mu(1-\mu)},$$

em que $z = y/(1-y)$ denota a chance observada. Note que, para μ fixado, z é uma função monotônica de y. Pode ser mostrado que $z \sim \text{Lindley}(\mu^{-1} - 1)$ pertence à família exponencial uniparamétrica de distribuições $f(z;\theta) = \exp\{\theta z - b(\theta) + c(z)\}$ (veja Apêndice A.1). Depois de algumas manipulações algébricas (veja Apêndice A.2) obtemos

$$\text{E}(u) = 0 \quad \text{e} \quad \text{Var}(u) = \frac{2-(1-\mu)^2}{\mu^2(1-\mu)^2}.$$

Toda a teoria desenvolvida para modelos lineares generalizados (McCullagh and Nelder, 1989) pode ser aplicada para modelar $\mathrm{E}(z)$. Entretanto, nosso interesse neste texto é modelar o parâmetro μ com funções de ligação apropriadas e dados correlacionados.

1.1 Modelos UL

Seja $\mathbf{y} = (y_1, \ldots, y_n)^\top$ um vetor $n \times 1$ contendo as respostas (taxas ou proporções). Vamos assumir que $y_i \sim \mathrm{UL}(\mu_i)$ independentes com componente de regressão $g(\mu_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$, para $i = 1, \ldots, n$, em que $g(\cdot)$ denota a função de ligação com domínio no $(0, 1)$ e diferenciável, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ contém os valores de variáveis explicativas e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ o vetor com os coeficientes da regressão. A função log-verossimilhança fica dada por

$$\mathrm{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} 2 \log \left\{ \frac{1 - g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})}{g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})} \right\} - \frac{y_i\{1 - g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})\}}{g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})\{1 - y_i\}} - 3 \log(1 - y_i),$$

em que $g^{-1}(\cdot)$ denota a função de ligação inversa. Denotando $d\eta_i/d\mu_i = g'(\mu_i)$, segue que a função escore do modelo de regressão e a matriz de informação de Fisher ficam dadas, respectivamente, por

$$\mathrm{U}_\beta = \mathbf{X}^\top \mathbf{W} \mathbf{D}^{-1} \mathbf{u}, \quad \text{e} \quad \mathrm{K}_{\beta\beta} = \mathbf{X}^\top \mathbf{W} \mathbf{X},$$

com $\mathbf{X}$ sendo uma matriz $n \times p$ de linhas $x_i^\top$, $\mathbf{W} = \mathrm{diag}\{w_1, \ldots, w_n\}$ com $w_i = \mathrm{Var}(u_i)\{g'(\mu_i)\}^{-2}$, $\mathbf{D} = \mathrm{diag}\{d_1, \ldots, d_n\}$ com $d_i = \mathrm{Var}(u_i)\{g'(\mu_{ij})\}^{-1}$ e $\mathbf{u} = (u_1, \ldots, u_n)^\top$ com $u_i = d\mathrm{L}(\boldsymbol{\mu})/d\mu_i$ para $i = 1, \ldots, n$. Assim, o processo iterativo escore de Fisher para obter a estimativa de máxima verossimilhança $\hat{\boldsymbol{\beta}}$ assume a forma de mínimos quadrados reponderados

$$\boldsymbol{\beta}^{(m+1)} = \{\mathbf{X}^\top \mathbf{W}^{(m)} \mathbf{X}\}^{-1} \mathbf{X}^\top \mathbf{W}^{(m)} \mathbf{t}^{(m)}, \ m = 0, 1, 2 \ldots,$$

em que $\mathbf{t} = \mathbf{X}\boldsymbol{\beta} - \mathbf{D}^{-1}\mathbf{u}$ desempenha o papel de uma variável resposta modificada. Para n grande, sob condições gerais de regularidade, tem-se que $\hat{\boldsymbol{\beta}} \sim \mathrm{N}_p(\boldsymbol{\beta}, \mathrm{K}_{\beta\beta}^{-1})$. Algumas propriedades de estimadores para o caso sem covariáveis e análise de resíduos do modelo de regressão por meio do resíduo de Cox-Snell (Cox e Snell, 1968) podem ser encontradas em Mazucheli et al. (2019).

Para avaliar a adequabilidade do ajuste do modelo e detectar observações aberrantes pode-se também considerar o resíduo quantílico (Dunn e Smyth, 1996), amplamente utilizado na área de modelos lineares generalizados e definido como

$$r_{q_i} = \Phi^{-1}\{F(y_i; \hat{\mu}_i)\},$$

em que $F(y_i; \hat{\mu}_i)$ denota a função densidade acumulada de $y_i \sim \mathrm{UL}(\mu_i)$ avaliada em $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$ e $\Phi(\cdot)$ a função densidade acumulada da distribuição normal padrão, para $i = 1, \ldots, n$. Segue que r_{q_i} tem distribuição assintótica $\mathrm{N}(0, 1)$ caso o modelo postulado seja adequado. As medidas de sensibilidade são obtidas de forma similar às derivações apresentadas ao longo da Seção 2.4 de modo que a matriz $\mathbf{B}$ é aproximada por $\boldsymbol{\Delta}^\top \mathrm{K}_{\beta\beta}^{-1} \boldsymbol{\Delta}$, com $\boldsymbol{\Delta} = \partial \dot{\mathrm{L}}(\boldsymbol{\beta}|\boldsymbol{\omega})/\partial\boldsymbol{\beta}\partial\boldsymbol{\omega}^\top$ denotando a matriz associada ao esquema de perturbação adotado.

Capítulo 2

Modelos UL-GEE

Estendendo a Seção 1.1, seja s_i o conjunto de índices relativos aos instantes em que as respostas são observadas na i-ésima unidade experimental, com cardinalidade denotada por $n(s_i)$. Então $\mathbf{y}_i^\top = \{y_{ij} : j \in s_i\}$ denota um vetor $n(s_i) \times 1$ contendo as respostas (taxas ou proporções), para $i = 1, \ldots, n$. Vamos assumir que $y_{ij} \sim \text{UL}(\mu_{ij})$ com componente de regressão $g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta}$, em que $g(\cdot)$ denota a função de ligação com domínio no $(0, 1)$ e diferenciável, $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^\top$ contém os valores das variáveis explicativas e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ o vetor com os coeficientes da regressão. A matriz de correlação de cada unidade experimental pode ser representada pela matriz $\mathbf{R}(\mathbf{u}_i)$ de dimensão $n(s_i) \times n(s_i)$, em que $\mathbf{u}_i^\top = \{u_{ij} : j \in s_i\}$ e

$$u_{ij} = \frac{z_{ij}}{\mu_{ij}^2} - \frac{1 + \mu_{ij}}{\mu_{ij}(1 - \mu_{ij})},$$

com $z_{ij} = y_{ij}/(1 - y_{ij})$, para $i = 1, \ldots, n$. Para $n(s_i) > 1$, os elementos da matriz $\mathbf{R}(\mathbf{u}_i)$ podem ser denotados como

$$\text{R}_{jj'}(\mathbf{u}_i) = \frac{\text{E}(u_{ij} u_{ij'})}{\sqrt{\text{Var}(u_{ij})}\sqrt{\text{Var}(u_{ij'})}},$$

para $j \neq j'$, em que $\text{R}_{jj'}(\mathbf{u}_i) = 1$, para $j = j'$, com

$$\text{Var}(u_{ij}) = \frac{2 - (1 - \mu_{ij})^2}{\mu_{ij}^2 (1 - \mu_{ij})^2},$$

para $i = 1, \ldots, n$. Note que u_{ij} consiste de uma combinação linear de z_{ij}, então $\mathbf{R}(\mathbf{u}_i)$ deve concordar com a matriz de correlação $\mathbf{R}(\mathbf{z}_i)$, com $\mathbf{z}_i^\top = \{z_{ij} : j \in s_i\}$, e devido à relação monotônica entre z_{ij} e y_{ij} é esperado uma boa concordância entre as matrizes de correlação $\mathbf{R}(\mathbf{z}_i)$ e $\mathbf{R}(\mathbf{y}_i)$ (Wicklin, 2013).

Baseado na teoria desenvolvida por Godambe (1997), na próxima seção derivamos uma classe de funções de estimação para $\boldsymbol{\beta}$ e um algoritmo de escore de Newton para estimar os coeficientes da regressão assumindo que a matriz $\mathbf{R}(\mathbf{u}_i)$ é substituída por alguma matriz de correlação estruturada (matriz de correlação de trabalho) que não envolve $\boldsymbol{\beta}$. Propriedades assintóticas dos estimadores obtidos para os coeficientes também são apresentadas. Denominamos essa classe de modelos UL-GEE.

2.1 Funções de estimação

As conhecidas equações de estimação generalizadas (GEE) (Liang e Zeger, 1986) podem ser estendidas para modelar taxas e proporções com distribuição marginal UL e alguma dependência entre as respostas dentro das unidades experimentais representada por uma matriz de correlação de trabalho. Detalhes sobre GEE são descritos, por exemplo, em Venezuela (2003). Em geral, tais equações de estimação podem ser derivadas como um caso particular da classe de funções de estimação ótima proposta por Crowder (1987) e definida como

$$\boldsymbol{\Psi}^*(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathrm{E}\left(\frac{\partial \mathbf{u}_i}{\partial \boldsymbol{\beta}^\top}\right)^\top \mathrm{Cov}(\mathbf{u}_i)^{-1}\mathbf{u}_i.$$

Depois de algumas manipulações algébricas (veja Apêndice A.3) obtemos

$$\boldsymbol{\Psi}^*(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{X}_i^\top \mathbf{D}_i \mathrm{Cov}(\mathbf{u}_i)^{-1}\mathbf{u}_i, \tag{2.1}$$

com $\mathbf{X}_i$ sendo uma matriz $n(s_i) \times p$ de linhas $\mathbf{x}_{ij}^\top$, $\mathbf{D}_i = \mathrm{diag}\{d_{ij} : j \in s_i\}$ com $d_{ij} = -\mathrm{Var}(u_{ij})\{g'(\mu_{ij})\}^{-1}$ e $\mathrm{Cov}(\mathbf{u}_i)$ denota a matriz de covariância de $\mathbf{u}_i$, para $i = 1, \ldots, n$.

Expressando $\mathrm{Cov}(\mathbf{u}_i) = \Sigma_{u_i}^{\frac{1}{2}}\mathbf{R}(\mathbf{u}_i)\Sigma_{u_i}^{\frac{1}{2}}$, em que $\Sigma_{u_i} = \mathrm{diag}\{\mathrm{Var}(u_{ij}) : j \in s_i\}$, a ideia consiste em substituir a matriz de correlação $\mathbf{R}(\mathbf{u}_i)$ por uma matriz de correlação de trabalho $\mathbf{R}_i(\boldsymbol{\rho})$, que envolve apenas do vetor de correlações $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_q)^\top$, não envolvendo $\boldsymbol{\beta}$. Então, a função de estimação da expressão (2.1) assume a forma alternativa

$$\boldsymbol{\Psi}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{X}_i^\top \mathbf{D}_i \boldsymbol{\Omega}_i^{-1}\mathbf{u}_i, \tag{2.2}$$

com $\boldsymbol{\Omega}_i = \Sigma_{u_i}^{\frac{1}{2}}\mathbf{R}_i(\boldsymbol{\rho})\Sigma_{u_i}^{\frac{1}{2}}$, para $i = 1, \ldots, n$.

Note que $\mathrm{E}(\mathbf{u}_i) = \{\mathrm{E}(u_{ij}) : j \in s_i\}^\top = 0_{n(s_i)\times 1}$, para todo i, assim temos uma função de estimação não viesada, ou seja, $\mathrm{E}\{\boldsymbol{\Psi}(\boldsymbol{\beta})\} = 0_{p\times 1}$. De Godambe (1997) a matriz de variabilidade de $\boldsymbol{\Psi}(\boldsymbol{\beta})$ fica definida como $\mathbf{V}_{n\Psi}(\boldsymbol{\beta}) = \mathrm{E}\{\boldsymbol{\Psi}(\boldsymbol{\beta})\boldsymbol{\Psi}^\top(\boldsymbol{\beta})\}$, e a respectiva matriz de sensibilidade fica dada por $\mathbf{S}_{n\Psi}(\boldsymbol{\theta}) = \mathrm{E}\{\boldsymbol{\Psi}'(\boldsymbol{\beta})\}$. A matriz de informação de Godambe de $\boldsymbol{\beta}$ consiste em uma função de estimação regular definida como

$$\mathbf{J}_{n\Psi}(\boldsymbol{\beta}) = \mathbf{S}_{n\Psi}(\boldsymbol{\beta})^\top \mathbf{V}_{n\Psi}^{-1}(\boldsymbol{\beta})\mathbf{S}_{n\Psi}(\boldsymbol{\theta}),$$

em que $\mathbf{V}_{n\Psi}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{V}_i(\boldsymbol{\beta})$ com $\mathbf{V}_i(\boldsymbol{\beta}) = \mathbf{X}_i^\top \mathbf{W}_i \mathbf{D}_i^{-1}\mathrm{Cov}(\mathbf{u}_i)\mathbf{D}_i^{-1}\mathbf{W}_i\mathbf{X}_i$, $\mathbf{S}_{n\Psi}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{S}_i(\boldsymbol{\beta})$ com $\mathbf{S}_i(\boldsymbol{\beta}) = \mathbf{X}_i^\top \mathbf{W}_i \mathbf{X}_i$ e $\mathbf{W}_i = \mathbf{D}_i\boldsymbol{\Omega}_i^{-1}\mathbf{D}_i$, para $i = 1, \ldots, n$ (veja as manipulações no Apêndice A.4).

Na próxima seção está derivado um processo iterativo reponderado para resolver $\boldsymbol{\Psi}(\hat{\boldsymbol{\beta}}) = 0_{p\times 1}$ da equação (2.2), e algumas propriedades assintóticas dos estimadores obtidos $\hat{\boldsymbol{\beta}}$ são apresentadas.

2.2 Processo iterativo

Analogamente a Tsuyuguchi et al. (2020) aplicamos o método escore de Newton, que é similar com o método escore de Fisher (veja, por exemplo, Jørgensen et al., 1996), para obtermos a estimativa $\hat{\boldsymbol{\beta}}$, em que $\boldsymbol{\Psi}'(\boldsymbol{\beta})$ fica substituída por sua esperança $\mathrm{E}\{\boldsymbol{\Psi}'(\boldsymbol{\beta})\} = \sum_{i=1}^{n} \mathbf{X}_i^\top \mathbf{W}_i \mathbf{X}_i$. Uma vantagem desse método está na existência, em cada passo do processo iterativo, da inversa $[\mathrm{E}\{\boldsymbol{\Psi}'(\boldsymbol{\beta}^{(m)})\}]^{-1}$, desde que cada $\mathbf{X}_i$ tenha posto coluna completo. Então, fixando $\boldsymbol{\rho}$, obtemos o seguinte processo iterativo:

$$\begin{aligned}\boldsymbol{\beta}^{(m+1)} &= \boldsymbol{\beta}^{(m)} - [\mathrm{E}\{\boldsymbol{\Psi}'(\boldsymbol{\beta}^{(m)})\}]^{-1}\boldsymbol{\Psi}(\boldsymbol{\beta}^{(m)}) \\ &= \boldsymbol{\beta}^{(m)} - \{\sum_{i=1}^{n} \mathbf{X}_i^\top \mathbf{W}_i^{(m)} \mathbf{X}_i\}^{-1}\{\sum_{i=1}^{n} \mathbf{X}_i^\top \mathbf{W}_i^{(m)} (\mathbf{D}_i^{(m)})^{-1}\mathbf{u}_i^{(m)}\},\ m = 0, 1, 2 \ldots, \end{aligned} \tag{2.3}$$

O processo iterativo da expressão (2.3) pode ser expresso como o seguinte processo iterativo reponderado:

$$\boldsymbol{\beta}^{(m+1)} = \{\sum_{i=1}^{n} \mathbf{X}_i^\top \mathbf{W}_i^{(m)} \mathbf{X}_i\}^{-1}\{\sum_{i=1}^{n} \mathbf{X}_i^\top \mathbf{W}_i^{(m)} \mathbf{t}_i^{(m)}\},\ m = 0, 1, 2 \ldots, \tag{2.4}$$

em que $\mathbf{t}_i = \mathbf{X}_i\boldsymbol{\beta} - \mathbf{D}_i^{-1}\mathbf{u}_i$ consiste na variável resposta modificada, para $i = 1, \ldots, n$. Abaixo descrevemos os estimadores de momentos de $\boldsymbol{\rho}$ fixando $\boldsymbol{\beta}$, para estruturas usuais de correlação desde que $n(s_i) > 1$:

1. Independente: Neste caso temos $\mathbf{R}_i(\boldsymbol{\rho}) = \mathbf{I}_{n(s_i)}$, com $\mathbf{I}_{n(s_i)}$ denotando a matriz identidade de ordem $n(s_i)$.

