**Diagnóstico para modelos lineares
generalizados para dados com
medidas repetidas
e valores faltantes**

Lucas Petri Damiani

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
MESTRE EM CIÊNCIAS

Programa: Estatística

Orientador: Profa. Dra. Denise A. Botter

Ao longo deste trabalho o autor recebeu auxílio financeiro do CNPq.

São Paulo, 20 de fevereiro de 2012

Diagnóstico para modelos lineares generalizados para dados com medidas repetidas e valores faltantes

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 10/05/2012. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof^{fa}. Dr^a. Denise Aparecida Botter (orientadora) - IME-USP.
- Prof. Dr. Gilberto Alvarenga Paula - IME-USP.
- Prof^{fa}. Dr^a. Maria Kelly Venezuela - INSPER.

Agradecimentos

Assim que peguei o esqueleto da dissertação, coloquei a capa, esbocei os títulos, e escrevi o título desse capítulo, não parei de pensar quantas pessoas e frases de efeito poderia colocar aqui. Sempre gostei de ler os agradecimentos dos autores em livros, teses e dissertações, e nunca gostei daqueles que agradecem a tudo e a todos, pois dessa forma, aparentemente, ninguém se sentia especial.

Enfim, deixei esse espaço em branco, pois quando chegada a hora, ele seria preenchido com carinho, cuidado e justiça! Mas a verdade é que com o passar do tempo, essa página é pequena para todos.

Muitíssimo obrigado ao pessoal que me aturou a maior parte do tempo no trabalho enquanto eu fazia as matérias: Flavio Sant’Ana Daher e Fábio Pizaruk. E aos que me aturaram no trabalho enquanto eu escrevia o texto: Roberta de Souza, João Ítalo Dias França, Patrícia Viana e Julio Pereira.

Experiências com orientadores de conhecidos meus não são animadoras, mas devo admitir que tive a sorte de ter Denise Botter e Mônica Sandoval como as minhas. Vocês são fantásticas.

Assim como tenho a sorte de ter minha mãe e irmão, sempre ao meu lado. Pessoas que amo e me fazem sentir orgulho.

E, por fim, devo deixar a dissertação dedicada à minha futura esposa, Bruna Bronhara, que em momento algum me deixou baixar a cabeça e ainda me ajudou a corrigir, escrever, reescrever e reescrever... Obrigado pela paciência. Te amo.

Resumo

A literatura dispõe de métodos de diagnóstico para avaliar o ajuste de modelos lineares generalizados (MLGs) para medidas repetidas baseado em equações de estimação generalizada (EEG). No entanto, tais métodos não contemplam a distribuição binomial nem bancos de dados com observações faltantes. O presente trabalho generalizou os métodos já desenvolvidos para essas duas situações. Na construção de gráficos de probabilidade meio-normal com envelope simulado para a distribuição binomial, foi proposto um método para geração de variáveis aleatórias com distribuição marginal binomial correlacionadas, baseado na convolução de variáveis com distribuição de Poisson independentes. Os métodos de diagnóstico desenvolvidos foram aplicados em dados reais e simulados.

Palavras-chave: Dados faltantes; Distribuições binomiais correlacionadas; Equações de estimação generalizadas; Medidas repetidas; Simulação de variáveis aleatórias; Técnicas de diagnóstico.

Abstract

Literature provides diagnostic methods to assess the fit of generalized linear models (GLM) for repeated measures based on generalized estimating equations (GEE). Still, such methods do not include the binomial distribution or databases with missing observations. This work generalizes the methods already developed for these two situations. A method for generating random variables with correlated marginal binomial distributions based on convolution of independent Poisson random variables has been proposed for the construction of half-normal probability plots. The diagnostic methods developed were applied to real and simulated data.

Keywords: Correlation structure; Diagnostic techniques; Generalized estimating equation; Missing data; Repeated measures; Simulation of random variables.

Sumário

Lista de Figuras	ix
1 Introdução	1
2 Equações de estimação para MLG com medidas repetidas	3
2.1 Distribuição binomial	11
3 Diagnóstico	13
3.1 Pontos Alavanca, Influentes e Aberrantes	13
3.2 Gráfico envelope	16
3.3 Simulando distribuições binomiais correlacionadas	18
4 Aplicações	25
4.1 Exemplos simulados	25
4.2 Exemplo 1 - Pronúncia da língua portuguesa	40
4.3 Exemplo 2 - Memória operacional	46
4.4 Exemplo 3 - Quimioterapia	55

5	Conclusão	59
A	Conjuntos de dados	61
A.1	Conjunto de dados do Exemplo 1	61
A.2	Conjunto de dados do Exemplo 2	73
A.3	Conjunto de dados do Exemplo 3	92
B	Algoritmos	97
B.1	Pontos de alavanca, influentes e aberrantes	97
B.2	Envelope	103
	Referências Bibliográficas	115

Lista de Figuras

3.1	Limite superior para ρ_{12} .	20
4.1	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 3$, $\mathbf{p}_1 = (0, 5; 0, 5; 0, 5; 0, 5; 0, 5)^T$ e estrutura de correlação uniforme ($\alpha = 0, 5$)	27
4.2	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 3$, $\mathbf{p}_2 = (0, 2; 0, 2; 0, 2; 0, 2; 0, 2)^T$ e estrutura de correlação uniforme ($\alpha = 0, 5$)	27
4.3	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 3$, $\mathbf{p}_3 = (0, 4; 0, 45; 0, 5; 0, 6)^T$ e estrutura de correlação uniforme ($\alpha = 0, 5$)	28
4.4	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 3$, $\mathbf{p}_4 = (0, 95; 0, 95; 0, 9; 0, 8)^T$ e estrutura de correlação uniforme ($\alpha = 0, 4$)	28
4.5	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 3$, $\mathbf{p}_5 = (0, 4; 0, 5; 0, 7; 0, 8)^T$ e estrutura de correlação uniforme ($\alpha = 0, 3$)	29

4.6	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_1 = (0, 5; 0, 5; 0, 5; 0, 5)^T$ e estrutura de correlação uniforme ($\alpha = 0, 5$)	30
4.7	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_2 = (0, 2; 0, 2; 0, 2; 0, 2)^T$ e estrutura de correlação uniforme ($\alpha = 0, 5$)	30
4.8	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_3 = (0, 4; 0, 45; 0, 5; 0, 6)^T$ e estrutura de correlação uniforme ($\alpha = 0, 5$)	31
4.9	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_4 = (0, 95; 0, 95; 0, 9; 0, 8)^T$ e estrutura de correlação uniforme ($\alpha = 0, 4$)	31
4.10	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_5 = (0, 4; 0, 5; 0, 7; 0, 8)^T$ e estrutura de correlação uniforme ($\alpha = 0, 3$)	32
4.11	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 50$, $\mathbf{p}_1 = (0, 5; 0, 5; 0, 5; 0, 5)^T$ e estrutura de correlação uniforme ($\alpha = 0, 5$)	32
4.12	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 50$, $\mathbf{p}_2 = (0, 2; 0, 2; 0, 2; 0, 2)^T$ e estrutura de correlação uniforme ($\alpha = 0, 5$)	33
4.13	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 50$, $\mathbf{p}_3 = (0, 4; 0, 45; 0, 5; 0, 6)^T$ e estrutura de correlação uniforme ($\alpha = 0, 5$)	33
4.14	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 50$, $\mathbf{p}_4 = (0, 95; 0, 95; 0, 9; 0, 8)^T$ e estrutura de correlação uniforme ($\alpha = 0, 4$)	34

4.15	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 50$, $\mathbf{p}_5 = (0, 4; 0, 5; 0, 7; 0, 8)^T$ e estrutura de correlação uniforme ($\alpha = 0, 3$)	34
4.16	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_1 = (0, 5; 0, 5; 0, 5; 0, 5)^T$ e estrutura de correlação autorregressiva ($\alpha = 0, 8$)	35
4.17	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_2 = (0, 2; 0, 2; 0, 2; 0, 2)^T$ e estrutura de correlação autorregressiva ($\alpha = 0, 8$)	36
4.18	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_3 = (0, 4; 0, 45; 0, 5; 0, 6)^T$ e estrutura de correlação autorregressiva ($\alpha = 0, 8$)	36
4.19	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_4 = (0, 95; 0, 95; 0, 9; 0, 8)^T$ e estrutura de correlação autorregressiva ($\alpha = 0, 5$)	37
4.20	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_5 = (0, 4; 0, 5; 0, 7; 0, 8)^T$ e estrutura de correlação autorregressiva ($\alpha = 0, 5$)	37
4.21	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $\mathbf{k} = (7, 10, 15)$, $\mathbf{p}_6 = (0, 7; 0, 5; 0, 4)$ e estrutura de correlação autorregressiva ($\alpha = 0, 5$)	39
4.22	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 50$, $\mathbf{k} = (7, 10, 15)$, $\mathbf{p}_6 = (0, 7; 0, 5; 0, 4)$ e estrutura de correlação autorregressiva ($\alpha = 0, 5$)	39
4.23	Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 20$, $\mathbf{k} = (7, 10, 15)$, $\mathbf{p}_6 = (0, 7; 0, 5; 0, 4)$ e estrutura de correlação autorregressiva ($\alpha = 0, 5$)	40

4.24	Diagnóstico - Pontos alavanca, influentes e aberrantes - Exemplo 1	44
4.25	Gráfico de probabilidade meio-normal com envelope simulado - Exemplo 1 . . .	45
4.26	Média de acertos por idade e número de sílabas - Exemplo 2.	47
4.27	Média de acertos por idade e número de sílabas estimado - Exemplo 2.	50
4.28	Diagnóstico - Pontos alavanca, influentes e aberrantes - Exemplo 2.	51
4.29	Gráfico de probabilidade meio-normal com envelope simulado - Exemplo 2. . .	52
4.30	Gráfico de probabilidade meio-normal com envelope simulado - Exemplo 2 - 3 e 4 anos	54
4.31	Gráfico de probabilidade meio-normal com envelope simulado - Exemplo 2 - 5 e 6 anos	54
4.32	Gráfico de probabilidade meio-normal com envelope simulado - Exemplo 3 . .	56
4.33	Diagnóstico com dados completos - Pontos de alavanca, influentes e aberrantes - Exemplo 3 - Dados completos	57
4.34	Diagnóstico com dados faltantes - Pontos de alavanca, influentes e aberrantes - Exemplo 3 - 15 observações faltantes	58

Capítulo 1

Introdução

Estudos com unidades experimentais avaliadas em diferentes situações ou tempos são comuns em diferentes áreas do conhecimento. Em estatística, são denominados estudos com medidas repetidas ou longitudinais. Nesses planejamentos, as repetições podem ocorrer em situações ou tempos fixos, ou então, em tempos não necessariamente controlados (Afune et al., 2005). Modelos Lineares Generalizados (MLG) para medidas repetidas ajustados por meio de equações de estimação generalizadas (EEG) são alternativas possíveis de modelagem de dados provenientes de estudos em que as situações ou os tempos são fixos. O desenvolvimento de técnicas de diagnóstico para modelos desse tipo é o tema desse trabalho.

Os MLG foram desenvolvidos por Nelder e Wedderburn (1972) e são alternativas para casos onde a distribuição da variável resposta não segue a distribuição normal, embora essa distribuição também pertença a essa classe. Atualmente, os MLG são muito utilizados na prática e encabeçam uma série de modelos derivados, entre eles os propostos no artigo de Liang e Zeger (1986), que introduziram um método para ajustar MLG com medidas repetidas.

Posteriormente, Prentice e Zhao (1991) também propuseram um método para lidar com medidas repetidas que utilizavam equações de estimação generalizadas. A literatura chama esse método de EEG2, enquanto a proposta de Liang e Zeger (1986) é conhecida por EEG1. A diferença entre os métodos está na forma de estimação dos parâmetros de correlação. Para

a EEG1, os parâmetros do modelo de regressão são estimados via equações de estimação generalizadas e os parâmetros que definem a matriz de correlações são estimados por meio do método dos momentos. Já para a EEG2, os dois conjuntos de parâmetros são estimados por meio de equações de estimação generalizadas. O Capítulo 2 baseia-se no método de estimação proposto em EEG1. Uma discussão mais completa sobre o assunto encontra-se em Hardin e Hilbe (2003).

A literatura apresenta diversos métodos de diagnóstico para modelos de regressão com observações independentes, como os encontrados em Paula (2010), por exemplo. Pontos de alavanca, influentes e aberrantes, assim como gráficos envelope com banda de confiança, são técnicas descritas por esse autor.