2. Não estruturada: Aqui a matriz de correlação $\mathbf{R}_i(\boldsymbol{\rho})$ não tem estrutura e temos $n(s_i)\{n(s_i) - 1\}/2$ parâmetros da estrutura de correlação não únicos para estimar em cada unidade experimental. Seja o conjunto $A_{jj'} = \{i : j, j' \in s_i,\ i = 1, \ldots, n,\ j \neq j'\}$, denotando $\mathbf{R}_i = \{\rho_{ijj'}\}$, os elementos de $\mathbf{R}_i$ pode ser estimados por

$$\hat{\rho}_{jj'} = \frac{1}{n(A_{jj'})} \sum_{i \in A_{jj'}} \frac{\hat{u}_{ij}}{\sqrt{\widehat{\mathrm{V}}\mathrm{ar}(u_{ij})}} \frac{\hat{u}_{ij'}}{\sqrt{\widehat{\mathrm{V}}\mathrm{ar}(u_{ij'})}}.$$

3. Simétrica: Neste caso $\mathbf{R}_i = \mathbf{R}_i(\rho)$, em que os elementos de $\mathbf{R}_i$ ficam dados por $\mathrm{R}_{ijj'} = 1$, para $j = j'$, e $\mathrm{R}_{ijj'} = \rho$, para $j \neq j'$. Um estimador consistente para ρ pode ser expresso como

$$\hat{\rho} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n(s_i)\{n(s_i) - 1\}} \sum_{j \in s_i} \sum_{\substack{j' \in s_i \\ j' \neq j}} \frac{\hat{u}_{ij}}{\sqrt{\widehat{\mathrm{V}}\mathrm{ar}(u_{ij})}} \frac{\hat{u}_{ij'}}{\sqrt{\widehat{\mathrm{V}}\mathrm{ar}(u_{ij'})}}.$$

4. Autoregressiva de primeira ordem: Aqui assumimos que $\mathbf{R}_i = \mathbf{R}_i(\rho)$, em que os elementos de $\mathbf{R}_i$ ficam dados por $\mathrm{R}_{ijj'} = 1$, para $j = j'$, e $\mathrm{R}_{ijj'} = \rho^{|j-j'|}$, para $j \neq j'$.

Seja o conjunto $A_j = \{i : j, j+1 \in s_i,\ i = 1, \dots, n\}$ e $B = \bigcap_{i=1}^{n} s_i$, um estimador consistente para ρ pode ser expresso por

$$\hat{\rho} = \frac{1}{\{n(B)-1\}} \sum_{j \in B} \frac{1}{n(A_j)} \sum_{i \in A_j} \frac{\hat{u}_{ij}}{\sqrt{\widehat{\mathrm{V}}\mathrm{ar}(u_{ij})}} \frac{\hat{u}_{i(j+1)}}{\sqrt{\widehat{\mathrm{V}}\mathrm{ar}(u_{i(j+1)})}}.$$

Então, denotando $\mathbf{X} = [\mathbf{X}_1^\top, \dots, \mathbf{X}_n^\top]^\top$ uma matriz $N \times p$ de linhas $\mathbf{X}_i$ e $g(\mathbf{y}) = [g(\mathbf{y}_1^\top), \dots, g(\mathbf{y}_n^\top)]^\top$ um vetor $N \times 1$ de elementos $g(\mathbf{y}_i) = \{g(y_{ij}) : j \in s_i\}^\top$, em que $N = \sum_{i=1}^{n} n(s_i)$, propomos o Algoritmo 1 para obter as estimativas dos coeficientes da regressão.

Algoritmo 1 Estimativas dos coeficientes

1: **Entradas:**
   $\mathbf{X}$ com posto coluna completo
2: **Inicializar:**
   $\boldsymbol{\beta}^{(0)} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top g(\mathbf{y})$
3: **repetir**
4: atualize $\boldsymbol{\rho}$ de alguma estrutura fixada
5: atualize $\boldsymbol{\beta}$ do processo iterativo
6: **até** $\max_i |\beta_i^{(m)} - \beta_i^{(m-1)}| / \beta_i^{(m-1)} < \epsilon$
7: **retornar** $\boldsymbol{\beta}^{(m)}$ e $\boldsymbol{\rho}^{(m)}$

2.3 Inferência

Similar a Tsuyuguchi et al. (2020), de Artes e Jørgensen (2000), temos que $\hat{\boldsymbol{\beta}}$ obtido do processo iterativo da expressão (2.4) é tal que

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} \mathrm{N}_p\{0_{p\times 1}, \mathbf{J}_\Psi^{-1}(\boldsymbol{\beta})\},$$

em que $\mathbf{J}_\Psi(\boldsymbol{\beta}) = \lim_{n\to\infty} n^{-1}\mathbf{J}_{n\Psi}(\boldsymbol{\beta})$, com $\mathbf{J}_{n\Psi}(\boldsymbol{\beta}) = \{\sum_{i=1}^{n} \mathrm{S}_i(\boldsymbol{\beta})\}\{\sum_{i=1}^{n} \mathbf{V}_i(\boldsymbol{\beta})\}^{-1}\{\sum_{i=1}^{n} \mathrm{S}_i(\boldsymbol{\beta})\}$.

Um estimador consistente da matriz de covariância de $\hat{\boldsymbol{\beta}}$ fica dado por

$$\{\widehat{\mathbf{J}}_\Psi(\boldsymbol{\beta})\}^{-1} = \{\sum_{i=1}^{n} \widehat{\mathrm{S}}_i(\boldsymbol{\beta})\}^{-1}\{\sum_{i=1}^{n} \mathbf{X}_i^\top \widehat{\mathbf{D}}_i \widehat{\boldsymbol{\Omega}}_i^{-1} \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^\top \widehat{\boldsymbol{\Omega}}_i^{-1} \widehat{\mathbf{D}}_i \mathbf{X}_i\}\{\sum_{i=1}^{n} \widehat{\mathrm{S}}_i(\boldsymbol{\beta})\}^{-1},$$

com $\widehat{\mathrm{S}}_i(\boldsymbol{\beta}) = \mathbf{X}_i^\top \widehat{\mathbf{W}}_i \mathbf{X}_i$, em que $\widehat{\mathbf{W}}_i = \widehat{\mathbf{D}}_i \widehat{\boldsymbol{\Omega}}_i^{-1} \widehat{\mathbf{D}}_i$, $\widehat{\boldsymbol{\Omega}}_i = \widehat{\boldsymbol{\Sigma}}_{u_i}^{\frac{1}{2}} \mathbf{R}_i(\hat{\boldsymbol{\rho}}) \widehat{\boldsymbol{\Sigma}}_{u_i}^{\frac{1}{2}}$, $\widehat{\mathbf{D}}_i = \mathrm{diag}\{\hat{d}_j : j \in s_i\}$ e $\hat{d}_{ij} = \{g(\hat{\mu}_{ij})\}^{-1} \widehat{\mathrm{V}}\mathrm{ar}(u_{ij})$, para $i = 1, \dots, n$.

O teste de hipótese $\mathrm{H}_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{m}$ contra $\mathrm{H}_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{m}$, em que $\mathbf{C}$ uma matriz $r \times p$ de posto linha r ($r \leq p$) pode ser avaliado com uma estatística do tipo Wald, cuja expressão fica dada por $\xi_W = (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{m})^\top [\mathbf{C}\{\widehat{\mathbf{J}}_\Psi(\boldsymbol{\beta})\}^{-1}\mathbf{C}^\top]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{m})$. Para amostras grandes e sob condições usuais de regularidade, segue que $\xi_W \sim \chi_r^2$, em que χ_r^2 denota a distribuição qui-quadrado com r graus de liberdade.

2.4 Diagnóstico

A verificação da adequação modelo consiste em uma série de procedimentos de diagnóstico para avaliar as suposições feitas para o modelo, bem como detectar a existência de observações discrepantes e a sensibilidade dos coeficientes estimados sob perturbações feitas no modelo ou nos dados. No contexto de equações de estimação há uma vasta literatura, porém focada nos procedimentos desenvolvidos para equações de estimação generalizadas (veja, por exemplo, Preisser e Qaqish, 1996; Venezuela et al., 2011; Hardin e Hilbe, 2012 e Manghi et al., 2019). Uma extensão de tais procedimentos para equações de estimação de Godambe foi abordado para modelos Birnbaum-Saunders-GEE (Tsuyuguchi et al., 2020). Então, baseados nesse trabalho, derivamos nesta seção alguns procedimentos de diagnóstico para modelos UL-GEE.

2.4.1 Análise de resíduos

Para verificar as suposições feitas para o modelo UL-GEE, particularmente sobre a distribuição marginal com a estrutura de correlação proposta, e para detectar a presença de observações aberrantes, consideramos o resíduo quantílico marginal (Dunn e Smyth, 1996) definido como

$$r_{q_{ij}} = \Phi^{-1}\{F(y_{ij}; \hat{\mu}_{ij})\},$$

em que

$$F(y_{ij}; \hat{\mu}_{ij}) = 1 - \left(\frac{1 - y_{ij}\hat{\mu}_{ij}}{1 - y_{ij}}\right) \exp\left\{-\frac{(1 - \hat{\mu}_{ij})y_{ij}}{(1 - y_{ij})\hat{\mu}_{ij}}\right\},$$

denota a função densidade acumulada de $y_{ij} \sim \text{UL}(\mu_{ij})$ avaliada em $\hat{\mu}_{ij} = g^{-1}(\hat{\eta}_{ij})$ e $\Phi(\cdot)$ a função densidade acumulada da distribuição normal padrão, para $i = 1, \ldots, n$. Sob a hipótese de independência entre u_{ij} e $u_{ij'}$, para $j \neq j'$, temos que $r_{q_{ij}}$ tem distribuição assintótica N$(0, 1)$. Entretanto, na prática, pode-se ter observações correlacionadas nas unidades experimentais, então alguma banda de confiança empírica gerada, por exemplo com simulações de Monte Carlo, precisa ser adicionada no gráfico normal de probabilidade dos resíduos quantílicos marginais $r_{q_{ij}}$. Assim, fugas da banda de confiança empírica podem indicar que a suposição de distribuição marginal UL com a matriz de correlação de trabalho proposta é inadequada para ajustar os dados. Em adição, o gráfico pode revelar observações aberrantes. Podemos aplicar, para gerar a banda empírica, o mesmo algoritmo proposto na próxima seção para o estudo de simulação.

Alternativamente, pode-se selecionar n resíduos, denotados por $r^*_{q_1}, \ldots, r^*_{q_n}$, com $r^*_{q_i}$ sendo escolhido aleatoriamente do subconjunto de resíduos $\{r_{q_{ij}} : j \in s_i\}$ da i-ésima unidade experimental, $i = 1, \ldots, n$. Assim, temos que $r^*_{q_1}, \ldots, r^*_{q_n}$ são independentes e têm distribuição assintótica N$(0, 1)$. Com tal subconjunto de resíduos podemos gerar vários gráficos de resíduos, como o resíduo quantílico contra o valor ajustado, o gráfico normal de probabilidade e o *worm plot* (Buuren e Fredriks, 2001) como no GAMLSS (veja, por exemplo, Stasinopoulos et al., 2017). Uma vez que existem $\prod_{i=1}^{n} n(s_i)$ possíveis subconjuntos de resíduos, podemos exibir os gráficos de m subconjuntos diferentes. Similarmente ao GAMLSS, no caso de modelos discretos em que são gerados m gráficos de resíduos quantílicos aleatorizados, sugerimos $m \geq 8$ gráficos. Contudo, estudos mais teóricos são necessários para a determinação do valor apropriado de gráficos em cada caso.

2.4.2 Estudos de sensibilidade

A ideia dos estudos de sensibilidade consiste em verificar a influência das observações nos coeficientes da regressão estimados sob perturbações feitas no modelo ou nos dados. O principal objetivo deste tipo de estudo está em detectar observações que têm influência desproporcional nas estimativas dos coeficientes, particularmente com mudanças inferências. Entretanto, tais observações podem ser mascaradas, requerendo uma análise cuidadosa dos gráficos de influência. Há vários procedimentos desenvolvidos para modelos de regressão, tais como a tradicional deleção de casos (veja, por exemplo, Cook e Weisberg, 1982), influência local (Cook, 1986), influência local conformal (Poon e Poon, 1999) e *forward search* (Atkinson e Riani, 2000), entre outros. No contexto de equações de estimação Hardin e Hilbe (2012) apresentam uma revisão para GEE, enquanto Tsuyuguchi et al. (2020) desenvolvem algumas extensões para modelos fora da família exponencial. Baseado nesse último trabalho derivamos nesta seção a curvatura normal conformal para modelos UL-GEE. Medidas baseadas na remoção de observações são aplicadas apenas na análise confirmatória das observações destacadas pela influência local conformal.

A função log-verossimilhança, usualmente aplicada em modelos em que a verossimilhança é completamente conhecida, fica substituída nas equações de estimação pela função de ajuste $\mathcal{F}(\boldsymbol{\beta})$ (Cadigan e Farrell, 2002), que assume-se duas vezes diferenciável em $\boldsymbol{\beta}$ com única estimativa de coeficiente interno e definida como

$$\boldsymbol{\Psi}(\hat{\boldsymbol{\beta}}) = \left\{ \frac{\partial \mathcal{F}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right\} \Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 0_{p\times 1}.$$

Então, uma medida de influência apropriada fica dada por $\mathrm{FD}_{\omega} = 2\{\mathcal{F}(\hat{\boldsymbol{\beta}}) - \mathcal{F}(\hat{\boldsymbol{\beta}}_{\omega})\}$, com $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_N)^{\top}$ denotando o vetor de perturbações e $\hat{\boldsymbol{\beta}}_{\omega}$ consiste na solução da equação de estimação perturbada $\boldsymbol{\Psi}(\hat{\boldsymbol{\beta}}_{\omega}|\boldsymbol{\omega}) = 0_{p\times 1}$. O vetor de não perturbação $\boldsymbol{\omega}_0$ fica definido de tal forma que $\boldsymbol{\Psi}(\boldsymbol{\beta}_{\omega}|\boldsymbol{\omega}_0) = \boldsymbol{\Psi}(\boldsymbol{\beta})$.

Poon e Poon (1999) derivaram a curvatura normal conformal na direção unitária ℓ, expressa por

$$\mathrm{B}_{\ell}(\boldsymbol{\beta}) = |\ell^{\top}\mathbf{B}\ell| / \sqrt{\mathrm{tr}(\mathbf{B}^2)},$$

em que $0 \leq \mathrm{B}_{\ell}(\boldsymbol{\beta}) \leq 1$, $\mathbf{B} = \boldsymbol{\Delta}^{\top}\{\ddot{\mathcal{F}}(\boldsymbol{\beta})\}^{-1}\boldsymbol{\Delta}$ uma matriz simétrica positiva semi-definida com $\boldsymbol{\Delta} = \partial\boldsymbol{\Psi}(\boldsymbol{\beta}|\boldsymbol{\omega})/\partial\boldsymbol{\omega}^{\top}$ sendo avaliada em $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, $\boldsymbol{\rho} = \hat{\boldsymbol{\rho}}$ e $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. Gráficos de influência baseados na medida agregada dos autovalores não nulos e os correspondentes autovetores da matriz $\mathbf{B}$ foram propostos por Poon e Poon (1999). Em particular, consideramos a medida agregada B_{ij}, que corresponde à curvatura normal conformal avaliada na direção ℓ_{ij} da (i, j)-ésima observação, em que ℓ_{ij} denota um vetor $N \times 1$ de zeros com um na (i, j)-ésima posição. Lee e Xu (2004) sugerem destacar possíveis observações influentes tais que $\mathrm{B}_{ij} > \bar{\mathrm{B}} + \mathrm{SD(B)}c^*$, com $\bar{\mathrm{B}}$ e SD(B) denotando, respectivamente, a média e o desvio-padrão de $\{\mathrm{B}_{ij}, j \in s_i; i = 1, \ldots, n\}$ com c^* sendo escolhido apropriadamente.

Para verificar o efeito das observações destacadas sob o esquema de perturbação adotado, aplicamos o MRC (do inglês *Maximum Relative Chance*) proposta por Lee (2006)

e expressa como

$$\text{MRC} = \max_{1\leq k\leq p}\left|\frac{\hat{\beta}_k - \hat{\beta}_k^0}{\hat{\beta}_k}\right|,$$

com $\hat{\beta}_k^0$ denotando a estimativa de β_k depois de remover as observações destacadas na análise de diagnóstico. O critério consiste em comparar o MRC com os valores de MRC obtidos usando observações não destacadas. Se o valor de MRC referente ao grupo de observações destacadas for muito superior aos valores de MRC do grupo de observações não destacadas, há um indício de confirmação da influência das observações destacadas nas estimativas dos coeficientes.

A fim de auxiliar na escolha da estrutura de correlação adequada, estendemos o Critério de Independência de Quase-verossimilhança (QIC) (veja, por exemplo, Hardin e Hilbe, 2012) para a classe UL-GEE. A respectiva medida pode ser expressa como

$$\text{QIC} = -2\sum_{i=1}^{n}\sum_{j\in s_i}\log\{f(y_{ij};\hat{\mu}_{ij})\} + 2\text{tr}[\{\widehat{\mathbf{J}}_\Psi(\boldsymbol{\beta})\}^{-1}\{\sum_{i=1}^{n}\hat{\mathbf{S}}_{iI}(\boldsymbol{\beta})\}],$$

em que $\hat{\mu}_{ij}$ consiste na estimativa para uma estrutura de correlação específica $\mathbf{R}_i(\boldsymbol{\rho})$ e $\mathbf{S}_{iI}(\boldsymbol{\beta})$ denota a matriz $\mathbf{S}_i(\boldsymbol{\beta})$ avaliada sob a estrutura de correlação independente. O critério consiste na seleção da estrutura de correlação tal que QIC seja minimizado. Na sequência vamos derivar a matriz $\boldsymbol{\Delta}$ para dois esquemas usuais de perturbação.

2.4.3 Perturbação: ponderação de casos

Sob o esquema de perturbação de ponderação de casos a função de estimação para $\boldsymbol{\beta}$ fica expressa como

$$\boldsymbol{\Psi}(\boldsymbol{\beta}|\boldsymbol{\omega}) = \sum_{i=1}^{n}\mathbf{X}_i^\top\mathbf{W}_i\mathbf{D}_i^{-1}\text{diag}(\boldsymbol{\omega}_i)\mathbf{u}_i, \tag{2.5}$$

em que $\boldsymbol{\omega}_i^\top = \{\omega_{ij} : j \in s_i\}$ denota as perturbações aplicadas nos elementos da i-ésima unidade experimental, $0 \leq \omega_{ij} \leq 1$, para $i = 1,\ldots,n$, com $\boldsymbol{\omega} = (\boldsymbol{\omega}_1^\top,\ldots,\boldsymbol{\omega}_n^\top)^\top$. O vetor de não perturbação $\boldsymbol{\omega}_0$ é dado por um vetor $N \times 1$ formado por 1's. Pode-se reescrever a função de estimação da equação (2.5) na seguinte forma matricial:

$$\boldsymbol{\Psi}(\boldsymbol{\beta}|\boldsymbol{\omega}) = \mathbf{X}^\top\mathbf{W}\mathbf{D}^{-1}\text{diag}(\boldsymbol{\omega})\mathbf{u},$$

com $\mathbf{W} = \text{blockdiag}\{\mathbf{W}_1,\ldots,\mathbf{W}_n\}$, $\mathbf{D} = \text{blockdiag}\{\mathbf{D}_1,\ldots,\mathbf{D}_n\}$, $\mathbf{X} = (\mathbf{X}_1^\top,\ldots,\mathbf{X}_n^\top)^\top$ e $\mathbf{u} = (\mathbf{u}_1^\top,\ldots,\mathbf{u}_n^\top)^\top$. Consequentemente, obtemos

$$\boldsymbol{\Delta} = \left.\frac{\partial\boldsymbol{\Psi}(\boldsymbol{\beta}|\boldsymbol{\omega})}{\partial\boldsymbol{\omega}^\top}\right|_{(\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\boldsymbol{\rho}=\hat{\boldsymbol{\rho}},\boldsymbol{\omega}=\boldsymbol{\omega}_0)} = \mathbf{X}^\top\widehat{\mathbf{W}}\widehat{\mathbf{D}}^{-1}\text{diag}(\hat{\mathbf{u}}).$$

Então, a matriz $\mathbf{B}$ é aproximada por $\boldsymbol{\Delta}^\top[\text{E}\{\ddot{\mathcal{F}}(\hat{\boldsymbol{\beta}})\}]^{-1}\boldsymbol{\Delta}$ que fica expressa como $\text{diag}(\hat{\mathbf{u}})\widehat{\mathbf{D}}^{-1}\widehat{\mathbf{W}}\mathbf{X}\{\mathbf{X}^\top\widehat{\mathbf{W}}\mathbf{X}\}^{-1}\mathbf{X}^\top\widehat{\mathbf{W}}\widehat{\mathbf{D}}^{-1}\text{diag}(\hat{\mathbf{u}})$.

2.4.4 Perturbação: resposta

Neste caso, usualmente realiza-se a seguinte perturbação em cada resposta observada:

$$y_{\omega_{ij}} = y_{ij} + \omega_{ij}\sigma_{ij},$$

em que $\omega_{ij} \in R$, de modo que $0 < y_{ij} < 1$ e σ_{ij} denota o desvio-padrão de y_{ij}. Note que a curvatura normal conformal será avaliada em $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\rho}}, \boldsymbol{\omega}_0)$ de modo que $0 \leq y_{\omega_{ij}} \leq 1$. Sob tal esquema de pertubação, a função de estimação para $\boldsymbol{\beta}$ pode ser expressa de forma matricial como

$$\boldsymbol{\Psi}(\boldsymbol{\beta}|\boldsymbol{\omega}) = \mathbf{X}^\top \mathbf{W}\mathbf{D}^{-1}\text{diag}(\boldsymbol{\omega})\mathbf{u}_\omega,$$

com $\boldsymbol{\omega} = (\boldsymbol{\omega}_1^\top, \ldots, \boldsymbol{\omega}_n^\top)^\top$, $\boldsymbol{\omega}_i^\top = \{\omega_{ij} : j \in s_i\}$, $\mathbf{u}_\omega = (\mathbf{u}_{\omega_1}^\top, \ldots, \mathbf{u}_{\omega_n}^\top)^\top$, $\mathbf{u}_{\omega_i}^\top = \{u_{\omega_{ij}} : j \in s_i\}$ e

$$u_{\omega_{ij}} = \frac{z_{\omega_{ij}}}{\mu_{ij}^2} - \frac{(1+\mu_{ij})}{\mu_{ij}(1-\mu_{ij})},$$

com $z_{\omega_{ij}} = y_{\omega_{ij}}/(1 - y_{\omega_{ij}})$. Pode ser mostrado que $\partial u_{\omega_{ij}}/\partial\omega_{ij} = f_{ij}$, em que $f_{ij} = \sigma_{ij}/\{\mu_{ij}^2(1 - y_{ij})^2\}$, para $i = 1, \ldots, n$.

Logo, obtemos

$$\boldsymbol{\Delta} = \left.\frac{\partial\boldsymbol{\Psi}(\boldsymbol{\beta}|\boldsymbol{\omega})}{\partial\boldsymbol{\omega}^\top}\right|_{(\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\boldsymbol{\rho}=\hat{\boldsymbol{\rho}},\boldsymbol{\omega}=\boldsymbol{\omega}_0)} = \mathbf{X}^\top\widehat{\mathbf{W}}\widehat{\mathbf{D}}^{-1}\text{diag}(\hat{\mathbf{f}}),$$

em que $\mathbf{f} = (\mathbf{f}_1^\top, \ldots, \mathbf{f}_n^\top)^\top$ com $\mathbf{f}_i^\top = \{f_{ij} : j \in s_i\}$, para $i = 1, \ldots, n$. Então, a matriz $\mathbf{B}$ é também aproximada por $\boldsymbol{\Delta}^\top[\text{E}\{\ddot{\mathcal{F}}(\hat{\boldsymbol{\beta}})\}]^{-1}\boldsymbol{\Delta}$ que fica expressa por $\text{diag}(\hat{\mathbf{f}})\widehat{\mathbf{D}}^{-1}\widehat{\mathbf{W}}\mathbf{X}\{\mathbf{X}^\top\widehat{\mathbf{W}}\mathbf{X}\}^{-1}\mathbf{X}^\top\widehat{\mathbf{W}}\widehat{\mathbf{D}}^{-1}\text{diag}(\hat{\mathbf{f}})$.

Capítulo 3

Simulação e aplicação

Neste capítulo apresentamos estudos de simulação para avaliar as propriedades empíricas dos estimadores e também uma aplicação usando um conjunto de dados reais para ilustrar a metodologia desenvolvida no capítulo anterior.

3.1 Estudo de simulação

Para verificarmos o comportamento em relação ao tamanho da amostra dos estimadores obtidos do processo descrito na Seção 2.2, apresentamos um estudo de simulação (veja, os códigos, no Apêndice B.1) baseado no seguinte modelo UL-GEE:

1. $y_{ij}|x_{ij} \sim \text{UL}(\mu_{ij})$,
2. $\Phi^{-1}(\mu_{ij}) = \beta_0 + \beta_1 x_{ij}$,
3. $\mathbf{R}_i(\boldsymbol{\rho}) = \mathbf{R}_i(\rho)$,

em que $\Phi(\cdot)$ denota a função densidade acumulada da distribuição normal padrão, a matriz de correlação $\mathbf{R}_i(\rho)$ segue as estruturas simétrica e autoregressiva entre os elementos de $\mathbf{z}_i^\top = \{z_{ij} : j \in s_i\}$. O coeficiente de correlação ρ e x_{ij}'s são valores fixos gerados de uma distribuição uniforme no intervalo $[0, 1]$, para $s_i = \{1, \ldots, s\}$ e $i = 1, \ldots, n$. Os valores escolhidos dos coeficientes foram $\beta_0 = -3$ e $\beta_1 = 6$, $\rho = -0.1, 0.3, 0.7$, para $n = 10, 50, 500$ e $s = 3, 5, 10$, enquanto o viés (em valor absoluto) e o erro quadrático médio (EQM) foram calculados para cada cenário considerado. O modelo e valores dos coeficientes foram escolhidos de modo que, sob cada esquema de simulação, todo o espaço paramétrico é coberto uniformemente e não apenas uma região específica. Também consideramos para cada cenário uma correlação negativa para ilustrar um caso que o modelo misto não pode lidar.

O viés e o EQM para o coeficiente $\hat{\theta}$ foram calculados, respectivamente, como $|\bar{\hat{\theta}} - \theta_0|$ e $R^{-1}\sum_{r=1}^{R}(\hat{\theta}^{(r)} - \theta_0)^2$, em que $\bar{\hat{\theta}} = R^{-1}\sum_{r=1}^{R}\hat{\theta}^{(r)}$ com $\hat{\theta}^{(r)}$ sendo a estimativa de θ da r-ésima réplica e θ_0 denota o verdadeiro valor do coeficiente. Um total de $R = 5000$ réplicas foi considerado para θ sendo β_0, β_1 ou ρ (veja os resultados nas Tabelas 3.1-3.4). Denotando $\text{N}_s\{0_{s\times 1}, \mathbf{R}_i(\rho)\}$ a distribuição normal s-variada de média $0_{s\times 1}$ e matriz de variância $\mathbf{R}_i(\rho)$,

$\Phi(\mathbf{z}_i^*) = \{\Phi(z_{i1}^*), \ldots, \Phi(z_{is}^*)\}^\top$ com $\Phi(\cdot)$ denotando a função densidade acumulada da $N(0,1)$ e $F^{-1}(\cdot;\boldsymbol{\mu}_i) = \{F^{-1}(\cdot;\mu_{i1}), \ldots, F^{-1}(\cdot;\mu_{is})\}^\top$ com $F(\cdot;\mu_{ij})$ sendo a função densidade acumulada da distribuição $\text{UL}(\mu_{ij})$. Aplicamos o Algoritmo 2 para simular s valores correlacionados da distribuição marginal $\text{UL}(\mu_{i1}), \ldots, \text{UL}(\mu_{is})$ usando cópulas Gaussianas (veja, por exemplo, Wicklin, 2013) e ajustamos o modelo UL-GEE com a estrutura de correlação especificada sendo correta ou incorreta para verificar o comportamento das estimativas dos coeficientes.

Implementamos o estudo de simulação no software R (R Core Team, 2022) usando os seguintes pacotes adicionais: `lamW` (Adler, 2022), `expint` (Goulet, 2022), `pbapply` (Solymos, 2023), `mvnfast` (Fasiolo, 2023), `Pracma` (Borchers, 2022), `gamlss.dist` (Stasinopoulos, 2022), `gamlss` (Stasinopoulos, 2023), `Matrix` (Maechler, 2022), `RcppEigen` (Eddelbuettel, 2022) e `Rcpp` (Eddelbuettel, 2023). Para o número de réplicas usado em cada cenário, o tempo computacional gasto foi cerca de 19 horas rodando em um PC Desktop com AMD Ryzen 5 5600G (6x 3.90 GHz) e 16.0 GB RAM.

Algoritmo 2 Gerar de UL correlacionada

1: **Entradas:**
   $\mathbf{R}_i$ positiva definida
2: **Inicializar:**
   $\mathbf{z}_i^*$ = vetor aleatório de $N_s\{0_{s\times 1}, \mathbf{R}_i(\rho)\}$
3: **retornar** $F^{-1}\{\Phi(\mathbf{z}_i^*); \boldsymbol{\mu}_i\}$

Nas Tabelas 3.1 e 3.3 estão descritos os resultados dos cenários em que os dados são gerados e ajustados sob a mesma estrutura de correlação, notamos que o viés e EQM de $\hat{\beta}_0$ e $\hat{\beta}_1$ decrescem conforme s e n aumentam, com o indicativo de consistência de ambos os estimadores. Entretanto, para $\hat{\rho}$ podemos observar que o estimador de momentos é viesado para o coeficiente de correlação. Tal resultado é esperado uma vez que os dados são gerados de observações UL's correlacionadas, enquanto o estimador de momentos é calculado para observações z_{ij}'s correlacionadas. Devido à relação monotônica entre y_{ij} e u_{ij}'s é esperado uma boa concordância entre o verdadeiro valor da correlação e o estimado. Mas, pequenas diferenças podem ainda aparecer mesmo para grandes amostras como podemos observar para $n = 500$, em que a convergência dos valores para $\hat{\rho}$ eram aproximadamente de -0.09, 0.27 e 0.67, respectivamente, enquanto os valores considerados para correlação dos dados gerados eram de -0.1, 0.3 e 0.7, respectivamente.

As Tabelas 3.2 e 3.4 descrevem os resultados dos estudos de simulação sob especificação incorreta da estrutura de correlação. Podemos notar um comportamento similar para o viés e EQM de $\hat{\beta}_0$ e $\hat{\beta}_1$ assim como observado nas Tabelas 3.1 e 3.3. Entretanto para $\hat{\rho}$, em geral, há indícios de inconsistência, particularmente quando os dados são gerados sob estrutura de correlação AR(1) e são ajustados sob estrutura de correlação simétrica.

			$\hat{\beta}_0$			$\hat{\beta}_1$			$\hat{\rho}$	
ρ	s	$\bar{\hat{\beta}}_0$	Viés	EQM	$\bar{\hat{\beta}}_1$	Viés	EQM	$\bar{\hat{\rho}}$	Viés	EQM
					AR(1) com $n = 500$					
	10	-3.0000	0.0000	0.0001	5.9999	0.0001	0.0002	-0.0849	0.0151	0.0004
-0.1	5	-3.0004	0.0004	0.0002	6.0006	0.0006	0.0005	-0.0849	0.0151	0.0007
	3	-3.0004	0.0004	0.0003	6.0005	0.0005	0.0008	-0.0842	0.0158	0.0012
	10	-3.0004	0.0004	0.0001	6.0003	0.0003	0.0002	0.2697	0.0303	0.0013
0.3	5	-3.0005	0.0005	0.0002	6.0001	0.0001	0.0004	0.2692	0.0308	0.0018
	3	-3.0004	0.0004	0.0004	6.0000	0.0000	0.0007	0.2693	0.0307	0.0025
	10	-3.0008	0.0008	0.0001	6.0004	0.0004	0.0001	0.6691	0.0309	0.0028
0.7	5	-3.0010	0.0010	0.0003	6.0004	0.0004	0.0002	0.6699	0.0301	0.0040
	3	-3.0010	0.0010	0.0004	6.0005	0.0005	0.0004	0.6673	0.0327	0.0051
					AR(1) com $n = 50$					
	10	-3.0005	0.0005	0.0010	6.0001	0.0001	0.0022	-0.0861	0.0139	0.0022
-0.1	5	-3.0015	0.0015	0.0021	6.0004	0.0004	0.0048	-0.0884	0.0116	0.0047
	3	-3.0020	0.0020	0.0036	6.0006	0.0006	0.0083	-0.0879	0.0121	0.0093
	10	-3.0006	0.0006	0.0012	6.0000	0.0000	0.0021	0.2651	0.0349	0.0052
0.3	5	-3.0033	0.0033	0.0025	6.0022	0.0022	0.0044	0.2611	0.0389	0.0100
	3	-3.0033	0.0033	0.0041	6.0007	0.0007	0.0076	0.2553	0.0447	0.0168
	10	-3.0037	0.0037	0.0016	6.0022	0.0022	0.0012	0.6368	0.0632	0.0151
0.7	5	-3.0050	0.0050	0.0027	6.0030	0.0030	0.0024	0.6107	0.0893	0.0232
	3	-3.0053	0.0053	0.0039	6.0031	0.0031	0.0044	0.5891	0.1109	0.0304
					AR(1) com $n = 10$					
	10	-3.0043	0.0043	0.0053	6.0027	0.0027	0.0119	-0.0910	0.0090	0.0099
-0.1	5	-3.0068	0.0068	0.0111	6.0013	0.0013	0.0254	-0.0932	0.0068	0.0210
	3	-3.0106	0.0106	0.0207	6.0026	0.0026	0.0479	-0.1029	0.0029	0.0375
	10	-3.0056	0.0056	0.0064	6.0027	0.0027	0.0111	0.2457	0.0543	0.0184
0.3	5	-3.0110	0.0110	0.0134	6.0020	0.0020	0.0237	0.2101	0.0899	0.0352
	3	-3.0187	0.0187	0.0227	6.0082	0.0082	0.0439	0.1670	0.1330	0.0592
	10	-3.0132	0.0132	0.0082	6.0070	0.0070	0.0061	0.5207	0.1793	0.0541
0.7	5	-3.0172	0.0172	0.0143	6.0076	0.0076	0.0135	0.4545	0.2455	0.0891
	3	-3.0255	0.0255	0.0227	6.0110	0.0110	0.0275	0.4041	0.2959	0.1236

Tabela 3.1: *Estimativas médias $\bar{\hat{\beta}}_0$, $\bar{\hat{\beta}}_1$ e $\bar{\hat{\rho}}$, vieses (em valor absoluto) e erros quadráticos médios (EQM) de $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\rho}$ do estudo de simulação em que os dados são gerados da distribuição multivariada unit-Lindley com estrutura de correlação AR(1) e ajustados sob o modelo UL-GEE com mesma estrutura de correlação.*