Venezuela (2003) apresentou uma extensão dos métodos para detecção de pontos influentes, alavanca e aberrantes para dados com medidas repetidas analisados por meio do método EEG1 e implementou algoritmos em *S-Plus* para bancos de dados completos. Além disso, desenvolveu códigos para a criação de gráficos envelope para quatro distribuições da família exponencial (normal, gama, Poisson e Bernoulli), sempre considerando bancos de dados completos. Para esse fim, Venezuela (2003) utilizou os algoritmo de Park et al. (1996) e Park e Shin (1998) para geração de variáveis correlacionadas, exceto no caso da distribuição normal, para a qual utilizou a função *rmvnorm* (Genz, 1992).

Diante disso, o presente trabalho tem como objetivo estender os algoritmos de Venezuela (2003) para bancos com dados faltantes, sejam esses faltantes devido ao desenho do experimento ou por omissão, e para possibilitar o diagnóstico de modelos ajustados com distribuição binomial. Para isso, foi necessário criar um método capaz de gerar distribuições binomiais não identicamente distribuídas e correlacionadas. Além disso, os códigos individuais para cada distribuição foram agrupados em uma única função. Os algoritmos foram escritos no *software R* versão 2.13.2.

As EEG para MLG com medidas repetidas são introduzidas no Capítulo 2. As técnicas de diagnóstico utilizadas e os algoritmos implementados estão descritos no Capítulo 3. Por fim, no Capítulo 4, aplicam-se as técnicas estudadas em exemplos simulados e reais.

Capítulo 2

Equações de estimação para modelos lineares generalizados com medidas repetidas

Os Modelos Lineares Generalizados com medidas repetidas são uma extensão dos modelos inicialmente propostos por Nelder e Wedderburn (1972). Neste capítulo apresenta-se a definição do modelo proposto por Liang e Zeger (1986) utilizando notação similar a de Venezuela et al. (2007).

O presente trabalho não se ateve a provas, condições de existência e teoria assintótica. Uma leitura mais completa encontra-se em Godambe (1991), Venezuela et al. (2007), Artes e Botter (2005) e Hardin e Hilbe (2003).

O MLG para medidas repetidas é utilizado para modelar experimentos, como o nome diz, que são medidos mais do que uma vez na mesma unidade amostral. Nesse modelo, assumi-se uma matriz de correlação de tamanho $(t \times t)$, onde t é o número de tempos que se avaliaram as unidades amostrais.

Mais precisamente, seja $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{it_i})^T$, o vetor de respostas para a i -ésima unidade experimental, $i = 1, \dots, n$. Associado à unidade i na ocasião j , $j = 1, \dots, t_i$, observa-se o vetor

4CAPÍTULO 2. EQUAÇÕES DE ESTIMAÇÃO PARA MLG COM MEDIDAS REPETIDAS

$\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \dots, x_{ijp})^T$ de valores de p covariáveis. Assim, seja $\mathbf{X}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{it_i})^T$, a matriz de valores das p covariáveis para os t_i momentos de avaliação. Assumindo que a distribuição de y_{ij} pertença à família exponencial, sua função densidade de probabilidade (ou função de probabilidade) é dada por:

$$f(y_{ij}; \theta_{ij}, \phi) = \exp\{\phi[y_{ij}\theta_{ij} - b(\theta_{ij})] + c(y_{ij}, \phi)\}, \quad (2.1)$$

sendo $b(\cdot)$ e $c(\cdot, \cdot)$ funções conhecidas e $\phi > 0$ o parâmetro de precisão (ϕ^{-1} é conhecido como parâmetro de dispersão). De (2.1) vem que $E(y_{ij}) = \mu_{ij} = d b(\theta_{ij}) / d \theta_{ij} = b'(\theta_{ij})$, $Var(y_{ij}) = \phi^{-1} V(\mu_{ij})$, sendo $V(\mu_{ij}) = d^2 b(\theta_{ij}) / d \theta_{ij}^2 = b''(\theta_{ij})$ denominada função de variância.

A média μ_{ij} associa-se ao preditor linear $\eta_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}$ por meio de uma função monótona e diferenciável, denominada função de ligação $g(\cdot)$, sendo $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$, $p < n$, um vetor de parâmetros a ser estimado, de modo que:

$$g(\mu_{ij}) = \eta_{ij}.$$

Como a i -ésima unidade experimental foi observada em t_i ocasiões, definimos a matriz de covariâncias de $\mathbf{y}_i$ como:

$$\text{Cov}(\mathbf{y}_i) = \phi^{-1} \mathbf{A}_i^{1/2} \mathbf{R}_i \mathbf{A}_i^{1/2} = \boldsymbol{\Sigma}_i,$$

em que $\mathbf{R}_i$ é a verdadeira matriz de correlações de $\mathbf{y}_i$ e

$$\mathbf{A}_i = \text{diag} \{b''(\theta_{ij})\}$$

com dimensão $(t_i \times t_i)$, $j = 1, \dots, t_i$.

Segundo Liang e Zeger (1986) e Venezuela et al. (2007), a equação de estimação para

estimar o vetor de parâmetros $\boldsymbol{\beta}$, considerando a verdadeira matriz de correlações $\mathbf{R}_i$, é dada por

$$\Psi_n(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{D}_i^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = 0, \quad (2.2)$$

sendo, $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{it_i})^T$ e

$$\mathbf{D}_i^T = -\mathbf{X}_i^T \boldsymbol{\Delta}_i \mathbf{A}_i$$

com

$$\boldsymbol{\Delta}_i = \text{diag} \left\{ \frac{\partial \theta_{ij}}{\partial \eta_{ij}} \right\}.$$

Sob condições gerais de regularidade pode-se demonstrar que $\hat{\boldsymbol{\beta}}$, raiz da equação (2.2), é um estimador consistente para $\boldsymbol{\beta}$ e que

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(0, \mathbf{J}_*^{-1})$$

com

$$\mathbf{J}_* = \lim_{n \rightarrow \infty} \frac{\mathbf{J}_{n*}}{n},$$

sendo $\mathbf{J}_{n*}$ a matriz de informação de Godambe de $\boldsymbol{\beta}$ associada a $\Psi_n(\boldsymbol{\beta})$ dada por

$$\mathbf{J}_{n*} = \left\{ \sum_{i=1}^n \mathbf{S}_i(\boldsymbol{\beta}) \right\} \left\{ \sum_{i=1}^n \mathbf{V}_i(\boldsymbol{\beta}) \right\}^{-1} \left\{ \sum_{i=1}^n \mathbf{S}_i(\boldsymbol{\beta}) \right\}, \quad (2.3)$$

em que

$$\begin{aligned} \mathbf{S}_i(\boldsymbol{\beta}) &= E_{\boldsymbol{\beta}} \left[\frac{\partial}{\partial \boldsymbol{\beta}^T} \mathbf{D}_i^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) \right] \\ &= \mathbf{D}_i^T \boldsymbol{\Sigma}_i^{-1} E_{\boldsymbol{\beta}} \left[\frac{\partial}{\partial \boldsymbol{\beta}^T} (\mathbf{y}_i - \boldsymbol{\mu}_i) \right] \\ &= -\mathbf{D}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_i \end{aligned}$$

e

$$\begin{aligned} \mathbf{V}_i(\boldsymbol{\beta}) &= E_{\boldsymbol{\beta}} [\mathbf{D}_i^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) (\mathbf{y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_i] \\ &= \mathbf{D}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_i = -\mathbf{S}_i(\boldsymbol{\beta}). \end{aligned}$$

Assim, um estimador consistente para a matriz de covariâncias de $\hat{\boldsymbol{\beta}}$, denominado *naive* ou *model-based*, é dado por

$$\hat{\mathbf{J}}_{n*}^{-1} = \left\{ \sum_{i=1}^n \hat{\mathbf{D}}_i \hat{\boldsymbol{\Sigma}}_i^{-1} \hat{\mathbf{D}}_i \right\}^{-1}, \quad (2.4)$$

sendo $\hat{\mathbf{D}}_i$ e $\hat{\boldsymbol{\Sigma}}_i^{-1}$ as estimativas de $\mathbf{D}_i$ e $\boldsymbol{\Sigma}_i^{-1}$, respectivamente, avaliadas em $\hat{\boldsymbol{\beta}}$. Como, em geral, na prática $\mathbf{R}_i$ é desconhecida, Liang e Zeger (1986) definem $\mathbf{R}(\boldsymbol{\alpha})$, uma matriz de correlação de trabalho de dimensão $(t \times t)$, que depende de um vetor de parâmetros $\boldsymbol{\alpha}$ de dimensão $(s \times 1)$

e é a mesma para todas as unidades experimentais quando essas são avaliadas em t ocasiões.

No caso da i -ésima unidade experimental não possuir todas as t observações, utiliza-se a matriz $\mathbf{R}_i(\boldsymbol{\alpha})$, obtida da matriz de correlação de trabalho $\mathbf{R}(\boldsymbol{\alpha})$, retirando-se desta as colunas e linhas correspondentes às observações faltantes. Quando a unidade experimental é observada em todas as t ocasiões, $\mathbf{R}_i(\boldsymbol{\alpha}) = \mathbf{R}(\boldsymbol{\alpha})$.

Para estimar $\boldsymbol{\beta}$ com base em $\mathbf{R}_i(\boldsymbol{\alpha})$ define-se a *equação de estimação generalizada* como

$$\Psi_n(\boldsymbol{\beta}, \boldsymbol{\alpha}(\boldsymbol{\beta}, \phi)) = \sum_{i=1}^n \mathbf{D}_i^T \boldsymbol{\Omega}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = 0, \quad (2.5)$$

sendo

$$\boldsymbol{\Omega}_i = \boldsymbol{\Omega}_i(\boldsymbol{\alpha}, \phi) = \phi^{-1} \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}.$$

Para a estimação dos parâmetros $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ e ϕ utiliza-se um processo iterativo que combina o método *scoring* de Fisher para estimar $\boldsymbol{\beta}$ com o método dos momentos para estimar $\boldsymbol{\alpha}$ e ϕ . Venezuela et al. (2007) mostra que, dadas as estimativas de $\boldsymbol{\alpha}$ e ϕ , o vetor de parâmetros $\boldsymbol{\beta}$ é estimado da seguinte forma

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left\{ \left[\sum_{i=1}^n \hat{\mathbf{D}}_i^T \hat{\boldsymbol{\Omega}}_i^{-1} \hat{\mathbf{D}}_i \right]^{-1} \left[\sum_{i=1}^n \hat{\mathbf{D}}_i^T \hat{\boldsymbol{\Omega}}_i^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i) \right] \right\}^{(m)}, \quad (2.6)$$

em que $m = 0, 1, 2, 3, \dots$ é o número de iterações. O índice m indica que as matrizes e vetores envolvidos são atualizados pelas estimativas de $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ e ϕ da m -ésima iteração. A estimativa inicial $\boldsymbol{\beta}^{(0)}$ é obtida de um MLG assumindo independência entre as observações. A estimativa final de $\boldsymbol{\beta}$ é encontrada quando a diferença das estimativas $\boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}^{(m-1)}$ é inferior a certo limite ou com base em um número máximo de iterações escolhido. Este trabalho utilizou a função *gee* do pacote *gee* do *software* R na versão 2.13.2, em que o limite padrão para assumir

8CAPÍTULO 2. EQUAÇÕES DE ESTIMAÇÃO PARA MLG COM MEDIDAS REPETIDAS

convergência do processo é 0,001 e o número máximo de iterações permitidos é 25.

Como $\hat{\mathbf{D}}_i = -\hat{\mathbf{A}}_i \hat{\mathbf{\Delta}}_i \mathbf{X}_i$, reescreve-se a expressão (2.6) da seguinte forma

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \left\{ \left[\sum_{i=1}^n \mathbf{X}_i^T \hat{\mathbf{W}}_i \mathbf{X}_i \right]^{-1} \left[\sum_{i=1}^n \mathbf{X}_i^T \hat{\mathbf{W}}_i \mathbf{z}_i \right] \right\}^{(m)} \quad (2.7)$$

com $\hat{\mathbf{W}}_i = \hat{\mathbf{\Delta}}_i \hat{\mathbf{A}}_i \hat{\boldsymbol{\Omega}}^{-1} \hat{\mathbf{\Delta}}_i \hat{\mathbf{A}}_i$ e $\mathbf{z}_i = \hat{\boldsymbol{\eta}}_i + (\hat{\mathbf{A}}_i \hat{\mathbf{\Delta}}_i)^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)$.

Mais detalhes estão em Venezuela et al. (2007) e Artes e Botter (2005).