			$\hat{\beta}_0$			$\hat{\beta}_1$			$\hat{\rho}$	
ρ	s	$\bar{\hat{\beta}}_0$	Viés	EQM	$\bar{\hat{\beta}}_1$	Viés	EQM	$\bar{\hat{\rho}}$	Viés	EQM
					Simétrica com $n = 500$					
	10	-3.0000	0.0000	0.0001	5.9999	0.0001	0.0002	-0.0158	0.0842	0.0071
-0.1	5	-3.0004	0.0004	0.0002	6.0006	0.0006	0.0005	-0.0316	0.0684	0.0048
	3	-3.0005	0.0005	0.0003	6.0006	0.0006	0.0008	-0.0534	0.0466	0.0027
	10	-3.0004	0.0004	0.0001	6.0002	0.0002	0.0002	0.0726	0.2274	0.0518
0.3	5	-3.0005	0.0005	0.0002	6.0001	0.0001	0.0005	0.1364	0.1636	0.0273
	3	-3.0004	0.0004	0.0004	6.0000	0.0000	0.0007	0.2051	0.0949	0.0103
	10	-3.0005	0.0005	0.0002	6.0005	0.0005	0.0002	0.3238	0.3762	0.1424
0.7	5	-3.0005	0.0005	0.0003	6.0001	0.0001	0.0003	0.4868	0.2132	0.0475
	3	-3.0008	0.0008	0.0004	6.0004	0.0004	0.0004	0.5948	0.1052	0.0144
					Simétrica com $n = 50$					
	10	-3.0005	0.0005	0.0010	6.0001	0.0001	0.0023	-0.0173	0.0827	0.0072
-0.1	5	-3.0015	0.0015	0.0021	6.0002	0.0002	0.0049	-0.0345	0.0655	0.0058
	3	-3.0022	0.0022	0.0036	6.0008	0.0008	0.0083	-0.0583	0.0417	0.0073
	10	-3.0004	0.0004	0.0013	5.9998	0.0002	0.0023	0.0686	0.2314	0.0547
0.3	5	-3.0031	0.0031	0.0026	6.0022	0.0022	0.0046	0.1295	0.1705	0.0337
	3	-3.0030	0.0030	0.0041	6.0005	0.0005	0.0078	0.1922	0.1078	0.0230
	10	-3.0050	0.0050	0.0019	6.0027	0.0027	0.0019	0.3032	0.3968	0.1639
0.7	5	-3.0059	0.0059	0.0030	6.0033	0.0033	0.0032	0.4391	0.2609	0.0803
	3	-3.0061	0.0061	0.0041	6.0039	0.0039	0.0049	0.5230	0.1770	0.0481
					Simétrica com $n = 10$					
	10	-3.0042	0.0042	0.0054	6.0027	0.0027	0.0124	-0.0232	0.0768	0.0073
-0.1	5	-3.0068	0.0068	0.0113	6.0010	0.0010	0.0260	-0.0450	0.0550	0.0095
	3	-3.0110	0.0110	0.0208	6.0030	0.0030	0.0486	-0.0775	0.0225	0.0230
	10	-3.0061	0.0061	0.0066	6.0034	0.0034	0.0120	0.0544	0.2456	0.0655
0.3	5	-3.0108	0.0108	0.0138	6.0014	0.0014	0.0252	0.0881	0.2119	0.0599
	3	-3.0185	0.0185	0.0231	6.0080	0.0080	0.0454	0.1139	0.1861	0.0680
	10	-3.0183	0.0183	0.0100	6.0098	0.0098	0.0095	0.2255	0.4745	0.2405
0.7	5	-3.0197	0.0197	0.0156	6.0082	0.0082	0.0167	0.3083	0.3917	0.1797
	3	-3.0256	0.0256	0.0236	6.0109	0.0109	0.0293	0.3472	0.3528	0.1584

Tabela 3.2: *Estimativas médias $\bar{\hat{\beta}}_0$, $\bar{\hat{\beta}}_1$ e $\bar{\hat{\rho}}$, vieses (em valor absoluto) e erros quadráticos médios (EQM) de $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\rho}$ do estudo de simulação em que os dados são gerados da distribuição multivariada unit-Lindley com estrutura de correlação AR(1) e ajustados sob o modelo UL-GEE estrutura de correlação simétrica.*

			$\hat{\beta}_0$			$\hat{\beta}_1$			$\hat{\rho}$	
ρ	s	$\bar{\hat{\beta}}_0$	Viés	EQM	$\bar{\hat{\beta}}_1$	Viés	EQM	$\bar{\hat{\rho}}$	Viés	EQM
					Simétrica com $n = 500$					
	10	-3.0000	0.0000	0.0001	6.0000	0.0000	0.0002	-0.0846	0.0154	0.0002
-0.1	5	-3.0001	0.0001	0.0002	6.0000	0.0000	0.0004	-0.0850	0.0150	0.0003
	3	-3.0002	0.0002	0.0003	5.9999	0.0001	0.0008	-0.0851	0.0149	0.0007
	10	-3.0006	0.0006	0.0002	6.0004	0.0004	0.0002	0.2692	0.0308	0.0016
0.3	5	-3.0002	0.0002	0.0003	5.9999	0.0001	0.0004	0.2691	0.0309	0.0019
	3	-3.0002	0.0002	0.0004	6.0000	0.0000	0.0007	0.2682	0.0318	0.0025
	10	-3.0010	0.0010	0.0002	6.0003	0.0003	0.0001	0.6695	0.0305	0.0048
0.7	5	-3.0012	0.0012	0.0003	6.0004	0.0004	0.0002	0.6662	0.0338	0.0049
	3	-3.0009	0.0009	0.0004	6.0004	0.0004	0.0004	0.6656	0.0344	0.0051
					Simétrica com $n = 50$					
	10	-3.0007	0.0007	0.0007	6.0004	0.0004	0.0020	-0.0847	0.0153	0.0003
-0.1	5	-3.0009	0.0009	0.0019	6.0002	0.0002	0.0046	-0.0859	0.0141	0.0012
	3	-3.0017	0.0017	0.0034	5.9999	0.0001	0.0078	-0.0880	0.0120	0.0051
	10	-3.0015	0.0015	0.0018	6.0004	0.0004	0.0020	0.2635	0.0365	0.0073
0.3	5	-3.0036	0.0036	0.0029	6.0016	0.0016	0.0042	0.2573	0.0427	0.0104
	3	-3.0041	0.0041	0.0042	6.0016	0.0016	0.0075	0.2574	0.0426	0.0158
	10	-3.0066	0.0066	0.0022	6.0034	0.0034	0.0012	0.6006	0.0994	0.0261
0.7	5	-3.0053	0.0053	0.0029	6.0027	0.0027	0.0022	0.5890	0.1110	0.0284
	3	-3.0062	0.0062	0.0040	6.0030	0.0030	0.0042	0.5815	0.1185	0.0311
					Simétrica com $n = 10$					
	10	-3.0031	0.0031	0.0039	5.9991	0.0009	0.0110	-0.0817	0.0183	0.0006
-0.1	5	-3.0061	0.0061	0.0107	6.0013	0.0013	0.0259	-0.0925	0.0075	0.0043
	3	-3.0130	0.0130	0.0204	6.0039	0.0039	0.0489	-0.1017	0.0017	0.0216
	10	-3.0086	0.0086	0.0094	6.0038	0.0038	0.0103	0.2153	0.0847	0.0242
0.3	5	-3.0106	0.0106	0.0151	6.0022	0.0022	0.0231	0.1897	0.1103	0.0342
	3	-3.0161	0.0161	0.0234	6.0030	0.0030	0.0432	0.1651	0.1349	0.0524
	10	-3.0242	0.0242	0.0124	6.0119	0.0119	0.0061	0.4230	0.2770	0.1057
0.7	5	-3.0263	0.0263	0.0173	6.0129	0.0129	0.0128	0.4147	0.2853	0.1121
	3	-3.0272	0.0272	0.0229	6.0120	0.0120	0.0246	0.3897	0.3103	0.1309

Tabela 3.3: *Estimativas médias $\bar{\hat{\beta}}_0$, $\bar{\hat{\beta}}_1$ e $\bar{\hat{\rho}}$, vieses (em valor absoluto) e erros quadráticos médios (EQM) de $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\rho}$ do estudo de simulação em que os dados são gerados da distribuição multivariada unit-Lindley com estrutura de correlação simétrica e ajustados sob o modelo UL-GEE com mesma estrutura de correlação.*

			$\hat{\beta}_0$			$\hat{\beta}_1$			$\hat{\rho}$	
ρ	s	$\bar{\hat{\beta}}_0$	Viés	EQM	$\bar{\hat{\beta}}_1$	Viés	EQM	$\bar{\hat{\rho}}$	Viés	EQM
					AR(1) com $n = 500$					
	10	-3.0001	0.0001	0.0001	6.0001	0.0001	0.0002	-0.0848	0.0152	0.0004
-0.1	5	-3.0000	0.0000	0.0002	5.9999	0.0001	0.0005	-0.0853	0.0147	0.0006
	3	-3.0002	0.0002	0.0004	5.9999	0.0001	0.0008	-0.0849	0.0151	0.0010
	10	-3.0006	0.0006	0.0002	6.0004	0.0004	0.0002	0.2694	0.0306	0.0017
0.3	5	-3.0001	0.0001	0.0003	5.9998	0.0002	0.0004	0.2689	0.0311	0.0022
	3	-3.0003	0.0003	0.0004	6.0000	0.0000	0.0007	0.2682	0.0318	0.0028
	10	-3.0008	0.0008	0.0002	6.0003	0.0003	0.0001	0.6684	0.0316	0.0045
0.7	5	-3.0011	0.0011	0.0003	6.0003	0.0003	0.0003	0.6663	0.0337	0.0050
	3	-3.0009	0.0009	0.0004	6.0004	0.0004	0.0005	0.6656	0.0344	0.0055
					AR(1) com $n = 50$					
	10	-3.0010	0.0010	0.0008	6.0009	0.0009	0.0023	-0.0851	0.0149	0.0021
-0.1	5	-3.0007	0.0007	0.0019	5.9998	0.0002	0.0047	-0.0860	0.0140	0.0041
	3	-3.0017	0.0017	0.0034	6.0001	0.0001	0.0079	-0.0874	0.0126	0.0083
	10	-3.0006	0.0006	0.0019	6.0000	0.0000	0.0022	0.2613	0.0387	0.0084
0.3	5	-3.0034	0.0034	0.0030	6.0014	0.0014	0.0046	0.2573	0.0427	0.0131
	3	-3.0042	0.0042	0.0043	6.0017	0.0017	0.0078	0.2587	0.0413	0.0192
	10	-3.0078	0.0078	0.0026	6.0034	0.0034	0.0015	0.6021	0.0979	0.0255
0.7	5	-3.0055	0.0055	0.0032	6.0026	0.0026	0.0026	0.5899	0.1101	0.0290
	3	-3.0059	0.0059	0.0042	6.0028	0.0028	0.0047	0.5807	0.1193	0.0327
					AR(1) com $n = 10$					
	10	-3.0053	0.0053	0.0042	6.0024	0.0024	0.0123	-0.0824	0.0176	0.0089
-0.1	5	-3.0069	0.0069	0.0106	6.0022	0.0022	0.0254	-0.0921	0.0079	0.0185
	3	-3.0132	0.0132	0.0203	6.0041	0.0041	0.0484	-0.1018	0.0018	0.0355
	10	-3.0082	0.0082	0.0099	6.0034	0.0034	0.0114	0.2130	0.0870	0.0296
0.3	5	-3.0106	0.0106	0.0155	6.0026	0.0026	0.0245	0.1875	0.1125	0.0436
	3	-3.0162	0.0162	0.0239	6.0032	0.0032	0.0447	0.1630	0.1370	0.0639
	10	-3.0281	0.0281	0.0144	6.0134	0.0134	0.0077	0.4247	0.2753	0.1066
0.7	5	-3.0261	0.0261	0.0185	6.0118	0.0118	0.0151	0.4115	0.2885	0.1164
	3	-3.0271	0.0271	0.0238	6.0118	0.0118	0.0269	0.3866	0.3134	0.1363

Tabela 3.4: *Estimativas médias $\bar{\hat{\beta}}_0$, $\bar{\hat{\beta}}_1$ e $\bar{\hat{\rho}}$, vieses (em valor absoluto) e erros quadráticos médios (EQM) de $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\rho}$ do estudo de simulação em que os dados são gerados da distribuição multivariada unit-Lindley com estrutura de correlação simétrica e ajustados sob o modelo UL-GEE com estrutura de correlação AR(1).*

3.2 Aplicação

Como ilustração dos modelos UL-GEE propostos no texto, analisamos um conjunto de dados do censo demográfico do Brasil de 1991, 2000 e 2010 para as 27 unidades federativas (veja, os códigos, no Apêndice B.2). Os dados foram extraídos do Atlas do Desenvolvimento Humano no Brasil, disponível em http://www.atlasbrasil.org.br/consulta. Particularmente, a relação entre a proporção de pessoas em domicílios com abastecimento de água e esgoto inadequados (`psewage`) e o coeficiente de Gini (`Gini`). Ou seja, queremos avaliar como a desigualdade social afeta a evolução do abastecimento de água e esgoto, porém levando em consideração a característica temporal dos dados. O conjunto de dados é exibido na Tabela 3.5. A Figura 3.1 apresenta os *boxplots*, para cada ano (`tempo`), do logit(`psewage`), seus gráficos de dispersões (com tendência) contra `Gini` e os histogramas empíricos marginais de `psewage` contra os teóricos da distribuição UL. Notamos pelos gráficos apresentados uma boa concordância marginal da distribuição UL com os dados, e que para cada ano logit(`psewage`) tem relação crescente com o coeficiente de Gini, com indícios de interação entre `tempo` e `Gini`.

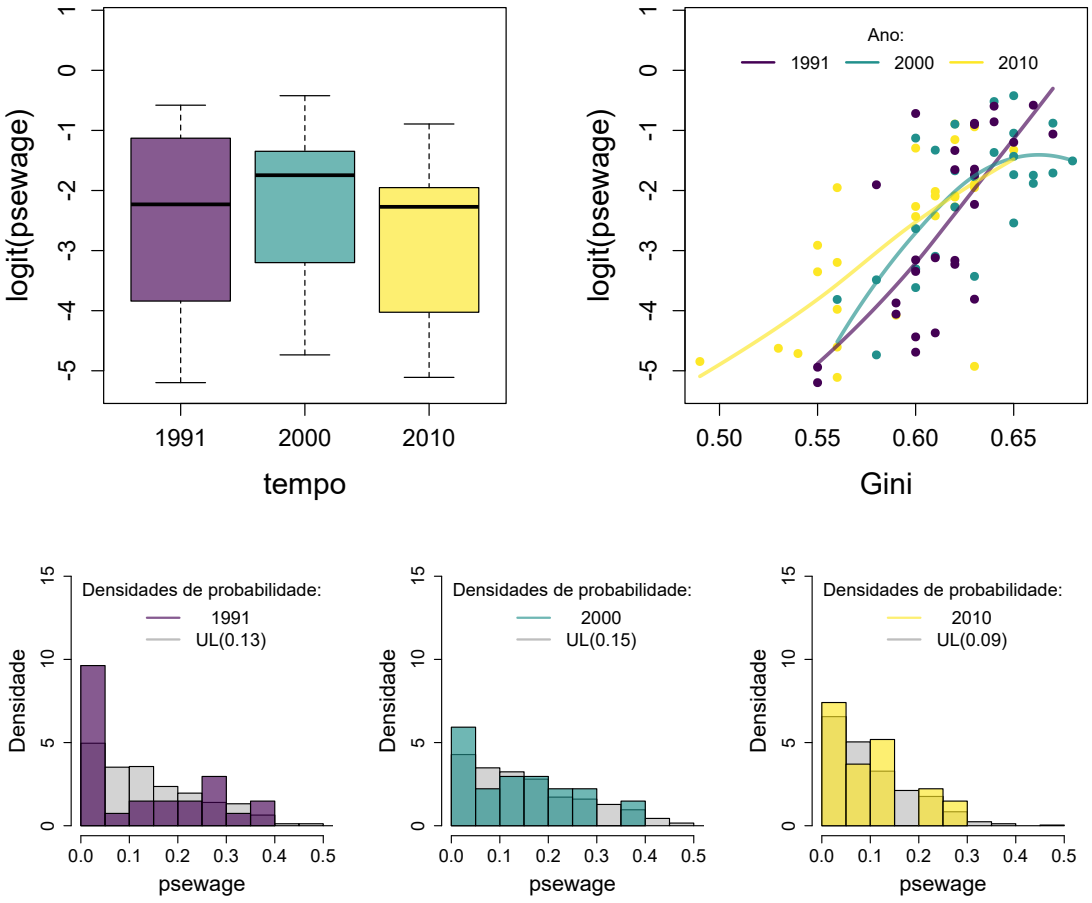

Figura 3.1: *O painel esquerdo superior apresenta os boxplots do logit(`psewage`), enquanto o painel direito superior exibe os gráficos de dispersão entre logit(`psewage`) e `Gini` com suavizações para tendências e o painel inferior mostra os histogramas empíricos marginais de `psewage` e os teóricos da distribuição UL para cada ano avaliados nas 27 unidades federativas do Brasil.*

	Censo					
	1991		2000		2010	
UF	Gini	psewage	Gini	psewage	Gini	psewage
AC	0.63	0.1491	0.64	0.3730	0.63	0.2809
AL	0.63	0.2929	0.68	0.1813	0.63	0.1307
AP	0.58	0.1296	0.62	0.2898	0.60	0.2151
AM	0.62	0.2085	0.67	0.2933	0.65	0.2098
BA	0.67	0.2571	0.66	0.1322	0.62	0.0935
CE	0.66	0.3589	0.67	0.1533	0.61	0.1099
DF	0.62	0.0406	0.63	0.0314	0.63	0.0072
ES	0.60	0.0409	0.60	0.0354	0.56	0.0099
GO	0.59	0.0204	0.60	0.0670	0.55	0.0338
MA	0.60	0.3278	0.65	0.2600	0.62	0.2399
MT	0.60	0.0117	0.62	0.0932	0.56	0.0393
MS	0.60	0.0340	0.62	0.1582	0.55	0.0516
MG	0.61	0.0423	0.61	0.0433	0.56	0.0184
PA	0.64	0.2980	0.63	0.1453	0.61	0.1175
PB	0.60	0.0091	0.60	0.0354	0.53	0.0097
PR	0.62	0.1607	0.65	0.3960	0.62	0.2905
PE	0.65	0.2319	0.66	0.1486	0.62	0.1083
PI	0.64	0.3551	0.65	0.0731	0.61	0.0815
RJ	0.61	0.0125	0.60	0.0262	0.59	0.0167
RN	0.63	0.2888	0.64	0.2031	0.60	0.0940
RS	0.59	0.0170	0.58	0.0297	0.54	0.0089
RO	0.62	0.0381	0.60	0.2446	0.56	0.1243
RR	0.63	0.0217	0.61	0.2095	0.63	0.1244
SC	0.55	0.0071	0.56	0.0216	0.49	0.0078
SP	0.55	0.0055	0.58	0.0087	0.56	0.0060
SE	0.63	0.1619	0.65	0.1498	0.62	0.1102
TO	0.63	0.0970	0.65	0.1929	0.60	0.0807

Tabela 3.5: *Proporção de pessoas em domicílios com abastecimento inadequado de água e esgoto (psewage) e coeficiente de Gini (Gini) para as 27 unidades federativas do Brasil dos censos de 1991, 2000 e 2010.*

Baseados na análise descritiva, propomos o seguinte modelo UL-GEE:

1. $\texttt{psewage}_{ij} | \texttt{Gini}_{ij} \sim \text{UL}(\mu_{ij})$
2. $$\log\left\{\frac{\mu_{ij}}{1-\mu_{ij}}\right\} = \begin{cases} \alpha_1 + \beta_1 \texttt{Gini}_{ij} & \text{para } \texttt{1991} \\ \alpha_2 + \beta_2 \texttt{Gini}_{ij} + \tau_2 \texttt{Gini}_{ij}^2 & \text{para } \texttt{2000} \\ \alpha_3 + \beta_3 \texttt{Gini}_{ij} & \text{para } \texttt{2010} \end{cases}$$
3. $\mathbf{R}_i = \mathbf{R}_i(\rho)$,

em que μ_{ij} denota a proporção esperada de pessoas em domicílios com abastecimento

inadequado de água e esgoto da i-ésima unidade da federação no j-ésimo ano, para $i = 1, \ldots, 27$ e $j = 1, 2, 3$, enquanto $\mathbf{R}_i(\rho)$ é a estrutura de correlação. Baseados na análise de diagnóstico e pela simplicidade escolhemos a estrutura de correlação simétrica com o segundo menor valor de QIC de 5928.91. As estimativas dos coeficientes e seus respectivos erros padrões aproximados do modelo selecionado são apresentados na Tabela 3.6. O menor valor foi obtido para a matriz de correlação não estruturada, QIC = 4877.56, mas essa estrutura custa três estimativas no componente de correlação, enquanto a simétrica custa apenas um.

Coeficiente	Estimativa	Erro padrão	Valor-z	Valor-P
α_1	-18.41	3.59	-5.13	< 0.0001
α_2	-103.98	25.09	-4.14	< 0.0001
α_3	-11.69	2.00	-5.86	< 0.0001
β_1	26.50	5.61	4.72	< 0.0001
β_2	316.01	79.63	3.97	0.0001
β_3	15.57	3.30	4.71	< 0.0001
τ_2	-243.83	63.14	-3.86	0.0001
ρ	0.45			

Tabela 3.6: *Estimativas dos coeficientes e seus erros padrões aproximados do modelo UL-GEE com estrutura de correlação simétrica ajustado para explicar a proporção de domicílios com abastecimento inadequado de água e esgoto nas 27 unidades federativas do Brasil nos anos de 1991, 2000 e 2010 dado o coeficiente de Gini.*

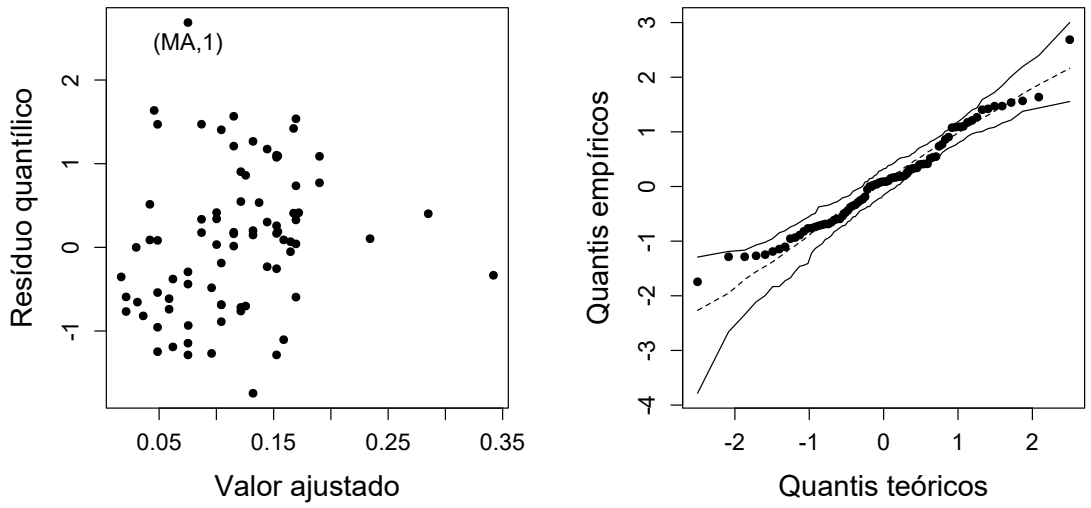

Figura 3.2: *Gráfico de dispersão entre os resíduos quantílicos e os valores ajustados (painel esquerdo) e gráfico normal de probabilidade do resíduo quantílico com uma banda de confiança empírica de 99% (painel direito) do modelo UL-GEE com estrutura de correlação simétrica ajustado para explicar a proporção de domicílios com abastecimento inadequado de água e esgoto nas 27 unidades federativas do Brasil nos anos de 1991, 2000 e 2010 dado o coeficiente de Gini.*

A Figura 3.2 (esquerda) exibe o gráfico de dispersão entre os resíduos quantílicos e

os valores ajustados com a observação (MA,1) (unidade da federação MA no ano 1991) destacada como uma possível observação aberrante, enquanto na Figura 3.2 (direita) temos o gráfico normal de probabilidade do resíduo quantílico com uma banda de confiança empírica de 99%. Ambos os gráficos indicam que o modelo proposto não é inadequado. Em adição, a Figura 3.3 apresenta os *worm plots* de $m = 12$ subconjuntos de resíduos quantílicos independentes, confirmando a adequabilidade do modelo proposto.

Na Figura 3.4 temos a análise de sensibilidade baseada nos gráficos dos índices da curvatura conformal local B_{ij}, sob os esquemas de perturbação de ponderação de casos e da resposta adotando $c^* = 4$. Duas observações, denominadas (AP,1) (unidade da federação AP no ano 1991) e (SC,2) (unidade da federação SC no ano 2000), são destacadas. Na Figura 3.5 é exibido o gráfico de dispersão entre logit(`psewage`) e `Gini` para cada ano com a identificação das observações destacadas, notamos que as três observações destacadas pelos gráficos de diagnóstico são extremas em relação às observações no mesmo ano.

Para confirmar o impacto das três observações destacadas comparamos os valores de MRC e do coeficiente de correlação estimado pela remoção de tais observações com os respectivos valores de 15 subconjuntos aleatórios de 3 observações cada, tomadas do grupo de observações não destacadas. Note que nesses ajustes as unidades federativas deixam de ter o mesmo número de observações. Como podemos ver na Tabela 3.7 o valor de MRC das observações do grupo destacado é 4 vezes o maior valor de MRC do grupo não destacado e a estimativa do coeficiente de correlação parece ser inflada no grupo das observações destacadas. Remover as três observações destacadas não muda a significância dos testes marginais.

Amostra de observações	MRC	$\hat{\rho}$
1	0.01	0.42
2	0.12	0.55
3	0.19	0.38
4	0.10	0.42
5	0.05	0.42
6	0.05	0.43
7	0.08	0.44
8	0.04	0.45
9	0.08	0.44
10	0.09	0.38
11	0.01	0.51
12	0.13	0.41
13	0.06	0.41
14	0.02	0.49
15	0.04	0.43
Destacadas	0.76	0.27

Tabela 3.7: *Comparação dos valores de MRC e estimativas do coeficiente de correlação entre as observações destacadas e* 15 *subconjuntos de amostras aleatórias de tamanho* 3 *tomadas do grupo de observações não destacadas*

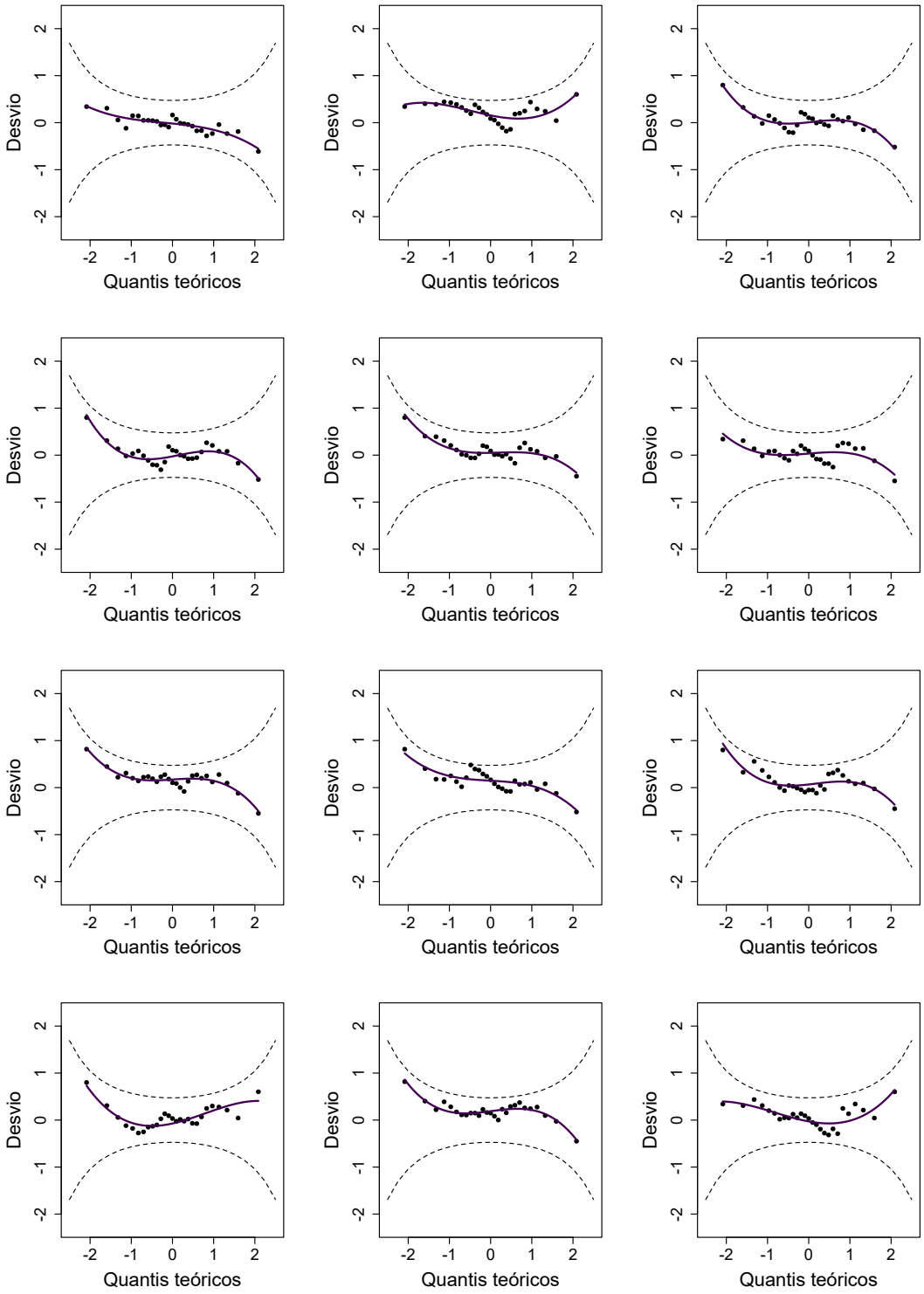

Figura 3.3: *Worm plots de $m = 12$ subconjuntos de resíduos quantílicos independentes do modelo UL-GEE com estrutura de correlação simétrica ajustado para explicar a proporção de domicílios com abastecimento inadequado de água e esgoto nas 27 unidades federativas do Brasil nos anos de 1991, 2000 e 2010 dado o coeficiente de Gini.*

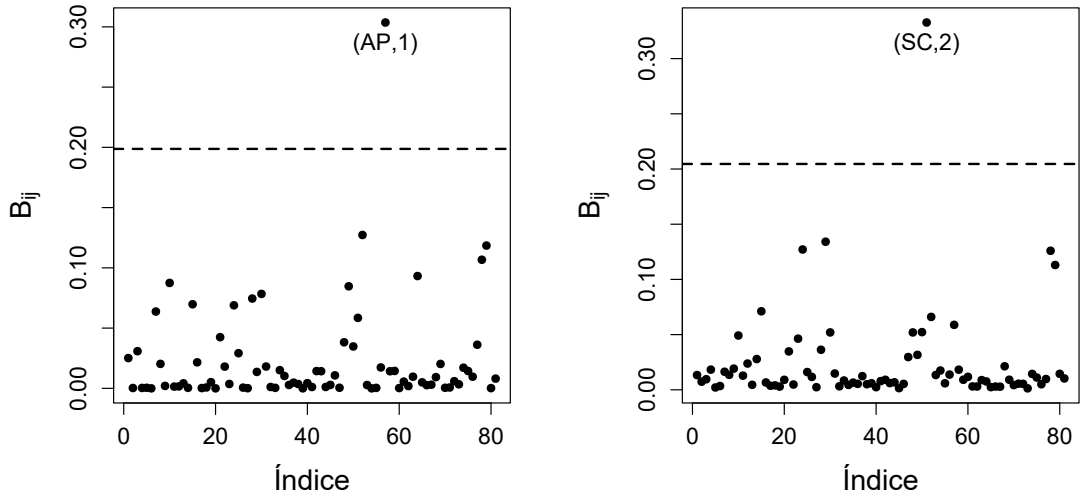

Figura 3.4: *Gráficos dos índices da medida de influência normal conformal B_{ij} sob o esquema de perturbação de ponderação de casos (painel esquerdo) e sob esquema de perturbação da resposta (painel direito) do modelo UL-GEE com estrutura de correlação simétrica ajustado para explicar a proporção de domicílios com abastecimento inadequado de água e esgoto nas 27 federativas do Brasil nos anos de 1991, 2000 e 2010 dado o coeficiente de Gini.*

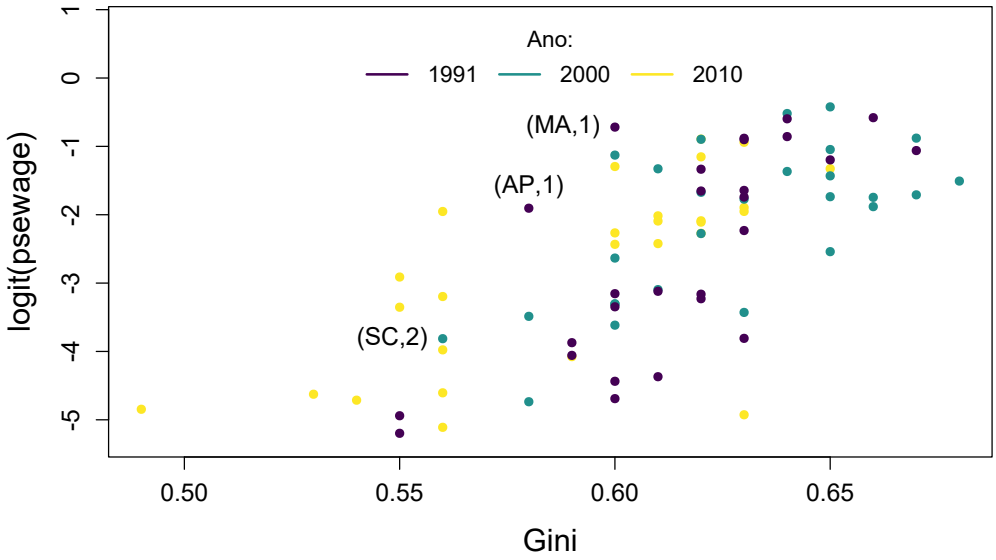

Figura 3.5: *Gráfico de dispersão entre logit(psewage) e `Gini` para cada ano e 27 unidades federativas do Brasil com identificação das observações destacadas nos gráficos de diagnóstico.*

Dado que a inferência é baseada em $n = 27$ unidades experimentais, fizemos um estudo de simulação para verificar a distribuição empírica dos estimadores dos coeficientes do modelo selecionado baseado em simulações Monte Carlo do modelo ajustado. Os gráficos normais de probabilidade são exibidos na Figura 3.6 para as estimativas padronizadas e podemos notar uma boa concordância de todos com a distribuição normal padrão.