Liang e Zeger (1986) mostram que, sob determinadas condições de regularidade, e assumindo que $\hat{\boldsymbol{\alpha}}$ e $\hat{\phi}$ sejam estimadores consistentes de $\boldsymbol{\alpha}$ e ϕ , respectivamente, tem-se que $\hat{\boldsymbol{\beta}}$, raiz da equação (2.5), é um estimador consistente de $\boldsymbol{\beta}$ e

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(0, \mathbf{J}^{-1})$$

com

$$\mathbf{J} = \lim_{n \rightarrow \infty} \frac{\mathbf{J}_n}{n},$$

sendo $\mathbf{J}_n$ a matriz de informação de Godambe de $\boldsymbol{\beta}$ associada a Ψ_n definida como em (2.3), com

$$\begin{aligned}
\mathbf{S}_i(\boldsymbol{\beta}) &= E_{\beta} \left[\frac{\partial}{\partial \boldsymbol{\beta}^T} \mathbf{D}_i^T \boldsymbol{\Omega}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) \right] \\
&= \mathbf{D}_i^T \boldsymbol{\Omega}_i^{-1} E_{\beta} \left[\frac{\partial}{\partial \boldsymbol{\beta}^T} (\mathbf{y}_i - \boldsymbol{\mu}_i) \right] \\
&= -\mathbf{D}_i^T \boldsymbol{\Omega}_i^{-1} \mathbf{D}_i
\end{aligned}$$

e

$$\begin{aligned}
\mathbf{V}_i(\boldsymbol{\beta}) &= E_{\beta} [\mathbf{D}_i^T \boldsymbol{\Omega}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) (\mathbf{y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Omega}_i^{-1} \mathbf{D}_i] \\
&= E_{\beta} [\mathbf{D}_i^T \boldsymbol{\Omega}_i^{-1} \text{Cov}(\mathbf{y}_i) \boldsymbol{\Omega}_i^{-1} \mathbf{D}_i].
\end{aligned}$$

Estima-se consistentemente a matriz de covariâncias de $\hat{\boldsymbol{\beta}}$, usando

$$\hat{\mathbf{J}}_n^{-1} = \left\{ \sum_{i=1}^n -\hat{\mathbf{D}}_i^T \hat{\boldsymbol{\Omega}}_i^{-1} \hat{\mathbf{D}}_i \right\}^{-1} \left\{ \hat{\mathbf{D}}_i^T \hat{\boldsymbol{\Omega}}_i^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i) (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)^T \hat{\boldsymbol{\Omega}}_i^{-1} \hat{\mathbf{D}}_i \right\} \left\{ \sum_{i=1}^n -\hat{\mathbf{D}}_i^T \hat{\boldsymbol{\Omega}}_i^{-1} \hat{\mathbf{D}}_i \right\}^{-1} \quad (2.8)$$

sendo que $\hat{\mathbf{D}}_i$, $\hat{\boldsymbol{\Omega}}_i$ e $\hat{\boldsymbol{\mu}}_i$ são obtidas a partir das estimativas de $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ e ϕ .

Esse estimador é denominado *robusto*, *empírico* ou *sanduíche*. Quando a matriz de correlações de trabalho for igual à verdadeira matriz de correlações dos dados ($\mathbf{y}_i$), o estimador *robusto* (2.8) será igual ao estimador *naive* apresentado em (2.4).

Liang e Zeger (1986) introduzem algumas estruturas para a matriz de correlações de trabalho $\mathbf{R}(\boldsymbol{\alpha})$ e apresentam o estimador $\hat{\boldsymbol{\alpha}}$ para cada caso. Essas estruturas são listadas abaixo e estão implementadas no pacote *gee* do *software* R. Entre parênteses estão os nomes das mesmas adotados no pacote *gee*.

Independente (*independence*): define-se $\mathbf{R}(\boldsymbol{\alpha}) = \mathbf{I}_t$, sendo $\mathbf{I}_t$ a matriz identidade de ordem t .

Não estruturada (*unstructured*): utilizada quando a matriz de correlações não apresenta uma estrutura bem definida. Define-se $\mathbf{R}(\boldsymbol{\alpha})$ como

$$\mathbf{R}(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & & & & \\ \alpha_{21} & 1 & & & \\ \alpha_{31} & \alpha_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ \alpha_{t1} & \dots & \dots & \alpha_{t(t-1)} & 1 \end{bmatrix}.$$

Autorregressiva de ordem M (*AR-M*): estrutura regularmente utilizada para repetições ao longo do tempo. Impõe que a correlação é maior entre as observações que distam M repetições, e regride conforme a distância entre os tempos se afastam de M . Para $M = 1$, a matriz de correlações de trabalho é definida da seguinte forma:

$$\mathbf{R}(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & & & & & \\ \alpha & 1 & & & & \\ \alpha^2 & \alpha & 1 & & & \\ \alpha^3 & \alpha^2 & \alpha & 1 & & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \\ \alpha^t & \dots & \alpha^3 & \alpha^2 & \alpha & 1 \end{bmatrix}.$$

Dependência estacionária de ordem M (*stat-M-dep*): possui redução graduada, idêntica à estrutura autorregressiva de ordem 1 até a ordem especificada (M). Para distância entre as observações de ordem $M+1$ ou maior, assume-se que não há correlação entre as observações.

A matriz abaixo exemplifica uma estrutura de dependência estacionária de ordem $M = 2$.

$$\mathbf{R}(\alpha) = \begin{bmatrix} 1 & & & & & & \\ \alpha & 1 & & & & & \\ \alpha^2 & \alpha & 1 & & & & \\ 0 & \alpha^2 & \alpha & 1 & & & \\ 0 & 0 & \alpha^2 & \alpha & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \\ 0 & 0 & 0 & \dots & \alpha^2 & \alpha & 1 \end{bmatrix}.$$

Uniforme ("exchangeable"): como o nome diz, impõe que todas as correlações duas a duas sejam iguais a α .

$$\mathbf{R}(\alpha) = \begin{bmatrix} 1 & & & & \\ \alpha & 1 & & & \\ \alpha & \alpha & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ \alpha & \dots & \dots & \alpha & 1 \end{bmatrix}.$$

Por fim, vale lembrar que se pode assumir $\mathbf{R}(\alpha)$ como uma matriz conhecida.

2.1 Distribuição binomial

Nesta seção, considera-se uma situação em que para cada unidade experimental i observa-se, em cada ocasião j , o número de sucessos y_{ij} em k_{ij} repetições independentes de um experimento com probabilidade de sucesso p_{ij} . Assume-se então que $y_{ij} \sim \text{Bin}(k_{ij}, p_{ij})$, para $i = 1, \dots, n$ e $j = 1, \dots, t_i$. A função de probabilidade de y_{ij} é dada por

$$\begin{aligned}
p(y_{ij}) &= \binom{k_{ij}}{y_{ij}} p_{ij}^{y_{ij}} (1 - p_{ij})^{k_{ij} - y_{ij}} \\
&= \exp \left\{ \log \left[\binom{k_{ij}}{y_{ij}} \right] + y_{ij} \log \left(\frac{p_{ij}}{1 - p_{ij}} \right) + k_{ij} \log(1 - p_{ij}) \right\}. \quad (2.9)
\end{aligned}$$

Verifica-se que (2.9) está escrita na forma da família exponencial dada em (2.1), sendo $\phi = 1$, $\theta_{ij} = \log(p_{ij}/(1 - p_{ij})) = \log(\mu_{ij}/(k_{ij} - \mu_{ij}))$, $c(y_{ij}^*, \phi) = \log \binom{k_{ij}}{y_{ij}}$,

$$b(\theta_{ij}) = -k_{ij} \log(1 - p_{ij}) = k_{ij} \log(1 + e^{\theta_{ij}}).$$

Calculando $b'(\theta_{ij})$ e $b''(\theta_{ij})$ obtemos $b'(\theta_{ij}) = \mu_{ij} = k_{ij} p_{ij}$ e $b''(\theta_{ij}) = k_{ij} p_{ij} (1 - p_{ij})$.

Considerando a função de ligação canônica temos

$$g(\mu_{ij}) = \log \left(\frac{\mu_{ij}}{k_{ij} - \mu_{ij}} \right) = \log \left(\frac{p_{ij}}{1 - p_{ij}} \right) = \eta_{ij}.$$

Para ajustar um MLG com medidas repetidas no R, com distribuição marginal binomial, utiliza-se o pacote *gee*. Até a versão do pacote utilizada nesse trabalho (4.13-17), ϕ deve ser especificado como conhecido para que se utilize a definição apresentada do modelo, inserindo os termos $(\dots, \text{scale.fix} = \text{TRUE}, \text{scale.value} = 1)$ na linha de comando.

Capítulo 3

Diagnóstico

Técnicas de diagnóstico para MLG com observações independentes são apresentadas em Paula (2010). Para detectar pontos de alavanca, observações influentes e discrepantes, são utilizados os elementos da diagonal principal da matriz de projeção (*matriz chapéu*), a distância de Cook (Cook, 1977) e os resíduos padronizados, além dos gráficos de probabilidade meio-normal com envelope simulado.

Venezuela (2003) e Venezuela et al. (2007) propõem técnicas de diagnóstico para os MLG com medidas repetidas e desenvolvem um algoritmo para aplicar essas técnicas para as distribuições de Poisson, gama e binomial, sendo essa última limitada a respostas dicotômicas ($k = 1$). Além disso, consideram que as unidades experimentais são todas avaliadas nas t ocasiões, ou seja, que não há observações faltantes.

O presente trabalho tem o intuito de estender o algoritmo para aceitar respostas faltantes e para que o modelo para distribuição binomial aceite $k \geq 1$.

3.1 Pontos Alavanca, Influentes e Aberrantes

Para definir o resíduo padronizado, Venezuela et al. (2007) partem da convergência do processo iterativo dado em (2.7) para estimar β , ou seja,

$$\hat{\beta} = \left[\sum_{i=1}^n \mathbf{X}_i^T \hat{\mathbf{W}}_i \mathbf{X}_i \right]^{-1} \left[\sum_{i=1}^n \mathbf{X}_i^T \hat{\mathbf{W}}_i \mathbf{z}_i \right], \quad (3.1)$$

com $\hat{\mathbf{W}}_i = \hat{\Delta}_i \hat{\mathbf{A}}_i \hat{\Omega}^{-1} \hat{\Delta}_i \hat{\mathbf{A}}_i$ e $\mathbf{z}_i = \hat{\eta}_i + (\hat{\mathbf{A}}_i \hat{\Delta}_i)^{-1}(\mathbf{y}_i - \hat{\mu}_i)$.

Para eliminar o somatório em (3.1), Venezuela et al. (2007) definem $\hat{\mathbf{W}} = \text{diag}(\hat{\mathbf{W}}_1, \dots, \hat{\mathbf{W}}_n)$ como uma matriz bloco diagonal de dimensão $(N \times N)$, sendo $N = nt$. Nesse caso, (3.1) pode ser simplificada como

$$\hat{\beta} = \left(\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X} \right)^{-1} \mathbf{X}^T \hat{\mathbf{W}} \mathbf{z}, \quad (3.2)$$

em que $\mathbf{X}$ é a matriz de planejamento de dimensão $(N \times p)$ e $\mathbf{z} = (\mathbf{z}_1^T, \dots, \mathbf{z}_n^T)^T$.

Assim como no Capítulo 2, não é necessário supor planejamento completo, com todas as t repetições por unidade experimental. Para estender as técnicas de diagnóstico define-se, então, $N = \sum_{i=1}^n t_i$.

A mudança na definição da estrutura da matriz $\hat{\mathbf{W}}$, que a transforma em uma matriz bloco diagonal de diferentes tamanhos é a diferença básica entre o algoritmo proposto em Venezuela (2003) e o estendido neste trabalho. Ao permitir tamanhos diferentes para as matrizes $\mathbf{A}_i$ e Δ_i , considera-se a situação com dados faltantes que não era contemplada no algoritmo original.

A equação em (3.2) permite mostrar que $\hat{\beta}$ pode ser interpretado como a solução de mínimos quadrados da regressão normal linear de $\hat{\mathbf{W}}^{1/2} \mathbf{z}$ tendo como matriz de planejamento $\hat{\mathbf{W}}^{1/2} \mathbf{X}$. A partir daí, define-se a matriz chapéu, ou matriz de projeção ortogonal, como sendo uma matriz bloco diagonal $(N \times N)$ dada por $\mathbf{H} = \text{diag}(\mathbf{H}_1, \dots, \mathbf{H}_n)$, com

$$\mathbf{H}_i = \hat{\mathbf{W}}_i^{1/2} \mathbf{X}_i (\mathbf{X}_i^T \hat{\mathbf{W}}_i \mathbf{X}_i)^{-1} \mathbf{X}_i^T \hat{\mathbf{W}}_i^{1/2}.$$

A diagonal de $\mathbf{H}$ é, então, utilizada para detectar pontos de alavanca, assim como Paula (2010) escreve para os MLG e Tan et al. (1997) para os modelos de regressão logística com medidas repetidas. Seja $\mathbf{h}$ um vetor de dimensão $(N \times 1)$, cujos elementos h_{ij} , $i = 1, \dots, n$; $j = 1, \dots, t_i$ são os elementos da diagonal principal da matriz $\mathbf{H}$. Um ponto de alavanca é definido como uma observação que se destaca por apresentar um valor diferente dos demais com respeito às variáveis explicativas. Um alto valor de h_{ij} indica a influência de $\mathbf{x}_{ij}$ sobre o correspondente valor ajustado $\hat{y}_{ij}$.