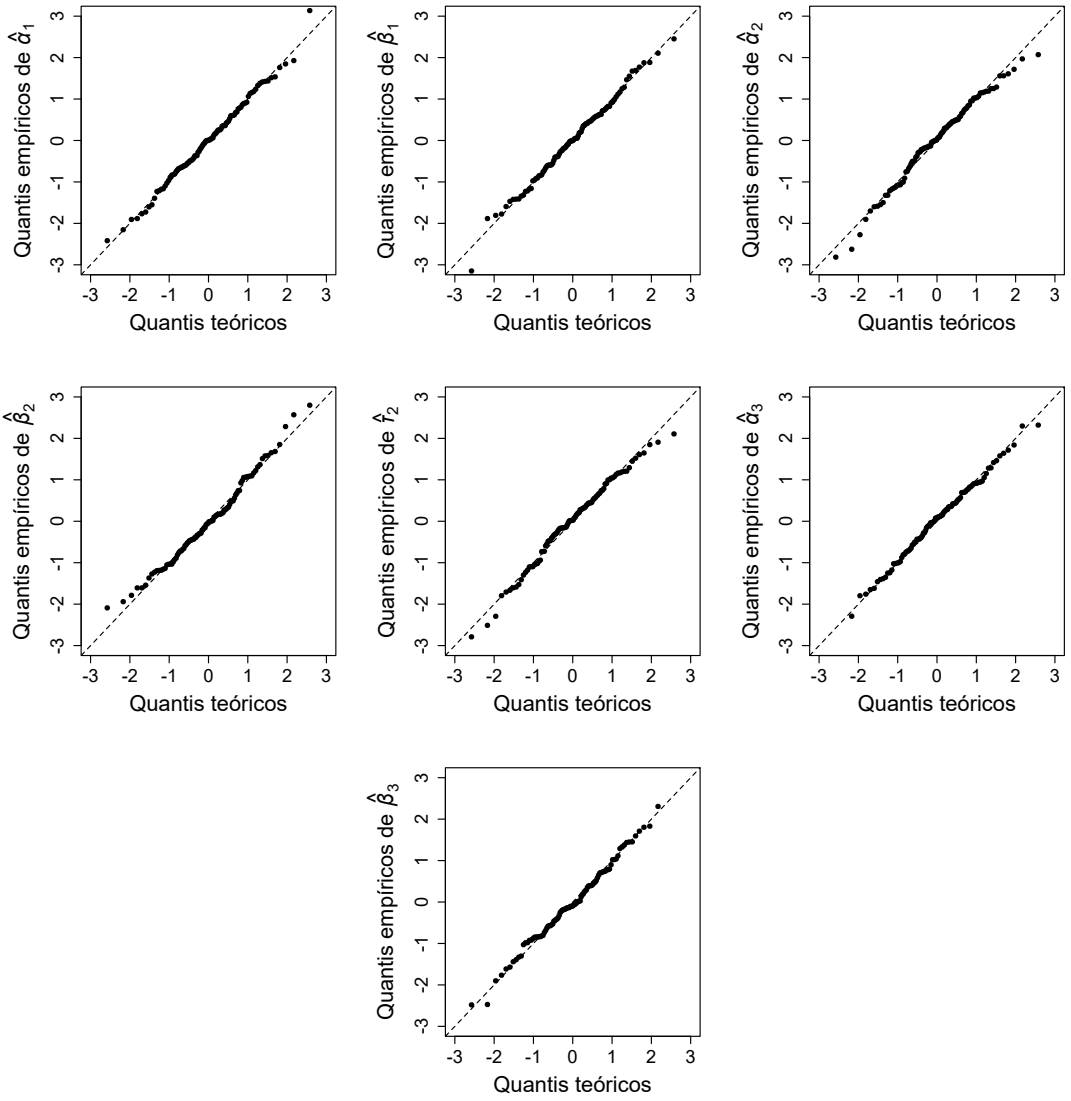

Figura 3.6: *Gráficos normais de probabilidade para distribuição empírica das estimativas padronizadas dos coeficientes do modelo UL-GEE com estrutura de correlação simétrica ajustado para explicar a proporção de domicílios com abastecimento inadequado de água e esgoto nas 27 unidades federativas do Brasil nos anos de 1991, 2000 e 2010 dado o coeficiente de Gini.*

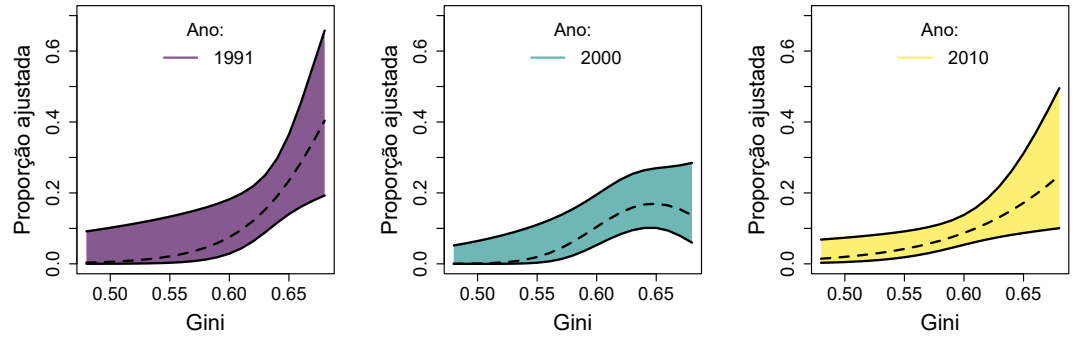

Figura 3.7: *Bandas de confiança de 95% para a média da proporção de domicílios com abastecimento inadequado de água e esgoto dado o coeficiente de Gini para 1991, 2000 e 2010.*

Finalmente, na Figura 3.7 temos as bandas de confiança de 95% para a média da proporção (veja, por exemplo, Piegorch e Casella, 1988) de domicílios com abastecimento inadequado de água e esgoto dado o coeficiente de Gini para 1991, 2000 e 2010. Podemos observar um comportamento semelhante nos anos 1991 e 2010, em que a proporção aumenta conforme o Gini aumenta, mas com grande variabilidade para altos valores de Gini. Contudo, para o ano de 2000 notamos a mesma tendência, porém com uma estabilidade para altos valores de Gini. Este último efeito pode ser devido a uma redução da proporção domicílios com abastecimento inadequado de água e esgoto de 1991 para 2000 em algumas unidades federativas com altos valores de coeficiente de Gini.

Capítulo 4

Modelos UL-GEE2

Para estruturas de correlação mais gerais não é possível obter o estimador de momentos dos coeficientes da estrutura de correlação em forma fechada. Por exemplo, em estudos de medidas espaço-temporais em que as unidades experimentais estão sujeitas ao efeito de mais de uma fonte de correlação. Também consiste em uma maneira alternativa de obter a variância dos estimadores de momentos apresentados no texto.

Assim, derivamos nesta seção equações de estimação para modelagem conjunta do parâmetro de posição e da estrutura de correlação utilizando funções de ligação apropriadas. Um processo iterativo simultâneo é desenvolvido para estimação dos coeficientes da regressão e algumas propriedades inferenciais dos estimadores propostos são apresentadas.

4.1 Funções de estimação

Estendendo a Seção 2.1, seja s_i o conjunto de índices relativos aos instantes em que as respostas são observadas na i-ésima unidade experimental, com cardinalidade denotada por $n(s_i)$. Então $\mathbf{y}_i^\top = \{y_{ij} : j \in s_i\}$ denota um vetor $n(s_i) \times 1$ contendo as respostas (taxas ou proporções), para $i = 1, \ldots, n$. Vamos assumir que $y_{ij} \sim \mathrm{UL}(\mu_{ij})$ com componentes de regressão $g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta}$ e $h(\rho_{ijj'}) = \lambda_{ijj'} = \boldsymbol{\zeta}_{ijj'}^\top \boldsymbol{\gamma}$, em que $g(\cdot)$ e $h(\cdot)$ denotam as funções de ligação diferenciáveis de modo que $g(\cdot)$ tenha domínio no $(0, 1)$ e $h(\cdot)$ tenha domínio no $(-1, 1)$. Note que para cada unidade experimental os parâmetros de correlação modelados $\boldsymbol{\rho}_i = \{\rho_{ijj'} : j, j' \in s_i, j' > j\}^\top$ consistem de um vetor com cardinalidade $n(s_i) \times \{n(s_i) - 1\}/2$, denotada por $m(s_i)$. Também, $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^\top$ e $\boldsymbol{\zeta}_{ijj'} = (\zeta_{ijj'1}, \ldots, \zeta_{ijj'q})^\top$ contêm os valores das variáveis explicativas com $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ e $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^\top$ os vetores com os coeficientes da regressão. Para cada unidade experimental a matriz de correlação de $\mathbf{u}_i = \{u_{ij} : j \in s_i\}^\top$ com $u_{ij} = d\mathrm{L}(\boldsymbol{\mu})/d\mu_{ij}$, para $i = 1, \ldots, n$, pode ser representada por $\mathbf{R}(\mathbf{u}_i)$ de dimensão $n(s_i) \times n(s_i)$. Para $n(s_i) > 1$, os elementos da matriz de correlação podem ser denotados como

$$\mathrm{R}_{kk'}(\mathbf{u}_i) = \frac{\mathrm{E}(u_{ik} u_{ik'})}{\sqrt{\mathrm{Var}(u_{ik})}\sqrt{\mathrm{Var}(u_{ik'})}},$$

para $k \neq k'$, e $\mathrm{R}_{kk'}(\mathbf{u}_i) = 1$, para $k = k'$. $\mathrm{E}(u_{ij}) = 0$ e $\mathrm{E}(u_{ij}^2) = \mathrm{Var}(u_{ij})$, em que

$$\mathrm{Var}(u_{ij}) = \frac{2 - (1 - \mu_{ij})^2}{\mu_{ij}^2 (1 - \mu_{ij})^2},$$

para $i = 1, \ldots, n$. Yan e Fine (2004) propõem a transformação z de Fisher reescalonada como inversa da função de ligação dos parâmetros de correlação, expressa por

$$\rho_{ijj'} = h^{-1}(\lambda_{ijj'}) = \frac{\exp\{\boldsymbol{\zeta}_{ijj'}^\top \boldsymbol{\gamma}\} - 1}{\exp\{\boldsymbol{\zeta}_{ijj'}^\top \boldsymbol{\gamma}\} + 1}.$$

Denotando por $v_{ijj'}$ uma variável aleatória tal que $\mathrm{E}(v_{ijj'}) = 0$ e por $\mathbf{b}_i = (\mathbf{u}_i^\top, \mathbf{v}_i^\top)^\top$ um vetor de dimensão $n(s_i) \times \{n(s_i) + 1\}/2$, com $\mathbf{v}_i = \{v_{ijj'} : j, j' \in s_i, j' > j\}^\top$, considerando o vetor $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$, temos o seguinte conjunto de equações de estimação:

$$\boldsymbol{\Gamma}^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathrm{E} \begin{pmatrix} \partial \mathbf{u}_i / \partial \boldsymbol{\beta}^\top & 0_{n(s_i) \times q} \\ 0_{m(s_i) \times p} & \partial \mathbf{v}_i / \partial \boldsymbol{\gamma}^\top \end{pmatrix}^\top \mathrm{Cov}(\mathbf{b}_i)^{-1} \mathbf{b}_i.$$

Uma proposta para $v_{ijj'}$ pode ser expressa da seguinte forma:

$$v_{ijj'} = \frac{u_{ij} u_{ij'}}{\sqrt{\mathrm{Var}(u_{ij})}\sqrt{\mathrm{Var}(u_{ij'})}} - \rho_{ijj'},$$

em que,

$$\frac{\partial v_{ijj'}}{\partial \gamma_\ell} = -\frac{d\rho_{ijj'}}{d\lambda_{ijj'}} \frac{\partial \lambda_{ijj'}}{\partial \gamma_\ell}.$$

Denotando $d\lambda_{ijj'}/d\rho_{ijj'} = h'(\rho_{ijj'})$ e como $\partial \lambda_{ijj'}/\partial \gamma_\ell = \zeta_{ijj'\ell}$, então obtemos $\mathrm{E}\left(\partial v_{ijj'}/\partial \gamma_\ell\right) = m_{ijj'} \zeta_{ijj'\ell}$, com $m_{ijj'} = -\{h'(\rho_{ijj'})\}^{-1}$, para $i = 1, \ldots, n$ e $\ell = 1, \ldots, q$. Logo, a função de estimação $\boldsymbol{\Gamma}^*(\boldsymbol{\theta})$ pode ser reescrita como

$$\boldsymbol{\Gamma}^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \begin{pmatrix} \mathbf{X}_i^\top \mathbf{D}_i & 0_{p \times m(s_i)} \\ 0_{q \times n(s_i)} & \mathbf{Z}_i^\top \mathbf{M}_i \end{pmatrix} \mathrm{Cov}(\mathbf{b}_i)^{-1} \mathbf{b}_i, \tag{4.1}$$

com $\mathbf{X}_i$ sendo uma matriz $n(s_i) \times p$ com linhas $\mathbf{x}_{ij}^\top = (x_{ij1}, \ldots, x_{ijp})$, $\mathbf{Z}_i$ sendo uma matriz $m(s_i) \times q$ com linhas $\boldsymbol{\zeta}_{ijj'}^\top = (\zeta_{ijj'1}, \ldots, \zeta_{ijj'q})$, $\mathbf{D}_i = \mathrm{diag}\{d_{ij} : j \in s_i\}$, com $d_{ij} = -\mathrm{Var}(u_{ij})\{g'(\mu_{ij})\}^{-1}$, $\mathbf{M}_i = \mathrm{diag}\{m_{ijj'} : j, j' \in s_i, j' > j\}$, com $m_{ijj'} = -\{h'(\rho_{ijj'})\}^{-1}$. Também, $\mathrm{Cov}(\mathbf{b}_i)$ denota a matriz de covariância de $\mathbf{b}_i$ que pode ser expressa da seguinte maneira particionada

$$\mathrm{Cov}(\mathbf{b}_i) = \begin{bmatrix} \mathrm{Cov}(\mathbf{u}_i) & \mathrm{Cov}(\mathbf{u}_i, \mathbf{v}_i) \\ \mathrm{Cov}(\mathbf{v}_i, \mathbf{u}_i) & \mathrm{Cov}(\mathbf{v}_i) \end{bmatrix},$$

com $\mathrm{Cov}(\mathbf{u}_i) = \boldsymbol{\Sigma}_{u_i}^{\frac{1}{2}} \mathbf{R}(\mathbf{u}_i) \boldsymbol{\Sigma}_{u_i}^{\frac{1}{2}}$, em que $\boldsymbol{\Sigma}_{u_i} = \mathrm{diag}\{\mathrm{Var}(u_{ij}) : j \in s_i\}$. Propomos para cada unidade experimental a substituição da matriz de correlação $\mathbf{R}(\mathbf{u}_i)$ por uma matriz de correlação de trabalho não estruturada $\mathbf{R}(\boldsymbol{\rho}_i)$ em que os elementos são expressos apenas em função dos parâmetros de correlação modelados, não envolvendo os coeficientes $\boldsymbol{\beta}_i$ relacionados com os parâmetros de posição. Similar a Yan e Fine (2004), consideramos convenientemente

que $\text{Cov}(\mathbf{u}_i, \mathbf{v}_i) = 0_{n(s_i)\times m(s_i)}$ e $\text{Cov}(\mathbf{v}_i) = \mathbf{I}_{m(s_i)}$, com $\mathbf{I}_{m(s_i)}$ denotando a matriz identidade de ordem $m(s_i)$. Como ilustração, considere o exemplo de uma unidade experimental ℓ em que $n(s_\ell) = 4$ e sem perda de generalidade considere também que $s_\ell = \{1, 2, 3, 4\}$, temos que $\mathbf{b}_\ell = (\mathbf{u}_\ell^\top, \mathbf{v}_\ell^\top)^\top$ com $\mathbf{u}_\ell = (u_{\ell 1}, u_{\ell 2}, u_{\ell 3}, u_{\ell 4})^\top$ e $\mathbf{v}_\ell = (v_{\ell 12}, v_{\ell 13}, v_{\ell 14}, v_{\ell 23}, v_{\ell 24}, v_{\ell 34})^\top$. Então, com $\Sigma_{u_\ell} = \text{diag}\{\text{Var}(u_{\ell 1}), \text{Var}(u_{\ell 1}), \text{Var}(u_{\ell 3}), \text{Var}(u_{\ell 4})\}$ a matriz de covariância de trabalho relativa a $\text{Cov}(\mathbf{b}_\ell)$ é expressa por

$$\left[\begin{array}{cc} \Sigma_{u_\ell}^{\frac{1}{2}} \begin{pmatrix} 1 & \rho_{\ell 12} & \rho_{\ell 13} & \rho_{\ell 14} \\ \rho_{\ell 12} & 1 & \rho_{\ell 23} & \rho_{\ell 24} \\ \rho_{\ell 13} & \rho_{\ell 23} & 1 & \rho_{\ell 34} \\ \rho_{\ell 14} & \rho_{\ell 24} & \rho_{\ell 34} & 1 \end{pmatrix} \Sigma_{u_\ell}^{\frac{1}{2}} & 0_{3\times 6} \\ \hline 0_{6\times 3} & \mathbf{I}_6 \end{array}\right].$$

Desse modo, a função de estimação da expressão (4.1) assume a forma alternativa

$$\boldsymbol{\Gamma}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \begin{pmatrix} \mathbf{X}_i^\top \mathbf{D}_i \boldsymbol{\Omega}_i^{-1} & 0_{p\times m(s_i)} \\ 0_{q\times n(s_i)} & \mathbf{Z}_i^\top \mathbf{M}_i \end{pmatrix} \mathbf{b}_i = \sum_{i=1}^{n} \mathbf{Q}_i^\top \mathbf{W}_i \mathbf{A}_i^{-1} \mathbf{b}_i, \tag{4.2}$$

com $\boldsymbol{\Omega}_i = \Sigma_{u_i}^{\frac{1}{2}} \mathbf{R}(\boldsymbol{\rho}_i) \Sigma_{u_i}^{\frac{1}{2}}$, $\mathbf{Q}_i = \text{blockdiag}\{\mathbf{X}_i, \mathbf{Z}_i\}$, $\mathbf{W}_i = \text{blockdiag}\{\mathbf{D}_i \boldsymbol{\Omega}_i^{-1} \mathbf{D}_i, \mathbf{M}_i^2\}$ e $\mathbf{A}_i = \text{blockdiag}\{\mathbf{D}_i, \mathbf{M}_i\}$, para $i = 1, \ldots, n$. Nas próximas seções derivamos um processo iterativo conjunto para resolver $\boldsymbol{\Gamma}(\hat{\boldsymbol{\theta}}) = 0_{(p+q)\times 1}$ da equação (4.2) e apresentamos algumas propriedades assintóticas dos estimadores obtidos $\hat{\boldsymbol{\theta}}$.