Supondo que a influência de cada ponto seja a mesma sobre os valores ajustados, espera-se que cada valor da diagonal principal de $\mathbf{H}$ esteja próximo ao $\text{tr}(\mathbf{H})/N = p/N$, já que $\mathbf{H}$ é idempotente. Assim, um ponto é considerado de alavanca quando $h_{ij} \geq 2p/N$.

Analogamente, a i -ésima unidade experimental pode ser definida como um ponto de alavanca se $\mathbf{h}_i = t_i^{-1} \left(\sum_{j=1}^{t_i} h_{ij} \right) \geq 2p/N$.

Esses resultados foram implementados em dois gráficos: o primeiro cruzando h_{ij} pelo índice i e o segundo cruzando $\mathbf{h}_i$ também pelo índice i . No algoritmo desenvolvido nesse trabalho (Apêndice B.1) implementou-se nos gráficos de diagnóstico o corte $2p/N$ para melhor identificação dos pontos de alavanca.

A definição do resíduo padronizado (r_{SD}) e da distância de Cook (CD) se mantêm iguais às apresentadas em Venezuela et al. (2007). São elas:

$$(r_{SD})_{ij} = \frac{\mathbf{e}_{(ij)}^T \hat{\mathbf{W}}_i^{1/2} (\hat{\mathbf{A}}_i \hat{\mathbf{\Delta}}_i)^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)}{\sqrt{1 - h_{ij}}} \quad (3.3)$$

e

$$(CD)_{ij} = (r_{SD})_{ij}^2 \frac{h_{ij}}{p(1 - h_{ij})},$$

em que $\mathbf{e}_{(ij)}^T$ é um vetor de dimensão t_i com a posição j tendo valor 1 e nas demais posições o

valor zero, $i = 1, \dots, n$ e $j = 1, \dots, t_i$.

Um valor de $(r_{SD})_{ij}$ distante de sua média (zero) é um indício de um ponto aberrante. Mas vale lembrar que a distribuição de (r_{SD}) não é necessariamente simétrica. No caso da distribuição binomial com probabilidade de sucesso distante de 0,5, é comum que a distribuição dos erros seja assimétrica, como no exemplo da Seção 4.3.

Já, a distância de Cook é utilizada para investigar pontos influentes (Cook (1979) e Pregibon (1981)). $(CD)_{ij}$ é sempre positivo e valores que se destacam dos demais são considerados pontos influentes.

As medidas citadas nesta seção foram implementadas em códigos individuais para as distribuições normal, Poisson, binomial ($k = 1$) e gama em Venezuela (2003). O algoritmo criado na linguagem R, presente no Anexo B.1, contempla o primeiro objetivo desse texto, inserindo em um único código as medidas para todas essas distribuições e estendendo o algoritmo para a distribuição binomial com $k > 1$.

Além disso, atualizaram-se os códigos para que o programa admitisse bancos de dados com informações faltantes para as variáveis respostas ao longo das ocasiões de avaliação.

3.2 Gráfico envelope

Outra técnica para avaliar o ajuste de um MLG com medidas repetidas utilizando EEG é o gráfico de probabilidade meio-normal com envelope simulado (Atkinson, 1985). O algoritmo para construção do gráfico encontra-se passo a passo em Neter et al. (1996) e Venezuela et al. (2007).

Para a construção do gráfico dispõe-se o l -ésimo resíduo padronizado (r_{SD}) versus o correspondente valor esperado da estatística de ordem, em valor absoluto, da distribuição normal padrão dado por

$$\Phi^{-1} \left(\frac{l + N - 1/8}{2N + 1/2} \right) \quad (3.4)$$

em que $\Phi(\cdot)$ é a função de distribuição acumulada da normal padrão, $l = 1, \dots, N$. Esse gráfico pode ser gerado sem que os resíduos apresentem distribuição normal. Quando isso ocorre, não se espera que os resíduos se disponham numa reta com 45° de inclinação, como seria esperado se os resíduos tivessem distribuição normal. Portanto, utilizar os valores da equação (3.4) ou o índice l , $l = 1, \dots, N$, no eixo das abscissas, fornece a mesma conclusão.

A construção do gráfico de probabilidade meio-normal com envelope simulado segue os sete passos descritos a seguir.

- P1. Para cada unidade experimental i , $i = 1, \dots, n$, simule um vetor de respostas de tamanho t_i , levando em consideração o vetor de médias e a matriz de covariâncias ajustados aos dados originais.
- P2. Ajuste, às respostas simuladas no passo 1, o mesmo modelo ajustado aos dados originais.
- P3. Calcule os resíduos padronizados conforme expresso na equação (3.3) e ordene os seus valores absolutos.
- P4. Repita os três primeiros passos mais 24 vezes. Aqui, define-se $(r_{SD})_{lm}$ como sendo o l -ésimo valor absoluto ordenado do resíduo padronizado pertencente à m -ésima simulação, $l = 1, \dots, N$ e $m = 1, \dots, 25$.
- P5. Determine o mínimo, a mediana e o máximo dos menores valores absolutos dos resíduos padronizados de todas as simulações, isto é, entre os resíduos $(r_{SD})_{1m}$, $m = 1, \dots, 25$.
- P6. Repita o passo anterior para os segundos menores valores absolutos dos resíduos das simulações, isto é, entre os resíduos $(r_{SD})_{2m}$; em seguida, para os terceiros resíduos $(r_{SD})_{3m}$, e assim sucessivamente, até os maiores valores absolutos dos resíduos padronizados das simulações, $(r_{SD})_{Nm}$, $m = 1, \dots, 25$. Ao final deste passo, obtêm-se três

vetores, de tamanho N , contendo os mínimos, as medianas e os máximos dos resíduos padronizados, em valores absolutos.

- P7. Disponha no gráfico de probabilidade meio-normal dos resíduos padornizados observados, os valores mínimos, medianos e máximos dos dados simulados versus os correspondentes valores esperados em (3.4), conectando os pontos para formar uma linha.

A construção do gráfico envelope é muito simples, uma vez que se saiba gerar variáveis aleatórias correlacionadas (Passo 1). Na literatura, há diversos métodos que atingem esse objetivo. Para distribuições normais correlacionadas existem mais do que um método, um deles encontra-se em (Ripley, 1987), o algoritmo descrito por Park et al. (1996) é útil para gerar distribuições de Bernoulli correlacionadas e o descrito em Park e Shin (1998) para gerar distribuições de Poisson e gama correlacionadas.

No algoritmo proposto em Venezuela (2003), foram implementados os gráficos envelope para as distribuições normal, binomial ($k = 1$), de Poisson e gama correlacionadas, supondo que todas as unidades experimentais tenham sido avaliadas em t ocasiões. Porém, verificou-se a necessidade de avaliação de outras distribuições. O presente trabalho propõe, então, a geração de distribuições binomiais correlacionadas para $k > 1$, não identicamente distribuídas, para que seja possível criar gráficos envelope para MLG estimados por EEG assumindo essa distribuição marginal para a variável resposta. A proposta é apresentada na seção que segue. Já o código completo encontra-se no Apêndice B.2. Nele, além da nova distribuição contemplada para realização do gráfico, incluíram-se os códigos para as outras distribuições presentes em Venezuela (2003), permitindo que o mesmo aceitasse unidades amostrais observadas com número de repetições diferentes.

3.3 Simulando distribuições binomiais correlacionadas

O algoritmo para gerar variáveis aleatórias com distribuição binomial não negativamente correlacionadas é baseado no algoritmo para gerar variáveis com distribuição de Bernoulli correlacionadas descrito por (Park et al., 1996). Para exemplificar, considerou-se o caso bidi-

mensional, embora sirva para mais dimensões. Deseja-se gerar um vetor $\mathbf{Z} = (Z_1, Z_2)$, tal que $Z_j \sim \text{Bernoulli}(p_j)$, $j = 1, 2$ e $\text{Corr}(Z_1, Z_2) = \rho_{12}$.

Tomando variáveis aleatórias independentes com distribuição de Poisson ($X(\lambda) \sim \text{Poisson}(\lambda)$), definem-se U_1 e U_2 por.

$$U_1 = X_1(\lambda_1 - \lambda_{12}) + X_3(\lambda_{12}) \quad \text{e} \quad U_2 = X_2(\lambda_2 - \lambda_{12}) + X_3(\lambda_{12}).$$

Dada a independência de X_1 , X_2 e X_3 , vê-se que U_1 e U_2 têm distribuição de Poisson com médias λ_1 e λ_2 , respectivamente, e são não negativamente correlacionadas por causa do termo X_3 em comum. Define-se também a seguinte variável aleatória

$$Z_j = I_{\{0\}}(U_j),$$

$j = 1, 2$, onde $I_{\{0\}}(\cdot)$ é a função indicadora tal que $I_A(y) = 1$, se $y \in A$ e $I_A(y) = 0$, se $y \notin A$. Escolhendo λ_j de forma que $E[Z_j] = p_j$, tem-se que $\lambda_j = -\ln(p_j)$, pois $E[Z_j] = P(U_j = 0) = e^{-\lambda_j}$. Já λ_{12} deve ser escolhido de modo que $\text{Corr}(Z_1, Z_2) = \rho_{12}$. Como

$$\text{Cov}(Z_1, Z_2) = E[Z_1 Z_2] - E[Z_1]E[Z_2] = P(U_1 = U_2 = 0) - p_1 p_2 = p_1 p_2 (e^{\lambda_{12}} - 1),$$

tem-se

$$\text{Corr}(Z_1, Z_2) = \frac{p_1 p_2 (e^{\lambda_{12}} - 1)}{\sqrt{\text{Var}(Z_1)} \sqrt{\text{Var}(Z_2)}},$$

ou seja,

$$\rho_{12} = \frac{p_1 p_2 (e^{\lambda_{12}} - 1)}{\sqrt{p_1(1 - p_1)} \sqrt{p_2(1 - p_2)}}.$$

Logo,

$$\lambda_{12} = \ln \left\{ 1 + \rho_{12} \sqrt{\frac{(1 - p_1)(1 - p_2)}{p_1 p_2}} \right\}.$$

Venezuela (2003) mostra que o mesmo algoritmo serve também para gerar variáveis com

distribuição de Poisson e gama correlacionadas. Se o algoritmo fosse interrompido nesse instante, obteriam-se variáveis com distribuição de Poisson correlacionadas. Como a ideia se baseia na propriedade de que a soma de variáveis com distribuição de Poisson tem distribuição de Poisson, a mesma estrutura pode ser seguida para gerar variáveis com distribuição gama, já que esta possui a mesma propriedade, como mostra Park e Shin (1998).

O artigo de Prentice (1988) mostra que ρ_{12} não varia livremente em $[0,1]$, pois como $E[Z_1 Z_2] = P(Z_1 = Z_2 = 1) \leq P(Z_j = 1) = p_j$, então

$$\begin{cases} Cov(Z_1, Z_2) \leq p_1(1 - p_2) \\ Cov(Z_1, Z_2) \leq p_2(1 - p_1) \end{cases} \Rightarrow \begin{cases} \rho_{12} \leq \sqrt{\frac{p_1(1-p_2)}{p_2(1-p_1)}} \\ \rho_{12} \leq \sqrt{\frac{p_2(1-p_1)}{p_1(1-p_2)}}. \end{cases}$$

A Figura 3.1 apresenta o limite superior da correlação ρ_{12} em função de valores de p_j . Nota-se que, quanto maior é a diferença entre as probabilidades de sucesso das variáveis Bernoulli, menor é o limite.

Figura 3.1: Limite superior para ρ_{12} .

Assim, a proposta dessa disse para gerar amostras de variáveis aleatórias com distribuição

binomial correlacionadas é dada a seguir. Sejam $\mathbf{Z}_l = (Z_{l1}, Z_{l2}, \dots, Z_{lt})$, $l = 1, \dots, k$, k vetores aleatórios independentes tais que $Z_{lq} \sim \text{Bernoulli}(p_q)$ e $\text{Corr}(Z_{lq}, Z_{lr}) = \rho_{qr}, \forall l$. Definimos $\mathbf{W} = (W_1, \dots, W_t)$, com $W_q = \sum_{l=1}^{k_q} Z_{lq} \sim \text{Binomial}(k_q, p_q)$, sendo $k_q \leq k$ e $q = 1, \dots, t$.