4.2 Processo iterativo

Aplicamos o método escore de Newton para obtermos a estimativa $\hat{\boldsymbol{\theta}}$. Denotando $\boldsymbol{\Gamma}(\boldsymbol{\theta}) = \{\boldsymbol{\Gamma}(\boldsymbol{\beta})^\top, \boldsymbol{\Gamma}(\boldsymbol{\gamma})^\top\}^\top$, temos que

$$\text{E}\{\boldsymbol{\Gamma}'(\boldsymbol{\theta})\} = \sum_{i=1}^{n} \begin{Bmatrix} \mathbf{X}_i^\top \mathbf{D}_i \boldsymbol{\Omega}_i^{-1} \text{E}(\partial \mathbf{u}_i / \partial \boldsymbol{\beta}^\top) \\ \mathbf{Z}_i^\top \mathbf{M}_i \text{E}(\partial \mathbf{v}_i / \partial \boldsymbol{\gamma}^\top) \end{Bmatrix} = \sum_{i=1}^{n} \mathbf{Q}_i^\top \mathbf{W}_i \mathbf{Q}_i,$$

e obtemos o seguinte processo iterativo conjunto:

$$\begin{aligned} \boldsymbol{\theta}^{(m+1)} &= \boldsymbol{\theta}^{(m)} - [\text{E}\{\boldsymbol{\Gamma}'(\boldsymbol{\theta}^{(m)})\}]^{-1} \boldsymbol{\Gamma}(\boldsymbol{\theta}^{(m)}) \\ &= \{\sum_{i=1}^{n} \mathbf{Q}_i^\top \mathbf{W}_i^{(m)} \mathbf{Q}_i\}^{-1} \{\sum_{i=1}^{n} \mathbf{Q}_i^\top \mathbf{W}_i^{(m)} \mathbf{t}_i^{(m)}\}, \ m = 0, 1, 2 \ldots, \end{aligned} \tag{4.3}$$

em que $\mathbf{t}_i = \mathbf{Q}_i \boldsymbol{\theta} - \mathbf{A}_i^{-1} \mathbf{b}_i$ consiste na variável resposta modificada, para $i = 1, \ldots, n$. Note que $\mathbf{W}_i$ desempenha o papel da matriz de pesos e devido ao fato de ser bloco diagonal o processo iterativo descrito é decomposto em duas partes podendo ser implementado de forma que retorne atualizações de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ a cada passo.

Denotando por $\mathbf{X} = [\mathbf{X}_1^\top, \ldots, \mathbf{X}_n^\top]^\top$ uma matriz $N \times p$ de linhas $\mathbf{X}_i$, em que $N = \sum_{i=1}^{n} n(s_i)$, $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_n]^\top$ uma matriz $M \times q$ de linhas $\mathbf{Z}_i^\top$, em que $M = \sum_{i=1}^{n} m(s_i)$, $g(\mathbf{y}) = [g(\mathbf{y}_1^\top), \ldots, g(\mathbf{y}_n^\top)]^\top$ um vetor de dimensão $N \times 1$, com elementos $g(\mathbf{y}_i) = \{g(y_{ij}) : j \in s_i\}^\top$ e

$h(\bar{\boldsymbol{\rho}}) = [h(\bar{\boldsymbol{\rho}}_1), \ldots, h(\bar{\boldsymbol{\rho}}_n)]^\top$ um vetor de dimensão $M \times 1$, em que $\bar{\boldsymbol{\rho}}_i$ é um vetor de dimensão $m(s_i) \times 1$ com todos os elementos iguais à correlação amostral entre as variáveis resposta da i-ésima unidade experimental, para $i = 1, \ldots, n$. Então, propomos o Algoritmo 3 para obter as estimativas dos coeficientes da regressão.

Algoritmo 3 Estimativas dos coeficientes

1: **Entradas:**
   $\mathbf{X}$ e $\mathbf{Z}$ com posto coluna completo
2: **Inicializar:**
   $\boldsymbol{\beta}^{(0)} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top g(\mathbf{y})$ e $\boldsymbol{\gamma}^{(0)} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top h(\bar{\boldsymbol{\rho}})$
3: **repetir**
4: atualize $\boldsymbol{\theta}$ do processo iterativo
5: **até** $\max_i|\theta_i^{(m)} - \theta_i^{(m-1)}|/\theta_i^{(m-1)} < \epsilon$
6: **retornar** $\boldsymbol{\theta}^{(m)}$

4.3 Inferência

Note que $\mathrm{E}(\mathbf{b}_i) = \{\mathrm{E}(\mathbf{u}_i)^\top, \mathrm{E}(\mathbf{v}_i)^\top\}^\top$, em que $\mathrm{E}(\mathbf{u}_i) = \{\mathrm{E}(u_{ij}) : j \in s_i\}^\top = 0_{n(s_i)\times 1}$ e também que $\mathrm{E}(\mathbf{v}_i) = \{\mathrm{E}(v_{ijj'}) : j, j' \in s_i, j' > j\}^\top$, para todo i, com

$$\mathrm{E}(v_{ijj'}) = \frac{\mathrm{E}(u_{ij}u_{ij'})}{\sqrt{\mathrm{Var}(u_{ij})}\sqrt{\mathrm{Var}(u_{ij'})}} - \rho_{ijj'} = \mathrm{R}_{jj'}(\mathbf{u}_i) - \frac{\exp\{\boldsymbol{\zeta}_{ijj'}^\top\boldsymbol{\gamma}\} - 1}{\exp\{\boldsymbol{\zeta}_{ijj'}^\top\boldsymbol{\gamma}\} + 1},$$

de modo que $\mathrm{E}(\mathbf{v}_i) = 0_{m(s_i)\times 1}$, assintoticamente, caso o modelo seja corretamente especificado, assim temos uma função de estimação não viesada, ou seja, $\mathrm{E}\{\boldsymbol{\Gamma}(\boldsymbol{\theta})\} = 0_{(p+q)\times 1}$. De Godambe (1997), a matriz de variabilidade de $\boldsymbol{\Gamma}(\boldsymbol{\theta})$ fica definida como $\mathbf{V}_{n\Psi}(\boldsymbol{\theta}) = \mathrm{E}\{\boldsymbol{\Gamma}(\boldsymbol{\theta})\boldsymbol{\Gamma}^\top(\boldsymbol{\theta})\}$ e a respectiva matriz de sensibilidade fica dada por $\mathbf{S}_{n\Psi}(\boldsymbol{\theta}) = \mathrm{E}\{\boldsymbol{\Gamma}'(\boldsymbol{\theta})\}$. A matriz de informação de Godambe de $\boldsymbol{\theta}$ consiste em uma função de estimação regular definida como

$$\mathbf{J}_{n\Gamma}(\boldsymbol{\theta}) = \mathbf{S}_{n\Gamma}(\boldsymbol{\theta})^\top\mathbf{V}_{n\Gamma}^{-1}(\boldsymbol{\theta})\mathbf{S}_{n\Gamma}(\boldsymbol{\theta}),$$

em que $\mathbf{V}_{n\Gamma}(\boldsymbol{\theta}) = \sum_{i=1}^n \mathbf{V}_i(\boldsymbol{\theta})$, com $\mathbf{V}_i(\boldsymbol{\theta}) = \mathbf{Q}_i^\top\mathbf{W}_i\mathbf{A}_i^{-1}\mathrm{Cov}(\mathbf{b}_i)\mathbf{A}_i^{-1}\mathbf{W}_i\mathbf{Q}_i$, $\mathbf{S}_{n\Gamma}(\boldsymbol{\theta}) = \sum_{i=1}^n \mathbf{S}_i(\boldsymbol{\theta})$ com $\mathbf{S}_i(\boldsymbol{\theta}) = \mathbf{Q}_i^\top\mathbf{W}_i\mathbf{Q}_i$, para $i = 1, \ldots, n$. Similar a Tsuyuguchi et al. (2020), de Artes e Jørgensen (2000), temos que $\hat{\boldsymbol{\theta}}$ obtido do processo iterativo da equação (4.3) é tal que

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} \mathrm{N}_{p+q}\{0_{(p+q)\times 1}, \mathbf{J}_\Gamma^{-1}(\boldsymbol{\theta})\},$$

em que $\mathbf{J}_\Gamma(\boldsymbol{\theta}) = \lim_{n\to\infty} n^{-1}\mathbf{J}_{n\Gamma}(\boldsymbol{\theta})$, com $\mathbf{J}_{n\Gamma}(\boldsymbol{\theta}) = \{\sum_{i=1}^n \mathbf{S}_i(\boldsymbol{\theta})\}\{\sum_{i=1}^n \mathbf{V}_i(\boldsymbol{\theta})\}^{-1}\{\sum_{i=1}^n \mathbf{S}_i(\boldsymbol{\theta})\}$.

Um estimador consistente da matriz de covariância de $\hat{\boldsymbol{\theta}}$ fica dado por

$$\{\widehat{\mathbf{J}}_\Gamma(\boldsymbol{\theta})\}^{-1} = \{\sum_{i=1}^n \widehat{\mathbf{S}}_i(\boldsymbol{\theta})\}^{-1}\{\sum_{i=1}^n \mathbf{Q}_i^\top\widehat{\mathbf{W}}_i\widehat{\mathbf{A}}_i^{-1}\hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^\top\widehat{\mathbf{A}}_i^{-1}\widehat{\mathbf{W}}_i\mathbf{Q}_i\}\{\sum_{i=1}^n \widehat{\mathbf{S}}_i(\boldsymbol{\theta})\}^{-1},$$

com $\widehat{\mathbf{S}}_i(\boldsymbol{\theta}) = \mathbf{Q}_i^\top\widehat{\mathbf{W}}_i\mathbf{Q}_i$, em que $\widehat{\mathbf{W}}_i = \mathrm{blockdiag}\{\widehat{\mathbf{D}}_i\widehat{\boldsymbol{\Omega}}_i^{-1}\widehat{\mathbf{D}}_i, \widehat{\mathbf{M}}_i^2\}$ e $\widehat{\mathbf{A}}_i = \mathrm{blockdiag}\{\widehat{\mathbf{D}}_i, \widehat{\mathbf{M}}_i\}$, para

$i = 1, \ldots, n$.

O teste de hipótese $\mathrm{H}_0 : \mathbf{C}\boldsymbol{\theta} = \mathbf{m}$ contra $\mathrm{H}_1 : \mathbf{C}\boldsymbol{\theta} \neq \mathbf{m}$, em que $\mathbf{C}$ uma matriz $r \times (p+q)$ de posto linha r ($r \leq p+q$) pode ser avaliado com uma estatística do tipo Wald, cuja expressão fica dada por $\xi_W = (\mathbf{C}\hat{\boldsymbol{\theta}} - \mathbf{m})^\top [\mathbf{C}\{\widehat{\mathbf{J}}_\Gamma(\boldsymbol{\theta})\}^{-1}\mathbf{C}^\top]^{-1}(\mathbf{C}\hat{\boldsymbol{\theta}} - \mathbf{m})$. Para amostras grandes e sob condições usuais de regularidade, segue que $\xi_W \sim \chi^2_r$, em que χ^2_r denota a distribuição qui-quadrado com r graus de liberdade.

Capítulo 5

Considerações finais

Neste texto derivamos equações de estimação para analisar taxas e proporções correlacionadas no intervalo $(0, 1)$ assumindo que as distribuições marginais são *unit*-Lindley (Mazucheli et al., 2019). A dependência dentro de cada unidade experimental é estimada usando a mesma ideia de equações de estimação generalizadas propostas por Liang e Zeger (1986), e a inferência é baseada na abordagem de Godambe (1997) de funções de estimação. Vários resultados são derivados no texto, tais como um processo iterativo reponderado para estimação dos coeficientes da regressão conjuntamente com a estrutura de correlação das unidades experimentais, análises de resíduos baseadas nos resíduos quantílicos marginais e estudos de sensibilidade baseados na curvatura normal conformal. Em particular, propomos uma nova forma de obter os resíduos quantílicos de maneira independente para verificar a adequabilidade da distribuição marginal e da estrutura de correlação. Diferente de outros trabalhos em equações de estimação, os resultados derivados e implementados ao longo do texto podem ser aplicados para estudos não balanceados.

Nos estudos de simulação temos indicação de consistência para as estimativas dos coeficientes da regressão em todos os cenários considerados, mesmo sob especificação incorreta da estrutura de correlação. Uma aplicação a um conjunto de dados reais é apresentada em que a proporção de pessoas em domicílios com abastecimento de água e esgoto inadequados é relacionada com o coeficiente de Gini nas unidades da federação brasileiras para um estudo longitudinal baseado no censo demográfico do país nos anos de 1991, 2000 e 2010. Por fim, propomos uma possível extensão natural para estruturas de correlação gerais em que as correlações também são modeladas.

A ideia central do texto foi proposta pela Professora Hatice Tül Kübra Akdur da Universidade de Gazi, Ankara, Turquia. Todos os códigos R para ajuste e diagnóstico de modelos UL-GEE, e para obter as saídas das simulações e aplicações estão disponíveis, respectivamente, nos endereços https://github.com/silva-danilo/ulgee e https://github.com/silva-danilo/ulgee_sup.

5.1 Pesquisas futuras

Uma extensão deste texto é a adição de termos parcialmente aditivos (Ibacache et al., 2013) ou termos *single-index* (Wood, 2017) nos preditores tanto do parâmetro de posição como da estrutura de correlação, de forma a acomodar melhor as relações com as variáveis explicativas.

Quanto aos códigos é necessário investigar uma maneira de simular diretamente da estrutura de correlação desejada, talvez usando medidas de correlação baseada em postos como a correlação de Spearman, tau de Kendall e a recente medida de Chatterjee (2021). Também é possível implementar o processo iterativo de forma mais eficiente usando recursão e submeter o esqueleto do pacote ao CRAN (*The Comprehensive R Archive Network*).

Pode-se calcular predições para modelos UL-GEE e UL-GEE2 usando validação cruzada como feito em `cvGEE` (Rizopoulos, 2019) e atualizar o conjunto de dados com as informações do censo demográfico de 2022. Ainda é possível complementar a análise de sensibilidade utilizando a metodologia *forward search* (Atkinson e Riani, 2000) e algoritmos de seleção objetiva de covariáveis.

5.2 Artigo

Parte deste texto resultou no seguinte artigo Silva DV, Akdur HTK, Paula GA (2023). Analysis of correlated unit-Lindley data based on estimating equations. *Statistical Methods and Applications*. https://doi.org/10.1007/s10260-023-00699-w.

Apêndice A

Aqui apresentamos, para melhor organização, alguns resultados técnicos citados ao longo do texto.

A.1 Distribuição da chance

Seja $z = y/(1-y)$ a razão observada, em que $y \sim \text{UL}(\mu)$, $0 < \mu < 1$. De Mazucheli et al. (2019), a função densidade de probabilidade de z pode ser expressa na forma

$$f(z;\mu) = \frac{(1+z)(1-\mu)^2}{\mu}\exp\left\{-\frac{z(1-\mu)}{\mu}\right\},$$

em que $z > 0$. Podemos reescrever essa função na forma da família exponencial uniparamétrica de distribuições $f(z;\theta) = \exp\{\theta z - b(\theta) + c(z)\}$, com $\theta = 1-\mu^{-1}$, $b(\theta) = \log(1-\theta) - 2\log(-\theta)$ e $c(z) = \log(1+z)$ (veja, por exemplo, McCullagh e Nelder, 1989). Então, segue que $\text{E}(z) = b'(\theta)$ e $\text{Var}(z) = b''(\theta)$, com

$$b'(\theta) = -\left(\frac{2}{\theta} + \frac{1}{1-\theta}\right) = \frac{\mu(1+\mu)}{1-\mu} \quad \text{e} \quad b''(\theta) = \frac{2}{\theta^2} - \frac{1}{(1-\theta)^2} = \frac{\mu^2\{2-(1-\mu)^2\}}{(1-\mu)^2}.$$

Dessa forma

$$\text{E}(z^2) = b''(\theta) - b'(\theta)^2 = \frac{2\mu^2(2\mu+1)}{(1-\mu)^2}.$$

A.2 Condições de regularidade

Suponha que $y \sim \text{UL}(\mu)$, $0 < \mu < 1$. Como mostrado no Capítulo 1 a função escore de μ pode ser expressa como

$$u = \frac{d\text{L}(\mu)}{d\mu} = \frac{z}{\mu^2} - \frac{1+\mu}{\mu(1-\mu)},$$

em que $z = y/(1-y)$. Usando os resultados do Apêndice A.1, obtemos

$$\text{E}(u) = \frac{\text{E}(z)}{\mu^2} - \frac{1+\mu}{\mu(1-\mu)} = 0 \quad \text{e} \quad \text{E}(u^2) = \frac{\text{E}(z^2)}{\mu^4} + \frac{(1+\mu)^2}{\mu^2(1-\mu)^2} - \frac{2\text{E}(z)(1+\mu)}{\mu^3(1-\mu)} = \frac{2-(1-\mu)^2}{\mu^2(1-\mu)^2}.$$

Então,

$$\mathrm{Var}(u) = \mathrm{E}(u^2) = \frac{2-(1-\mu)^2}{\mu^2(1-\mu)^2},$$

e também,

$$\mathrm{E}(u') = \mathrm{E}\left(\frac{d^2\mathrm{L}(\mu)}{d\mu^2}\right) = \frac{-2\mathrm{E}(z)}{\mu^3} - \frac{1}{\mu(1-\mu)} + \frac{(1+\mu)(1-2\mu)}{\mu^2(1-\mu)^2} = -\mathrm{Var}(u).$$

A.3 Função de estimação ótima

Considere a função de estimação ótima de Crowder

$$\boldsymbol{\Psi}^*(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathrm{E}\left(\frac{\partial \mathbf{u}_i}{\partial \boldsymbol{\beta}^\top}\right)^\top \mathrm{Cov}(\mathbf{u}_i)^{-1}\mathbf{u}_i,$$

em que $(\partial \mathbf{u}_i/\partial \boldsymbol{\beta}^\top)$ é uma matriz $n(s_i) \times p$ com o (j,ℓ)-ésimo elemento dado por

$$\frac{\partial u_{ij}}{\partial \beta_\ell} = \frac{du_{ij}}{d\mu_{ij}}\frac{\partial \mu_{ij}}{\partial \beta_\ell} = u'_{ij}\frac{d\mu_{ij}}{d\eta_{ij}}\frac{\partial \eta_{ij}}{\partial \beta_\ell},$$

denotando $d\eta_{ij}/d\mu_{ij} = g'(\mu_{ij})$ e como $\partial\eta_{ij}/\partial\beta_\ell = x_{ij\ell}$, então obtemos

$$\mathrm{E}\left(\frac{\partial u_{ij}}{\partial \beta_\ell}\right) = d_{ij}x_{ij\ell},$$

com $d_{ij} = -\mathrm{Var}(u_{ij})\{g'(\mu_{ij})\}^{-1}$, para $i = 1,\ldots,n$ e $\ell = 1,\ldots,p$. Então, a função de estimação ótima $\boldsymbol{\Psi}^*(\boldsymbol{\beta})$ pode ser reescrita como

$$\boldsymbol{\Psi}^*(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{X}_i^\top \mathbf{D}_i \mathrm{Cov}(\mathbf{u}_i)^{-1}\mathbf{u}_i,$$

em que $\mathbf{X}_i$ é uma matriz $n(s_i) \times p$ com linhas $\mathbf{x}_{ij}^\top = (x_{ij1},\ldots,x_{ijp})$ e $\mathbf{D}_i = \mathrm{diag}\{d_{ij} : j \in s_i\}$, para $i = 1,\ldots,n$.