Para $q \neq r$, $q, r = 1, \dots, t$, sendo t é o número de binomiais correlacionadas, tem-se

$$\begin{aligned} \text{Cov}(W_q, W_r) &= \text{Cov}\left(\sum_{l_1=1}^{k_q} Z_{l_1 q}, \sum_{l_2=1}^{k_r} Z_{l_2 r}\right) \\ &= \sum_{l_1=1}^{k_q} \sum_{l_2=1}^{k_r} \text{Cov}(Z_{l_1 q}, Z_{l_2 r}) \\ &= \sum_{l=1}^{\min(k_q, k_r)} \text{Cov}(Z_{lq}, Z_{lr}). \end{aligned}$$

Portanto,

$$\begin{aligned} \text{Corr}(W_q, W_r) &= \frac{\min(k_q, k_r) \text{Cov}(Z_q, Z_r)}{\sqrt{\text{Var}(\sum_{l=1}^{k_q} Z_{lq})} \sqrt{\text{Var}(\sum_{m=1}^{k_r} Z_{mr})}} \\ &= \frac{\min(k_q, k_r)}{\sqrt{k_q} \sqrt{k_r}} \frac{\text{Cov}(Z_q, Z_r)}{\sqrt{\text{Var}(Z_q)} \sqrt{\text{Var}(Z_r)}} \\ &= \frac{\min(k_q, k_r)}{\sqrt{k_q} \sqrt{k_r}} \rho_{qr}. \end{aligned}$$

Como $D(k_q, k_r) = \frac{\min(k_q, k_r)}{\sqrt{k_q} \sqrt{k_r}} \in [0, 1]$, a correlação entre as variáveis W_q e W_r é menor ou igual a ρ_{qr} (essa correlação é exatamente igual a ρ_{qr} quando $k_q = k_r = k$, ou seja, quando $W_q \sim \text{Binomial}(k, p_q)$, $q = 1, \dots, t$). Para que a correlação entre W_q e W_r seja igual a ρ_{qr} quando $k_q < k$ ou $k_r < k$, devem ser construídas variáveis Z_{lq} e Z_{lr} de tal forma que a correlação entre elas seja $\tilde{\rho}_{qr} = \frac{\rho_{qr}}{D(k_q, k_r)}$. Agora, temos a seguinte limitação

$$0 \leq \rho_{qr} \leq \min \left(D(k_q, k_r) \sqrt{\frac{p_q(1-p_r)}{p_r(1-p_q)}}, D(k_q, k_r) \sqrt{\frac{p_r(1-p_q)}{p_q(1-p_r)}} \right).$$

Para exemplificar, deseja-se gerar um vetor aleatório $\mathbf{W} = (W_1, \dots, W_4)$, com $W_1 = \sum_{l=1}^5 Z_{l1} \sim \text{Binomial}(5; 0, 3)$, $W_2 = \sum_{l=1}^4 Z_{l2} \sim \text{Binomial}(4; 0, 4)$, $W_3 = \sum_{l=1}^4 Z_{l3} \sim \text{Binomial}(4; 0, 4)$, $W_4 = \sum_{l=1}^3 Z_{l4} \sim \text{Binomial}(3; 0, 5)$ e matriz de correlação

$$R = \begin{bmatrix} 1 & 0,5 & 0,5 & 0,5 \\ 0,5 & 1 & 0,5 & 0,5 \\ 0,5 & 0,5 & 1 & 0,5 \\ 0,5 & 0,5 & 0,5 & 1 \end{bmatrix}.$$

Como o $\max(k_q) = 5$, $q = 1, \dots, 4$, geram-se os vetores aleatórios independentes $\mathbf{Z}_l = (Z_{l1}, Z_{l2}, Z_{l3}, Z_{l4})$, $l = 1, \dots, 5$, tais que $Z_{l1} \sim \text{Bernoulli}(0, 3)$, $Z_{l2} \sim \text{Bernoulli}(0, 4)$, $Z_{l3} \sim \text{Bernoulli}(0, 4)$, $Z_{l4} \sim \text{Bernoulli}(0, 5)$ e matriz de correlação

$$\tilde{R} = \begin{bmatrix} 1 & \frac{0,5}{D(5,4)} & \frac{0,5}{D(5,4)} & \frac{0,5}{D(5,3)} \\ \frac{0,5}{D(4,5)} & 1 & \frac{0,5}{D(4,4)} & \frac{0,5}{D(4,3)} \\ \frac{0,5}{D(4,5)} & \frac{0,5}{D(4,4)} & 1 & \frac{0,5}{D(4,3)} \\ \frac{0,5}{D(3,5)} & \frac{0,5}{D(3,4)} & \frac{0,5}{D(3,5)} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0,559 & 0,559 & 0,645 \\ 0,559 & 1 & 0,500 & 0,577 \\ 0,559 & 0,500 & 1 & 0,577 \\ 0,645 & 0,577 & 0,577 & 1 \end{bmatrix}.$$

Assim, obtêm-se, por exemplo, as observações $Z_1 = (1, 0, 1, 1)$, $Z_2 = (0, 0, 0, 0)$, $Z_3 = (1, 1, 1, 1)$, $Z_4 = (0, 0, 1, 1)$ e $Z_5 = (0, 0, 0, 1)$, que conduzem a $\mathbf{W} = (2, 1, 3, 2)$. Note que o último elemento de Z_4 e os 3 últimos de Z_5 são desconsiderados na soma, uma vez que W_4 possui tamanho 3 e W_2 e W_3 têm tamanho 4.

Para testar o algoritmo, simulou-se uma amostra de 100 valores de $\mathbf{W} = (W_1, \dots, W_4)$ e replicou-se o procedimento 1000 vezes. Para cada réplica, obtiveram-se a média $\bar{\mathbf{W}}_m$ e a matriz de correlação R_m amostrais. Em seguida, calculou-se a média $\bar{\mathbf{W}} = \sum_{m=1}^{1000} \bar{\mathbf{W}}_m / 1000$, obtendo $\bar{\mathbf{W}} = (1, 5009; 1, 6012; 1, 6015; 1, 5005)$. A Tabela 3.1 apresenta o mínimo, o máximo, a média e o desvio padrão dos 1000 valores de cada elemento de R_m .

Parâmetro	Mínimo	Máximo	Média	Desvio padrão
ρ_{12}	0,137	0,740	0,5005	0,077
ρ_{13}	0,199	0,682	0,5002	0,078
ρ_{14}	0,241	0,722	0,5004	0,072
ρ_{23}	0,144	0,716	0,4996	0,080
ρ_{24}	0,158	0,725	0,4995	0,079
ρ_{34}	0,197	0,688	0,5000	0,079

Tabela 3.1: Estatísticas para os elementos da matriz de correlação.

No estudo de simulação apresentado, considerou-se $E(\mathbf{W}) = (1, 5; 1, 6; 1, 6; 1, 5)$ e um coeficiente de correlação igual a 0,5 entre todos os elementos do vetor $\mathbf{W}$. Os resultados desse estudo mostraram que os valores em $\bar{\mathbf{W}}$ são muito próximos dos valores esperados e que as estatísticas apresentadas na Tabela 3.1 indicam que as correlações de interesse não são viesadas.

A distribuição das correlações simuladas é simétrica, provavelmente normal, a precisão das correlações tendem a aumentar tanto com o aumento do tamanho das binomiais quanto com o número de observações. Para o caso simulado, espera-se que 95% dos valores estejam entre 0,35 e 0,65.

Capítulo 4

Aplicações

Inicialmente, neste capítulo, foram gerados dados de distribuições binomiais marginais correlacionadas, utilizando o algoritmo proposto na Seção 3.3. Foram adotados diferentes valores para o número de repetições (k), probabilidade de sucesso ($\mathbf{p}$) e diferentes estruturas de correlação. Foram avaliados os efeitos desses parâmetros nos gráficos envelope para o ajuste de modelos binomiais com medidas repetidas (Capítulo 2).

Basicamente, espera-se avaliar se o gráfico envelope, além de ser sensível às escolhas das distribuições, também é sensível à escolha da estrutura de correlação.

Em seguida, foram analisados três bancos de dados reais por meio do ajuste de modelos lineares generalizados com medidas repetidas. A validação desses modelos foi realizada segundo as técnicas de diagnóstico propostas no Capítulo 3. No último exemplo, comparou-se o diagnóstico aplicado ao banco de dados completo e incompleto.

4.1 Exemplos simulados

Todos os bancos de dados foram gerados sem dados faltantes, considerando as limitações na estrutura de correlação apresentadas na Seção 3.3.

Nos cinco primeiros bancos de dados foram gerados 100 valores de $t = 4$ distribuições marginais binomiais correlacionadas com $k_{ij} = 3$, $i = 1, \dots, 100$; $j = 1, \dots, 4$ e os seguintes valores de $\mathbf{p}$: $\mathbf{p}_1 = (0, 5; 0, 5; 0, 5; 0, 5)^T$, $\mathbf{p}_2 = (0, 2; 0, 2; 0, 2; 0, 2)^T$, $\mathbf{p}_3 = (0, 4; 0, 45; 0, 5; 0, 6)^T$, $\mathbf{p}_4 = (0, 95; 0, 95; 0, 9; 0, 8)^T$ e $\mathbf{p}_5 = (0, 4; 0, 5; 0, 7; 0, 8)^T$. Nos três primeiros bancos, adotou-se uma estrutura de correlação uniforme com $\alpha = 0, 5$. Já, nos dois últimos bancos, a diferença entre as probabilidades de sucesso limita a correlação permitida para a geração das variáveis (Figura 3.1). Nesses casos, os dados foram gerados com estrutura de correlação uniforme com $\alpha = 0, 4$ e $\alpha = 0, 3$, respectivamente.

Gerados os dados, ajustaram-se dois modelos de regressão binomial com ligação logito, sendo

$$\log \left(\frac{p_{ij}}{1 - p_{ij}} \right) = \beta_1 + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij}$$

com

$$x_{mij} = \begin{cases} 1, & \text{se a observação } i \text{ foi avaliada na repetição } j = m; \\ 0, & \text{caso contrário,} \end{cases}$$

para $i = 1, \dots, 100$, $j = 1, \dots, 4$ e $m = 2, 3, 4$. Em um dos modelos foi adotada a matriz de correlação de trabalho uniforme e, no outro, autorregressiva de ordem 1. Além disso, ajustou-se um modelo de regressão normal com estrutura de correlação de trabalho uniforme e ligação identidade, sendo $\mu_{ij} = k_{ij}p_{ij} = \gamma_1 + \gamma_2 x_{2ij} + \gamma_3 x_{3ij} + \gamma_4 x_{4ij}$ e $x_{mij} = 1$, $m = 2, 3, 4$, como definidos anteriormente.

As Figuras 4.1 a 4.5, apresentam os gráficos envelope construídos a partir dos modelos ajustados aos cinco bancos de dados.

(a) Binomial - uniforme

(b) Binomial - autorregressiva

(c) Normal - uniforme

Figura 4.1: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 3$, $\mathbf{p}_1 = (0, 5; 0, 5; 0, 5; 0, 5)^T$ e estrutura de correlação uniforme ($\alpha = 0, 5$)

(a) Binomial - uniforme

(b) Binomial - autorregressiva

(c) Normal - uniforme

Figura 4.2: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 3$, $\mathbf{p}_2 = (0, 2; 0, 2; 0, 2; 0, 2)^T$ e estrutura de correlação uniforme ($\alpha = 0, 5$)

Figura 4.3: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 3$, $\mathbf{p}_3 = (0, 4; 0, 45; 0, 5; 0, 6)^T$ e estrutura de correlação uniforme ($\alpha = 0, 5$)

Figura 4.4: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 3$, $\mathbf{p}_4 = (0, 95; 0, 95; 0, 9; 0, 8)^T$ e estrutura de correlação uniforme ($\alpha = 0, 4$)

Figura 4.5: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 3$, $\mathbf{p}_5 = (0, 4; 0, 5; 0, 7; 0, 8)^T$ e estrutura de correlação uniforme ($\alpha = 0, 3$)

Nas Figuras 4.1 a 4.5, nota-se que os modelos que consideram a distribuição binomial parecem bem ajustados para as duas estruturas de correlação de trabalho. No entanto, o mesmo não ocorre quando considera-se o modelo de regressão normal, principalmente nas situações em que os valores de p_{ij} estão afastados de 0, 5.

Os gráficos envelope das Figuras 4.6 a 4.10 e Figuras 4.11 a 4.15, foram construídos da mesma forma que os das Figuras 4.1 a 4.5 considerando, no entanto, bancos de dados gerados com $k_{ij} = 10$ e $k_{ij} = 50$, respectivamente. Pelas Figuras 4.6 a 4.15, observa-se, novamente, que os modelos ajustados com matriz de correlação de trabalho uniforme e autorregressiva de ordem 1 se ajustam bem aos dados. Já, para $k_{ij} = 10$, o modelo normal se apresenta melhor quando p_{ij} se aproxima de 0, 5 e, quando $k_{ij} = 50$, não apresenta bom ajuste para p_{ij} próximo de zero ou de um (Figura 4.14).

Figura 4.6: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_1 = (0, 5; 0, 5; 0, 5; 0, 5; 0, 5)^T$ e estrutura de correlação uniforme ($\alpha = 0, 5$)

Figura 4.7: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_2 = (0, 2; 0, 2; 0, 2; 0, 2; 0, 2)^T$ e estrutura de correlação uniforme ($\alpha = 0, 5$)

(a) Binomial - uniforme

(b) Binomial - autorregressiva

(c) Normal - uniforme

Figura 4.8: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_3 = (0, 4; 0, 45; 0, 5; 0, 6)^T$ e estrutura de correlação uniforme ($\alpha = 0, 5$)

(a) Binomial - uniforme

(b) Binomial - autorregressiva

(c) Normal - uniforme

Figura 4.9: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_4 = (0, 95; 0, 95; 0, 9; 0, 8)^T$ e estrutura de correlação uniforme ($\alpha = 0, 4$)

Figura 4.10: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_5 = (0, 4; 0, 5; 0, 7; 0, 8)^T$ e estrutura de correlação uniforme ($\alpha = 0, 3$)

Figura 4.11: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 50$, $\mathbf{p}_1 = (0, 5; 0, 5; 0, 5; 0, 5)^T$ e estrutura de correlação uniforme ($\alpha = 0, 5$)

Figura 4.12: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 50$, $\mathbf{p}_2 = (0, 2; 0, 2; 0, 2; 0, 2)^T$ e estrutura de correlação uniforme ($\alpha = 0, 5$)

Figura 4.13: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 50$, $\mathbf{p}_3 = (0, 4; 0, 45; 0, 5; 0, 6)^T$ e estrutura de correlação uniforme ($\alpha = 0, 5$)

Figura 4.14: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 50$, $\mathbf{p}_4 = (0, 95; 0, 95; 0, 9; 0, 8)^T$ e estrutura de correlação uniforme ($\alpha = 0, 4$)

Figura 4.15: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 50$, $\mathbf{p}_5 = (0, 4; 0, 5; 0, 7; 0, 8)^T$ e estrutura de correlação uniforme ($\alpha = 0, 3$)

As Figuras 4.16 a 4.20 apresentam gráficos envelope construídos da mesma forma que os anteriores, sendo que todos os bancos de dados foram gerados com $k_{ij} = 10$ e estrutura de correlação autorregressiva de ordem 1 com $\alpha = 0, 8$, para $\mathbf{p}_1$, $\mathbf{p}_2$ e $\mathbf{p}_3$ e $\alpha = 0, 5$, para $\mathbf{p}_4$ e

$\mathbf{p}_5$. Gráficos semelhantes, não exibidos neste trabalho, foram construídos com $k_{ij} = 3$ e 50. Nota-se que os envelopes obtidos exibem o mesmo padrão observado quando os dados são gerados com estrutura de correlação uniforme.

É importante salientar que os gráficos envelope não parecem auxiliar na escolha da melhor estrutura de correlação de trabalho. Para auxiliar na escolha dessa estrutura de correlação, Johnston (1996) sugere observar o quanto o erro padrão *naive* está próximo ao *robusto*. Nos casos simulados acima, de fato, a escolha da matriz correta proporcionou resultados do erro padrão *naive* ligeiramente mais próximos ao *robusto*.

Figura 4.16: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_1 = (0, 5; 0, 5; 0, 5; 0, 5; 0, 5)^T$ e estrutura de correlação autorregressiva ($\alpha = 0, 8$)

(a) Binomial - autorregressiva

(b) Binomial - uniforme

(c) Normal - autorregressiva

Figura 4.17: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_2 = (0, 2; 0, 2; 0, 2; 0, 2)^T$ e estrutura de correlação autorregressiva ($\alpha = 0, 8$)

(a) Binomial - autorregressiva

(b) Binomial - uniforme

(c) Normal - autorregressiva

Figura 4.18: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_3 = (0, 4; 0, 45; 0, 5; 0, 6)^T$ e estrutura de correlação autorregressiva ($\alpha = 0, 8$)

Figura 4.19: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_4 = (0, 95; 0, 95; 0, 9; 0, 8)^T$ e estrutura de correlação autorregressiva ($\alpha = 0, 5$)

Figura 4.20: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $k = 10$, $\mathbf{p}_5 = (0, 4; 0, 5; 0, 7; 0, 8)^T$ e estrutura de correlação autorregressiva ($\alpha = 0, 5$)

Por fim, simularam-se mais três bancos de dados de tamanhos $n = 100, 50$ e 20 , desta vez fixando-se $t = 3$, $\mathbf{k} = (7, 10, 15)^T$, $\mathbf{p}_6 = (0, 7; 0, 5; 0, 4)^T$ e estrutura de correlação autorregressiva de ordem 1 com $\alpha = 0, 5$.

Gerados os dados, ajustaram-se dois modelos de regressão binomial com ligação logito, sendo

$$\log \left(\frac{p_{ij}}{1 - p_{ij}} \right) = \beta_1 + \beta_2 x_{2ij} + \beta_3 x_{3ij},$$

com

$$x_{mij} = \begin{cases} 1, & \text{se a observação } i \text{ foi avaliada na repetição } j = m; \\ 0, & \text{caso contrário,} \end{cases}$$

para $i = 1, \dots, n$, $j = 1, \dots, 3$ e $m = 2, 3$. Em um dos modelos foi adotada a matriz de correlação de trabalho uniforme e, no outro, autorregressiva de ordem 1.

Ao variar os valores de k_{ij} , o modelo com distribuição marginal normal não faz sentido, e portanto, não foi ajustado. Diante disso, para os exemplos que seguem, a comparação limitou-se às estruturas de correlação para modelos com distribuição binomial. As Figuras 4.21 a 4.23 exibem os gráficos envelope para os modelos ajustados. Da mesma forma que nas figuras anteriores, observa-se que mesmo variando o valor de k_{ij} , os modelos ajustados com as duas estruturas de correlação de trabalho encontram-se bem ajustados, independente do tamanho amostral.

(a) Estrutura certa - autorregressiva

(b) Estrutura uniforme

Figura 4.21: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 100$, $\mathbf{k} = (7, 10, 15)$, $\mathbf{p}_6 = (0, 7; 0, 5; 0, 4)$ e estrutura de correlação autorregressiva ($\alpha = 0, 5$)

(a) Estrutura certa - autorregressiva

(b) Estrutura uniforme

Figura 4.22: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 50$, $\mathbf{k} = (7, 10, 15)$, $\mathbf{p}_6 = (0, 7; 0, 5; 0, 4)$ e estrutura de correlação autorregressiva ($\alpha = 0, 5$)

(a) Estrutura certa - autorregressiva

(b) Estrutura uniforme

Figura 4.23: Gráficos envelope para os modelos ajustados ao banco de dados gerado com $n = 20$, $\mathbf{k} = (7, 10, 15)$, $\mathbf{p}_6 = (0, 7; 0, 5; 0, 4)$ e estrutura de correlação autorregressiva ($\alpha = 0, 5$)

4.2 Exemplo 1 - Pronúncia da língua portuguesa

Em Botter et al. (2008), analisou-se um banco de dados envolvendo 56 brasileiros, de 3 faixas etárias (3 anos, 6 anos e idade adulta). Cada pessoa assistia a uma apresentação de *slides*, sendo que em cada um havia duas palavras lado a lado. Em cada par, as duas palavras possuíam a primeira sílaba idênticas. Uma das palavras do par era oxítone (Frac-forte), ou seja, a primeira sílaba era seguida por outra mais forte; e a outra era paroxítone (Forte-fraca). Foram utilizados quatro pares de palavra. Cada palavra era recortada de 3 formas diferentes, denominadas *gates* (G2, G3 ou G4), sendo que quanto maior o *gate*, mais se escutava do início da palavra.

Em cada *slide*, para cada *gate*, era pronunciada a primeira sílaba referente a uma das palavras e o entrevistado deveria adivinhar qual das duas era a palavra sendo pronunciada.

Com o objetivo de aplicarem-se os métodos de diagnóstico desenvolvidos no Capítulo 3 a dados reais, utilizou-se apenas a parte dos dados analisados em Botter et al. (2008) referente à entonação do tipo Fraca-forte. Como o mesmo indivíduo analisava cada palavra 2 vezes e, em

alguns casos, o indivíduo não respondeu em pelo menos uma das vezes, o número de palavras que cada indivíduo tentou adivinhar foi no máximo 8 em cada *gate*.

Idade	<i>Gate</i>					
	G2		G3		G4	
	Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
3 ANOS	0,45	0,26	0,54	0,25	0,55	0,23
6 ANOS	0,34	0,24	0,37	0,17	0,37	0,18
ADULTOS	0,50	0,17	0,54	0,20	0,69	0,26

Tabela 4.1: Média e desvio padrão da proporção de acertos por *gate* e idade - Exemplo 1, entonação Fraca-forte.

A Tabela 4.1 mostra as médias e desvios padrões das proporções observadas de acertos em cada repetição (*gate*) por grupo de idade, para essa entonação. Então, ajustou-se o modelo de regressão binomial com ligação logito para esse dados, sendo

$$\log \left(\frac{p_{ijk}}{1 - p_{ijk}} \right) = \beta_0 + \beta_1 x_{1ijk} + \beta_2 x_{2ijk} + \beta_3 x_{3ijk} + \beta_4 x_{4ijk},$$

com

$$x_{1ijk} = \begin{cases} 1, & \text{se o } i\text{-ésimo indivíduo submetido ao } j\text{-ésimo } gate \text{ pertencer} \\ & \text{à faixa etária 6 anos } (k = 2); \\ 0, & \text{caso contrário;} \end{cases}$$

$$x_{2ijk} = \begin{cases} 1, & \text{se o } i\text{-ésimo indivíduo submetido ao } j\text{-ésimo } gate \text{ pertencer} \\ & \text{à faixa etária adulta } (k = 3); \\ 0, & \text{caso contrário;} \end{cases}$$

$$x_{3ijk} = \begin{cases} 1, & \text{se o } i\text{-ésimo indivíduo pertencente à } k\text{-ésima faixa etária foi submetido} \\ & \text{ao } gate \text{ 3 } (j = 2); \\ 0, & \text{caso contrário;} \end{cases}$$

$$x_{4ijk} = \begin{cases} 1, & \text{se o } i\text{-ésimo indivíduo pertencente à } k\text{-ésima faixa etária foi submetido} \\ & \text{ao } gate\ 4\ (j = 3); \\ 0, & \text{caso contrário;} \end{cases}$$

e p_{ijk} , a probabilidade de acerto do indivíduo i pertencente à k -ésima faixa etária e submetido ao j -ésimo *gate*, $k = 1, 2, 3$, $j = 1, 2, 3$ e $i = 1, \dots, n_k$.

Para escolher a estrutura de correlação de trabalho, observa-se a matriz de correlação amostral ($\mathbf{R}$) dada por

$$\mathbf{R} = \begin{bmatrix} 1 & 0,566 & 0,525 \\ 0,566 & 1 & 0,617 \\ 0,525 & 0,617 & 1 \end{bmatrix}.$$

Para o ajuste do modelo, testou-se diversas estruturas de correlação (uniforme, autorregressiva e não estruturada). Os resultados foram similares, e optou-se pela não estruturada, por apresentar, erro padrão *naive* mais próximo ao *robusto* e gráficos envelope mais satisfatórios. A Tabela 4.2 mostra as estimativas dos parâmetros do modelo ajustado.

Parâmetro	Estimativa	E.P. Naive	z(Naive)	E.P. Robusto	z(Robusto)
β_0	-0,21	0,15	-1,37	0,19	-1.11
β_1	-0,50	0,20	-2,51	0,21	-2,43
β_2	0,30	0,19	1,57	0,23	1,33
β_3	0,18	0,11	1,60	-0,12	1,46
β_4	0,45	0,12	3,89	0,14	3,22
α_{12}	0,42				
α_{13}	0,37				
α_{23}	0,49				

Tabela 4.2: Estimativas e erros padrões (E.P.) dos parâmetros do modelo de regressao binomial com estrutura de correlação de trabalho não estruturada - Exemplo 1

As estimativas apontam efeito de idade e *gate*. Crianças com 6 anos de idade, apresentaram

resultados menos positivos, em relação as crianças de 3 anos e o *gate* 4 é aparentemente mais fácil do que os demais.

A Figura 4.24 apresenta a análise de diagnóstico do modelo ajustado. Não se observam pontos de alavanca. Quanto aos pontos influentes e aberrantes, porém, destacam-se as observações do indivíduo 18. É ele que apresenta a maior distância de Cook e o menor resíduo padronizado dentre as unidades experimentais.