A.4 Matrizes de sensibilidade e variabilidade

As matrizes de sensibilidade e variabilidade de $\boldsymbol{\Psi}(\boldsymbol{\beta})$ são, respectivamente, dadas por $\mathbf{S}_{n\Psi}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\mathbf{S}_i(\boldsymbol{\beta})$ e $\mathbf{V}_{n\Psi}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\mathbf{V}_i(\boldsymbol{\beta})$, em que

$$\mathbf{S}_i(\boldsymbol{\beta}) = \mathrm{E}\{\boldsymbol{\Psi}'_i(\boldsymbol{\beta})\} = \mathbf{X}_i^\top \mathbf{D}_i \boldsymbol{\Omega}_i^{-1}\mathrm{E}\left(\frac{\partial \mathbf{u}_i}{\partial \boldsymbol{\beta}^\top}\right) = \mathbf{X}_i^\top \mathbf{W}_i \mathbf{X}_i,$$

com $\mathbf{W}_i = \mathbf{D}_i\boldsymbol{\Omega}_i^{-1}\mathbf{D}_i$ e $\boldsymbol{\Omega}_i = \boldsymbol{\Sigma}_{u_i}^{\frac{1}{2}}\mathbf{R}_i(\boldsymbol{\rho})\boldsymbol{\Sigma}_{u_i}^{\frac{1}{2}}$, e

$$\mathbf{V}_i(\boldsymbol{\beta}) = \mathrm{E}\left\{\boldsymbol{\Psi}_i(\boldsymbol{\beta})\boldsymbol{\Psi}_i^\top(\boldsymbol{\beta})\right\} = \mathbf{X}_i^\top \mathbf{W}_i \mathbf{D}_i^{-1}\mathrm{Cov}(\mathbf{u}_i)\mathbf{D}_i^{-1}\mathbf{W}_i\mathbf{X}_i.$$

Apêndice B

Aqui apresentamos alguns códigos em R citados ao longo do texto. Para utilizá-los é necessário instalar o esqueleto do pacote executando no R o seguinte comando: `devtools::install_github("silva-danilo/ulgee")`.

B.1 Estudo de simulação

```
# load ulgee package
library(ulgee)

# prep. simulation
beta <- c(-3, 6)
R <- 5000
rho <- c(-0.1, 0.3, 0.7)
n <- c(500, 50, 10)
s <- c(10, 5, 3)

# AR1
set.seed(4598)
table_1 <- ul_sim(beta, R, rho, n, s, "AR1", "EXC")
print(xtable::xtable(table_1[,-2], digits=4),
      type="latex",
      include.rownames=F, include.colnames=F)

# EXC
set.seed(8878)
table_2 <- ul_sim(beta, R, rho, n, s, "EXC", "AR1")
print(xtable::xtable(table_2[,-2], digits=4),
      type="latex",
      include.rownames=F, include.colnames=F)
```

B.2 Aplicação

```
# load ulgee package
library(ulgee)

# set directory
#setwd("\home\...\application")

# prep. 2010
data_2010 <- readxl::read_excel("data/censo_2010.xlsx")
data_2010 <- data_2010[-c(1,29:31),]
names(data_2010) <- c("state", "gini", "pwater", "life", "peletro",
                      "psewage", "hdi", "income")
data_2010$time <- 3
data_2010$year <- 2010
data_2010$id <- 1:nrow(data_2010)
data_2010$psewage <- data_2010$psewage/100
data_2010$state <- c("AC","AL","AP","AM","BA","CE","DF","ES","GO","MA","MT",
                     "MS","MG","PA","PB","PR","PE","PI","RJ","RN","RS","RO",
                     "RR","SC","SP","SE","TO")

# prep. 2000
data_2000 <- readxl::read_excel("data/censo_2000.xlsx")
data_2000 <- data_2000[-c(1,29:31),]
names(data_2000) <- c("state", "gini", "pwater", "life", "peletro",
                      "psewage", "hdi", "income")
data_2000$time <- 2
data_2000$year <- 2000
data_2000$id <- 1:nrow(data_2000)
data_2000$psewage <- data_2000$psewage/100
data_2000$state <- c("AC","AL","AP","AM","BA","CE","DF","ES","GO","MA","MT",
                     "MS","MG","PA","PB","PR","PE","PI","RJ","RN","RS","RO",
                     "RR","SC","SP","SE","TO")

# prep. 1991
data_1991 <- readxl::read_excel("data/censo_1991.xlsx")
data_1991 <- data_1991[-c(1,29:31),]
names(data_1991) <- c("state", "gini", "pwater", "life", "peletro",
                      "psewage", "hdi", "income")
data_1991$time <- 1
data_1991$year <- 1991
```

```r
data_1991$id <- 1:nrow(data_1991)
data_1991$psewage <- data_1991$psewage/100
data_1991$state <- c("AC","AL","AP","AM","BA","CE","DF","ES","GO","MA","MT",
                     "MS","MG","PA","PB","PR","PE","PI","RJ","RN","RS","RO",
                     "RR","SC","SP","SE","TO")

# prep. data
data <- rbind(data_2010, data_2000, data_1991)

# prep. model
y <- data$psewage; time <- data$time; id <- data$id; state <- data$state
X <- cbind(as.numeric(time==1), as.numeric(time==1)*data$gini,
           as.numeric(time==2), as.numeric(time==2)*data$gini,
           as.numeric(time==2)*data$gini^2,
           as.numeric(time==3), as.numeric(time==3)*data$gini)

# model
fit_1 <- ulgee(y, X, time, id, "EXC", "logit", 1e-6, 40)

# estimation
round(fit_1$mu.coefs, 2)
round(diag(fit_1$vcov), 2)
round(fit_1$pvalues, 4)
round(fit_1$rho, 2)
round(fit_1$qic, 2)

# qic comparison
table <- matrix(NA, 4, 1)
corr_type <- c("EXC", "AR1", "UNS", "IDE")
rownames(table) <- corr_type
colnames(table) <- "qic"
for(i in 1:length(corr_type)){
  fit_i <- ulgee(y, X, time, id, corr_type[i], "logit", 1e-6, 50, F)
  table[i,1] <- fit_i$qic
}
round(table,2)

# diagnostic
set.seed(3489)
diag_quant(fit_1, X, 100, 1, label.id=data$state, label.time=data$time)

```

```r
# sensitivity
sens_conf(fit_1, 4, 1, 4, 1, label.id=data$state, label.time=data$time)

# prep position
pos <- as.numeric(data$id==3 & data$time==1)
pos <- pos + as.numeric(data$id==24 & data$time==2)
pos <- pos + as.numeric(data$id==10 & data$time==1)
pos <- pos==1

# sensitivity dropp
set.seed(9792)
table <- sens_mrc(fit_1, y, X, 10 , pos)
print(xtable::xtable(table, digits=2), type="latex",
      include.rownames=F, include.colnames=F)

# asymptotic check
set.seed(1111)
table <- sens_coef(fit_1, X, 100)
titles <- c(expression(paste("Quantis empíricos de  ",
            hat(italic("\u03b1"))[1])),
            expression(paste("Quantis empíricos de  ",
            hat(italic("\u03b2"))[1])),
            expression(paste("Quantis empíricos de  ",
            hat(italic("\u03b1"))[2])),
            expression(paste("Quantis empíricos de  ",
            hat(italic("\u03b2"))[2])),
            expression(paste("Quantis empíricos de  ",
            hat(italic("\u03c4"))[2])),
            expression(paste("Quantis empíricos de  ",
            hat(italic("\u03b1"))[3])),
            expression(paste("Quantis empíricos de  ",
            hat(italic("\u03b2"))[3])))
par(mar=c(5.5,5.5,2,2), mfrow=c(1,3))
for(i in 1:7){
  if(i==7) plot(1, type="n", xaxt="n", yaxt="n", xlab="", ylab="", bty="n",
                xlim=c(-3,3), ylim=c(-3,3))
  pcoef_i <- (table[,i]-mean(table[,i]))/sd(table[,i])
  qqnorm(pcoef_i, xlab="Quantis teóricos", ylab=titles[i], pch=16,
         cex.lab=1.8, cex.axis=1.5, main="", xlim=c(-3,3), ylim=c(-3,3))
  abline(a=0, b=1, xlab="", ylab="", lty=2, lwd=1)
}
```

Referências

Abd-Elrahman AM (2013). Utilizing ordered statistics in lifetime distributions production: a new lifetime distribution and applications. *Journal of Probability and Statistical Science 11*, 153–164.

Adler A (2022). lamW: Lambert-W Function. R package version 2.1.1. https://cran.r-project.org/package=lamW.

Akdur HTK (2021). Unit-lindley mixed-effect model for proportion data. *Journal of Applied Statistics 48*, 2389–2405.

Altun E, El-Morshedy M, Eliwa, MS (2021). A new regression model for bounded response variable: An alternative to the beta and unit-lindley regression models. *Plos One 16*, 1–15.

Artes R, Jørgensen B (2000). Longitudinal data estimating equations for dispersion model. *Scandinavian Journal of Statistics 27*, 321–334.

Atkinson A, Riani M (2000). *Robust Diagnostic Regression Analysis*. New York: Springer.

Barndorff-Nielsen OE, Jørgensen B (1991). Some parametric models on the simplex. *Journal of Multivariate Analysis 39*, 109–116.

Borchers HW (2022). pracma: Practical Numerical Math Functions. R package version 2.4.2. https://cran.r-project.org/package=pracma.

Buuren V, Fredriks M (2001). Worm plot: simple diagnostic device for modelling growth reference curves. *Statistics in Medicine 20*, 1259–1277.

Cadigan NG, Farrell PJ (2002). Generalized local influence with applications to fish stock cohort analysis. *Journal of Applied Statistics 51*, 469–483.

Chatterjee S (2021). A new coefficient of correlation. *Journal of the American Statistical Association 116*, 2009–2022.

Cook RD (1986). Assessment of local influence. *Journal of the Royal Statistical Society: Series B 48*, 133–169.

Cook RD, Weisberg S (1982). *Residuals and Influence in Regression*. London: Chapman and Hall/CRC.

Cox DR, Snell EJ (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B 30*, 248–275.

Crowder M (1987). On linear and quadratic estimating functions. *Biometrika 74*, 591–597.

Dunn PK, Smyth GK (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics 5*, 236–244.

Eddelbuettel D (2022). RcppEigen: 'Rcpp' Integration for the 'Eigen' Templated Linear Algebra Library. R package version 0.3.3.9.3. https://cran.r-project.org/package=RcppEigen.

Eddelbuettel D (2023). Rcpp: Seamless R and C++ Integration. R package version 1.0.10. https://cran.r-project.org/package=Rcpp.

Fasiolo M (2023). mvnfast: Fast Multivariate Normal and Student's t Methods. R package version 0.2.8. https://cran.r-project.org/package=mvnfast.

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

Galvis DM, Bandyopadhyay D, Lachos VH (2014). Augmented mixed beta regression models for periodontal proportion data. *Statistics in Medicine 33*, 3759–3771.

Ghitany ME, Atieh B, Nadarajah S (2008). Lindley distribution and its application. *Mathematics and Computers in Simulation 78*, 493–506.

Godambe VP (1997). Estimating functions: A synthesis of least squares and maximum likelihood methods. In *I.V. Basawa, V.P. Godambe, R.L. Taylor (Eds.), Selected Proceedings of the Symposium on Estimating Functions*, pp. 5–15. California: Institute of Mathematical Statistics.

Goulet V (2022). expint: Exponential Integral and Incomplete Gamma Function. R package version 0.1-8. https://cran.r-project.org/package=expint.

Grassia A (1977). On a family of distributions with argument between 0 and 1 obtained by transformation of the gamma distribution and derived compound distributions. *Australian Journal of Statistics 19*, 108–114.

Hardin JW, Hilbe JM (2012). *Generalized Estimating Equations, 2nd Edition.* New York: Chapman and Hall/CRC.

Ibacache P, Paula GA, Cysneiros FJ (2013). Semiparametric additive models under symmetric distributions. *TEST 22*, 103–121.

Jørgensen B, Lundbye-Christensen S, Song PX-K, Sun L (1996). State-space models for multivariate longitudinal data of mixed types. *Canadian Journal of Statistics 24*, 385–402.

Kumaraswamy P (1980). A generalized probability density function for double-bounded random processes. *Journal of Hydrology 46*, 79–88.

Lee SY, Lu B, Song XY (2006). Assessing local influence for nonlinear structural equation models with ignorable mussing data. *Computational Statistics and Data Analysis 50*, 1356–1377.

Lee SY, Xu L (2004). Influence analyses of nonlinear mixed-effects models. *Computational Statistics and Data Analysis 45*, 321–341.

Liang KY, Zeger SL (1986). Longitudinal analysis using generalized linear models. *Biometrika 73*, 13–22.

Lindley DV (1958). Fiducial distribuitions and bayes's theorem. *Journal of the Royal Statistical Society: Series B 20*, 102–107.

Maechler M (2022). Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.5-3. https://cran.r-project.org/package=Matrix.

Manghi RF, Cysneiros FJA, Paula GA (2019). Generalized additive partial linear models for analyzing correlated data. *Computational Statistics and Data Analysis 129*, 49–60.

Mazucheli J, Menezes AFB, Chakraborty S (2019). On the one parameter unit-lindley distribution and its associated regression model for proportion data. *Journal of Applied Statistics 46*, 700–714.

McCullagh P, Nelder JA (1989). *Generalized Linear Models, 2nd Edition.* London: Chapman and Hall/CRC.

Mousa AM, El-Sheikh AA, Abdel-Fattah MA (2016). A gamma regression for bounded continuous variables. *Advances and Applications in Statistics 49*, 305–326.

Petterle RR, Bonat WH, Scarpin CT (2019). Quasi-beta longitudinal regression model applied to water quality index data. *Journal of Agricultural, Biological and Environmental Statistics 24*, 346–368.

Piegorsch WW, Casella G (1988). Confidence bands for logistic regression with restricted predictor variables. *Biometrics 44*, 739–750.

Poon W, Poon YS (1999). Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical Society: Series B 61*, 51–61.

Preisser JS, Qaqish BF (1996). Deletion diagnostics for generalised estimating equations. *Biometrika 83*, 551–562.

Qiu Z, Song PX-K, Tan M (2008). Simplex mixed-effects models for longitudinal proportional data. *Scandinavian Journal of Statistics 35*, 577–596.

Queiroz FF, Ferrari SLP (2023). Power logit regression for modeling bounded data. *Statistical Modelling*. https://doi.org/10.1177/1471082x221140157.

R Core Team (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org.

Rizopoulos D (2019). cvGEE: Cross-Validated Predictions from GEE. R package version 0.3-0. https://CRAN.R-project.org/package=cvGEE.

Solymos P (2023). pbapply: Adding Progress Bar to '*apply' Functions. R package version 1.7-0. https://cran.r-project.org/package=pbapply.

Stasinopoulos M (2022). gamlss.dist: Distributions for Generalized Additive Models for Location Scale and Shape. R package version 6.0-5. https://cran.r-project.org/package=gamlss.dist.

Stasinopoulos M (2023). gamlss: Generalised Additive Models for Location Scale and Shape. R package version 5.4-12. https://cran.r-project.org/package=gamlss.

Stasinopoulos MD, Righy RA, Gillian ZA, Voudouris V, de Bastiani F (2017). *Flexible Regression and Smoothing Using GAMLSS in R.* Florida: Chapman and Hall/CRC.

Tsuyuguchi A, Paula GA, Barros M (2020). Analysis of correlated birnbaum-saunders data based on estimating equations. *TEST 29*, 661–681.

Venezuela MK (2003). Modelos lineares generalizados para análise de dados com medidas repetidas [dissertação]. São Paulo: Universidade de São Paulo, Instituto de Matemática e Estatística.

Venezuela MK, Sandoval MC, Botter DA (2011). Local influence in estimating equations. *Computational Statistics and Data Analysis 55*, 1867–1883.

Wicklin R (2013). *Simulating Data with SAS.* North Carolina: SAS Institute.

Wood SN (2017). *Generalized Additive Models, 2nd Edition.* Boca Raton: Chapman and Hall/CRC.

Yan J, Fine J (2004). Estimating equations for association structures. *Statistics in Medicine 23*, 859–874.