Figura 4.24: Diagnóstico - Pontos alavanca, influentes e aberrantes - Exemplo 1

Portanto, ajustou-se novamente o modelo, retirando-se o indivíduo em questão. A Tabela 4.3 mostra a variação ocorrida nos coeficientes ao retirarem-se as observações desse indivíduo.

Parâmetro	Estimativa		Variação
	Dados completos	Sem o indivíduo 18	
β_0	-0,21	-0,12	42%
β_1	-0,50	-0,58	-15%
β_2	0,30	0,20	33%
β_3	0,18	0,19	-8%
β_4	0,45	0,43	5%

Tabela 4.3: Variação nas estimativas dos parâmetros dos modelos ajustados com dados completos e retirando a unidade experimental 18 - Exemplo 1.

A exclusão dessas observações promove variações altas, de até 42% nos parâmetros. A Figura 4.25, por sua vez, mostra o gráfico envelope simulado para os ajustes dos dois modelos, com e sem o indivíduo 18, respectivamente. Observa-se que o modelo se ajusta melhor sem as observações desse indivíduo.

Figura 4.25: Gráfico de probabilidade meio-normal com envelope simulado - Exemplo 1

4.3 Exemplo 2 - Memória operacional

O trabalho realizado por Elia et al. (2006) investigou, entre outros objetivos, o desempenho de crianças entre 3 e 6 anos no teste de *Memória Operacional* a partir de repetição de não-palavras (itens que possuem uma sequência fonológica e não remetem a nenhum significado na língua portuguesa). Foram apresentadas oralmente 40 palavras a cada criança, sendo 10 monossílabas, 10 dissílabas, 10 trissílabas e 10 polissílabas, sendo que a criança deveria repeti-las imediatamente após a apresentação. Assim, temos como variável resposta o número de acertos de cada criança em cada nível de dificuldade.

O Banco de dados contém 136 crianças divididas igualmente entre quatro grupos de idade (3, 4, 5 e 6 anos completos). O Gráfico 4.26 sugere uma queda no número de acertos com o aumento do número de sílabas e que essa queda é menos acentuada quando a idade é maior.

Figura 4.26: Média de acertos por idade e número de sílabas - Exemplo 2.

Em Eliau et al. (2006), foi ajustado, para cada grupo de idade, um modelo de curva de crescimento (Graybill, 1976). Em todos os modelos confirmou-se o decréscimo do número médio de acertos com o aumento do número de sílabas.

Para o ajuste de um MLG para medidas repetidas com distribuição binomial, considerou-se o número de sílabas como variável categorizada, diferentemente dos modelos de curva de crescimento, em que o número de sílabas foi considerado como uma variável contínua. Ajustou-se o seguinte modelo

$$\log \left(\frac{p_{ijk}}{1 - p_{ijk}} \right) = \beta_0 + \beta_1 x_{1ijk} + \beta_2 x_{2ijk} + \beta_3 x_{3ijk} + \beta_4 x_{4ijk} + \beta_5 x_{5ijk} + \beta_6 x_{6ijk},$$

sendo p_{ijk} a proporção de acertos do i -ésimo indivíduo, de idade k , na j -ésima repetição (Sílabas),

$$x_{1ijk} = \begin{cases} 1, & \text{se o } i\text{-ésimo indivíduo tiver 4 anos de idade } (k = 2), \forall j; \\ 0, & \text{caso contrário;} \end{cases}$$

$$x_{2ijk} = \begin{cases} 1, & \text{se o } i\text{-ésimo indivíduo tiver 5 anos de idade } (k = 3), \forall j; \\ 0, & \text{caso contrário;} \end{cases}$$

$$x_{3ijk} = \begin{cases} 1, & \text{se o } i\text{-ésimo indivíduo tiver 6 anos de idade } (k = 4), \forall j; \\ 0, & \text{caso contrário;} \end{cases}$$

$$x_{4ijk} = \begin{cases} 1, & \text{se o } i\text{-ésimo indivíduo, de idade } k, \text{ responde à palavras dissílabas } (j = 2); \\ 0, & \text{caso contrário;} \end{cases}$$

$$x_{5ijk} = \begin{cases} 1, & \text{se o } i\text{-ésimo indivíduo, de idade } k, \text{ responde à palavras trissílabas } (j = 3); \\ 0, & \text{caso contrário;} \end{cases}$$

$$x_{6ijk} = \begin{cases} 1, & \text{se o } i\text{-ésimo indivíduo, de idade } k, \text{ responde à palavras polissílabas } (j = 4); \\ 0, & \text{caso contrário;} \end{cases}$$

Propôs-se como estrutura de correlação de trabalho uma matriz autorregressiva de ordem 1, com base baseada na matriz de correlação observada $\mathbf{R}$.

$$\mathbf{R} = \begin{bmatrix} 1 & 0,453 & 0,153 & 0,046 \\ 0,453 & 1 & 0,54 & 0,289 \\ 0,153 & 0,54 & 1 & 0,493 \\ 0,046 & 0,289 & 0,493 & 1 \end{bmatrix}$$

Os coeficientes do modelo ajustado e correspondentes erros padrões seguem na (Tabela 4.4).

Parâmetro	Estimativa	E.P. Naive	z(Naive)	E.P. Robusto	z(Robusto)
β_0	3,69	0,31	16,48	0,22	16,47
β_1	-0,14	0,18	-1,04	0,21	-,67
β_2	0,87	0,21	3,21	0,27	5,62
β_3	1,54	0,26	8,33	0,31	4,93
β_4	-1,43	0,28	-7,21	0,19	-7,60
β_5	-2,34	0,30	-10,82	0,22	-10,46
β_6	-3,12	0,31	-14,03	0,23	-13,25
α	0,35				

Tabela 4.4: Estimativas e erros padrões (E.P.) dos parâmetros do modelo de regressão binomial com estrutura de correlação de trabalho autorregressiva de ordem 1 - Exemplo 2.

O Gráfico 4.27 mostra as médias estimadas pelo modelo ajustado.

Figura 4.27: Média de acertos por idade e número de sílabas estimado - Exemplo 2.

Os gráficos de diagnóstico (Figura 4.28) e o envelope (Figura 4.29) mostram que o modelo não está bem ajustado. O gráfico de resíduos apresenta valores extremos e o gráfico envelope simulado cerca de 40 % dos pontos fora da banda de confiança.

Figura 4.28: Diagnóstico - Pontos alavanca, influentes e aberrantes - Exemplo 2.

Figura 4.29: Gráfico de probabilidade meio-normal com envelope simulado - Exemplo 2.

Analisando as matrizes de correlações observadas em cada grupo (Tabela 4.5), nota-se que as mesmas diferem bastante. Isto pode ter sido determinante para a falta de ajuste apresentada na Figura 4.29.

3 anos					4 anos				
	1	2	3	4		1	2	3	4
1	1	0,337	0,25	-0,203	1	1	0,479	0,094	0,073
2	0,337	1	0,674	0,01	2	0,479	1	0,479	0,243
3	0,25	0,674	1	0,261	3	0,094	0,479	1	0,453
4	-0,203	0,01	0,261	1	4	0,073	0,243	0,453	1
5 anos					6 anos				
	1	2	3	4		1	2	3	4
1	1	0,691	-0,025	-0,242	1	1	0,056	-0,116	0,087
2	0,691	1	0,171	-0,061	2	0,056	1	0,181	0,54
3	-0,025	0,171	1	0,432	3	-0,116	0,181	1	0,494
4	-0,242	-0,061	0,432	1	4	0,087	0,54	0,494	1

Tabela 4.5: Matrizes de correlação observadas, por grupo de idade - Exemplo 2.

Assim, ajustou-se um modelo para cada idade, ainda supondo que o número de sílabas seja uma variável categorizada, mas com diferentes estruturas de correlação de trabalho. Para 3 anos de idade, o melhor ajuste foi dado por uma matriz fixa R_3 , sendo R_3 dada por

$$R_3 = \begin{bmatrix} 1 & 0,25 & 0,25 & 0,0 \\ 0,25 & 1 & 0,25 & 0,25 \\ 0,25 & 0,25 & 1 & 0,25 \\ 0,0 & 0,25 & 0,25 & 1 \end{bmatrix}.$$

Para 4 anos de idade, considerou-se uma estrutura de correlação de dependência estacionária de ordem 1, $\hat{\alpha} = 0,425$. Para 5 e 6 anos, consideraram-se estruturas autorregressivas de ordem 1, com $\hat{\alpha} = 0,384$ e $\hat{\alpha} = 0,294$, respectivamente.

Os gráficos envelope dos modelos ajustados encontram-se nas Figuras 4.30 e 4.31. Todos os modelos ajustaram-se de forma satisfatória aos dados.

Figura 4.30: Gráfico de probabilidade meio-normal com envelope simulado - Exemplo 2 - 3 e 4 anos

Figura 4.31: Gráfico de probabilidade meio-normal com envelope simulado - Exemplo 2 - 5 e 6 anos

4.4 Exemplo 3 - Quimioterapia

O último exemplo aplicado tem o intuito de testar os algoritmos desenvolvidos neste trabalho para o caso da distribuição de Poisson com dados faltantes. Para tanto, utilizou-se um exemplo de Montgomery et al. (2001), que tinha como objetivo comparar três drogas utilizadas em quimioterapia. Cada droga foi aplicada em 10 ratos com leucemia e após o tratamento foi observado o número de glóbulos brancos (NGB), o número de glóbulos vermelhos (NGV) e o número de células cancerígenas (NCC), em quatro momentos distintos.

Esse mesmo exemplo foi utilizado por Venezuela (2003), pois o banco de dados não apresentava observações faltantes. Com o objetivo de comparar os resultados obtidos pelo algoritmo implementado em Venezuela (2003) e o que foi desenvolvido no presente trabalho, retiraram-se aleatoriamente 15 observações da variável resposta (NCC) do banco de dados original.

O modelo ajustado para o NCC, considera a distribuição de Poisson correlacionada com ligação log, da seguinte forma

$$\log(\mu_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \beta_5 x_{5ij}, \quad (4.1)$$

sendo

$$x_{1ij} = \begin{cases} 1, & \text{se o } i\text{-ésimo rato recebeu a droga 1, } \forall j; \\ 0, & \text{caso contrário;} \end{cases}$$

$$x_{2ij} = \begin{cases} 1, & \text{se o } i\text{-ésimo rato recebeu a droga 2, } \forall j; \\ 0, & \text{caso contrário;} \end{cases}$$

$$x_{3ij} = \begin{cases} 1, & \text{se o } i\text{-ésimo rato recebeu a droga 3, } \forall j; \\ 0, & \text{caso contrário;} \end{cases}$$

x_{4ij} = o número de glóbulos brancos do i -ésimo rato no j -ésimo tempo;

x_{5ij} = o número de glóbulos vermelhos do i -ésimo rato no j -ésimo tempo;

para $i = 1, \dots, 30$ e $j = 1, \dots, 4$.

Além disso, propõe-se uma matriz de correlação de trabalho autorregressiva de ordem 1. As Figuras 4.33 e 4.34 mostram as medidas de diagnóstico apresentadas na Seção 3.1 para bancos com dados completos e faltantes, e a Figura 4.32 mostra os gráficos envelope, também para os respectivos bancos de dados. Nas duas situações consideradas, os modelos parecem bem ajustados.

Figura 4.32: Gráfico de probabilidade meio-normal com envelope simulado - Exemplo 3

Figura 4.33: Diagnóstico com dados completos - Pontos de alavanca, influentes e aberrantes - Exemplo 3 - Dados completos

Figura 4.34: Diagnóstico com dados faltantes - Pontos de alavanca, influentes e aberrantes - Exemplo 3 - 15 observações faltantes

Capítulo 5

Conclusão

Venezuela et al. (2007) desenvolveram técnicas de diagnóstico para Modelos Lineares Generalizados com medidas repetidas estimados por Equações de Estimação Generalizadas, enfocando as distribuições normal com ligação identidade, gama com ligação inversa, Poisson com ligação log e Bernoulli com ligação logito. Neste trabalho agregaram-se em um único código no *software* R as medidas que haviam sido desenvolvidas em Venezuela et al. (2007) estendendo as mesmas para bancos de dados com observações faltantes. As mesmas medidas foram implementadas para distribuições binomiais correlacionadas considerando a função de ligação logito. Para esse fim, desenvolveu-se um método de geração de distribuições binomiais não identicamente distribuídas correlacionadas positivamente.

Para estudos futuros, pretende-se adequar os algoritmos desenvolvidos neste trabalho para a criação de uma biblioteca no *software* para o R. Deve-se, então, generalizar o código para que qualquer função de ligação possa ser utilizada e investir em outros métodos para geração de variáveis binomiais, que aceitem correlações negativas, por exemplo.

Outra pesquisa a ser realizada é a extensão do trabalho de Silva (2009), que considera medida repetida em dois fatores cruzados, para a distribuição marginal binomial. Dessa forma será possível analisar o exemplo da Seção 4.2, em que havia dois fatores de repetição na mesma unidade experimental (entonação e *gate*), de forma mais adequada.

Apêndice A

Conjuntos de dados

A.1 Conjunto de dados do Exemplo 1

id	Entonacao	Gate	Sucesso	Fracasso	Idade
1	SW	G2	5	3	3 ANOS
1	WS	G2	3	5	3 ANOS
1	SW	G3	6	2	3 ANOS
1	WS	G3	2	6	3 ANOS
1	SW	G4	6	1	3 ANOS
1	WS	G4	2	6	3 ANOS
2	SW	G2	3	1	3 ANOS
2	WS	G2	0	4	3 ANOS
2	SW	G3	4	0	3 ANOS
2	WS	G3	1	3	3 ANOS
2	SW	G4	4	0	3 ANOS
2	WS	G4	1	3	3 ANOS
3	SW	G2	7	1	3 ANOS
3	WS	G2	4	4	3 ANOS

id	Entonacao	Gate	Sucesso	Fracasso	Idade
3	SW	G3	2	6	3 ANOS
3	WS	G3	4	4	3 ANOS
3	SW	G4	4	3	3 ANOS
3	WS	G4	5	3	3 ANOS
4	SW	G2	0	2	3 ANOS
4	WS	G2	2	0	3 ANOS
4	SW	G3	1	1	3 ANOS
4	WS	G3	2	0	3 ANOS
4	SW	G4	0	2	3 ANOS
4	WS	G4	2	0	3 ANOS
5	SW	G2	5	1	3 ANOS
5	WS	G2	2	4	3 ANOS
5	SW	G3	3	3	3 ANOS
5	WS	G3	2	4	3 ANOS
5	SW	G4	3	2	3 ANOS
5	WS	G4	5	1	3 ANOS
6	SW	G2	2	2	3 ANOS
6	WS	G2	2	2	3 ANOS
6	SW	G3	3	1	3 ANOS
6	WS	G3	2	2	3 ANOS
6	SW	G4	1	2	3 ANOS
6	WS	G4	3	1	3 ANOS
7	SW	G2	3	5	3 ANOS
7	WS	G2	7	1	3 ANOS
7	SW	G3	2	6	3 ANOS
7	WS	G3	5	3	3 ANOS
7	SW	G4	0	7	3 ANOS
7	WS	G4	6	2	3 ANOS
8	SW	G2	3	5	3 ANOS

id	Entonacao	Gate	Sucesso	Fracasso	Idade
8	WS	G2	5	3	3 ANOS
8	SW	G3	3	5	3 ANOS
8	WS	G3	5	3	3 ANOS
8	SW	G4	4	3	3 ANOS
8	WS	G4	2	6	3 ANOS
9	SW	G2	4	2	3 ANOS
9	WS	G2	1	5	3 ANOS
9	SW	G3	5	1	3 ANOS
9	WS	G3	3	3	3 ANOS
9	SW	G4	3	2	3 ANOS
9	WS	G4	3	3	3 ANOS
10	SW	G2	2	4	3 ANOS
10	WS	G2	3	3	3 ANOS
10	SW	G3	2	4	3 ANOS
10	WS	G3	4	2	3 ANOS
10	SW	G4	4	2	3 ANOS
10	WS	G4	5	1	3 ANOS
11	SW	G2	1	3	3 ANOS
11	WS	G2	2	2	3 ANOS
11	SW	G3	2	2	3 ANOS
11	WS	G3	3	1	3 ANOS
11	SW	G4	0	4	3 ANOS
11	WS	G4	1	3	3 ANOS
12	SW	G2	3	3	3 ANOS
12	WS	G2	2	4	3 ANOS
12	SW	G3	5	1	3 ANOS
12	WS	G3	5	1	3 ANOS
12	SW	G4	4	1	3 ANOS
12	WS	G4	2	4	3 ANOS

id	Entonacao	Gate	Sucesso	Fracasso	Idade
13	SW	G2	7	1	3 ANOS
13	WS	G2	1	7	3 ANOS
13	SW	G3	7	1	3 ANOS
13	WS	G3	4	4	3 ANOS
13	SW	G4	6	1	3 ANOS
13	WS	G4	3	5	3 ANOS
14	SW	G2	4	4	3 ANOS
14	WS	G2	6	2	3 ANOS
14	SW	G3	3	5	3 ANOS
14	WS	G3	7	1	3 ANOS
14	SW	G4	3	4	3 ANOS
14	WS	G4	5	3	3 ANOS
15	SW	G2	5	3	3 ANOS
15	WS	G2	2	6	3 ANOS
15	SW	G3	6	2	3 ANOS
15	WS	G3	4	4	3 ANOS
15	SW	G4	4	3	3 ANOS
15	WS	G4	5	3	3 ANOS
16	SW	G2	4	4	3 ANOS
16	WS	G2	5	3	3 ANOS
16	SW	G3	4	4	3 ANOS
16	WS	G3	5	3	3 ANOS
16	SW	G4	6	1	3 ANOS
16	WS	G4	6	2	3 ANOS
17	SW	G2	6	2	3 ANOS
17	WS	G2	3	5	3 ANOS
17	SW	G3	6	2	3 ANOS
17	WS	G3	2	6	3 ANOS
17	SW	G4	5	2	3 ANOS

id	Entonacao	Gate	Sucesso	Fracasso	Idade
17	WS	G4	4	4	3 ANOS
18	SW	G2	8	0	3 ANOS
18	WS	G2	1	7	3 ANOS
18	SW	G3	8	0	3 ANOS
18	WS	G3	0	8	3 ANOS
18	SW	G4	7	0	3 ANOS
18	WS	G4	4	4	3 ANOS
19	SW	G2	4	4	3 ANOS
19	WS	G2	5	3	3 ANOS
19	SW	G3	6	2	3 ANOS
19	WS	G3	5	3	3 ANOS
19	SW	G4	5	2	3 ANOS
19	WS	G4	4	4	3 ANOS
20	SW	G2	3	5	6 ANOS
20	WS	G2	3	5	6 ANOS
20	SW	G3	5	3	6 ANOS
20	WS	G3	2	6	6 ANOS
20	SW	G4	3	4	6 ANOS
20	WS	G4	2	6	6 ANOS
21	SW	G2	5	1	6 ANOS
21	WS	G2	0	6	6 ANOS
21	SW	G3	6	0	6 ANOS
21	WS	G3	1	5	6 ANOS
21	SW	G4	6	0	6 ANOS
21	WS	G4	1	5	6 ANOS
22	SW	G2	4	4	6 ANOS
22	WS	G2	4	4	6 ANOS
22	SW	G3	5	3	6 ANOS
22	WS	G3	4	4	6 ANOS

id	Entonacao	Gate	Sucesso	Fracasso	Idade
22	SW	G4	4	3	6 ANOS
22	WS	G4	5	3	6 ANOS
23	SW	G2	3	5	6 ANOS
23	WS	G2	2	6	6 ANOS
23	SW	G3	4	4	6 ANOS
23	WS	G3	2	6	6 ANOS
23	SW	G4	2	5	6 ANOS
23	WS	G4	4	4	6 ANOS
24	SW	G2	6	2	6 ANOS
24	WS	G2	1	7	6 ANOS
24	SW	G3	7	1	6 ANOS
24	WS	G3	3	5	6 ANOS
24	SW	G4	7	0	6 ANOS
24	WS	G4	3	5	6 ANOS
25	SW	G2	3	1	6 ANOS
25	WS	G2	0	4	6 ANOS
25	SW	G3	4	0	6 ANOS
25	WS	G3	1	3	6 ANOS
25	SW	G4	3	0	6 ANOS
25	WS	G4	1	3	6 ANOS
26	SW	G2	3	5	6 ANOS
26	WS	G2	3	5	6 ANOS
26	SW	G3	3	5	6 ANOS
26	WS	G3	3	5	6 ANOS
26	SW	G4	2	5	6 ANOS
26	WS	G4	3	5	6 ANOS
27	SW	G2	2	6	6 ANOS
27	WS	G2	3	5	6 ANOS
27	SW	G3	5	3	6 ANOS

id	Entonacao	Gate	Sucesso	Fracasso	Idade
27	WS	G3	3	5	6 ANOS
27	SW	G4	3	4	6 ANOS
27	WS	G4	3	5	6 ANOS
28	SW	G2	5	3	6 ANOS
28	WS	G2	5	3	6 ANOS
28	SW	G3	8	0	6 ANOS
28	WS	G3	4	4	6 ANOS
28	SW	G4	7	0	6 ANOS
28	WS	G4	4	4	6 ANOS
29	SW	G2	2	6	6 ANOS
29	WS	G2	3	5	6 ANOS
29	SW	G3	3	5	6 ANOS
29	WS	G3	3	5	6 ANOS
29	SW	G4	3	4	6 ANOS
29	WS	G4	3	5	6 ANOS
30	SW	G2	5	3	6 ANOS
30	WS	G2	5	3	6 ANOS
30	SW	G3	5	3	6 ANOS
30	WS	G3	3	5	6 ANOS
30	SW	G4	6	1	6 ANOS
30	WS	G4	3	5	6 ANOS
31	SW	G2	3	3	6 ANOS
31	WS	G2	0	6	6 ANOS
31	SW	G3	5	1	6 ANOS
31	WS	G3	3	3	6 ANOS
31	SW	G4	6	0	6 ANOS
31	WS	G4	3	3	6 ANOS
32	SW	G2	5	3	6 ANOS
32	WS	G2	2	6	6 ANOS

id	Entonacao	Gate	Sucesso	Fracasso	Idade
32	SW	G3	4	4	6 ANOS
32	WS	G3	5	3	6 ANOS
32	SW	G4	2	5	6 ANOS
32	WS	G4	5	3	6 ANOS
33	SW	G2	3	3	6 ANOS
33	WS	G2	3	3	6 ANOS
33	SW	G3	4	2	6 ANOS
33	WS	G3	2	4	6 ANOS
33	SW	G4	5	1	6 ANOS
33	WS	G4	2	4	6 ANOS
34	SW	G2	2	6	6 ANOS
34	WS	G2	5	3	6 ANOS
34	SW	G3	4	4	6 ANOS
34	WS	G3	3	5	6 ANOS
34	SW	G4	5	2	6 ANOS
34	WS	G4	3	5	6 ANOS
35	SW	G2	5	3	6 ANOS
35	WS	G2	2	6	6 ANOS
35	SW	G3	7	1	6 ANOS
35	WS	G3	2	6	6 ANOS
35	SW	G4	6	1	6 ANOS
35	WS	G4	1	7	6 ANOS
36	SW	G2	3	1	6 ANOS
36	WS	G2	2	2	6 ANOS
36	SW	G3	3	1	6 ANOS
36	WS	G3	2	2	6 ANOS
36	SW	G4	3	0	6 ANOS
36	WS	G4	1	3	6 ANOS
37	SW	G2	0	2	6 ANOS

id	Entonacao	Gate	Sucesso	Fracasso	Idade
37	WS	G2	0	2	6 ANOS
37	SW	G3	2	0	6 ANOS
37	WS	G3	0	2	6 ANOS
37	SW	G4	2	0	6 ANOS
37	WS	G4	0	2	6 ANOS
38	SW	G2	3	1	6 ANOS
38	WS	G2	3	1	6 ANOS
38	SW	G3	4	0	6 ANOS
38	WS	G3	3	1	6 ANOS
38	SW	G4	3	0	6 ANOS
38	WS	G4	3	1	6 ANOS
39	SW	G2	6	2	ADULTOS
39	WS	G2	4	4	ADULTOS
39	SW	G3	7	1	ADULTOS
39	WS	G3	5	3	ADULTOS
39	SW	G4	6	1	ADULTOS
39	WS	G4	6	2	ADULTOS
40	SW	G2	7	1	ADULTOS
40	WS	G2	4	4	ADULTOS
40	SW	G3	8	0	ADULTOS
40	WS	G3	4	4	ADULTOS
40	SW	G4	6	1	ADULTOS
40	WS	G4	7	1	ADULTOS
41	SW	G2	3	5	ADULTOS
41	WS	G2	5	3	ADULTOS
41	SW	G3	4	4	ADULTOS
41	WS	G3	4	4	ADULTOS
41	SW	G4	3	4	ADULTOS
41	WS	G4	4	4	ADULTOS

id	Entonacao	Gate	Sucesso	Fracasso	Idade
42	SW	G2	6	2	ADULTOS
42	WS	G2	2	6	ADULTOS
42	SW	G3	2	6	ADULTOS
42	WS	G3	3	5	ADULTOS
42	SW	G4	4	3	ADULTOS
42	WS	G4	3	5	ADULTOS
43	SW	G2	6	2	ADULTOS
43	WS	G2	5	3	ADULTOS
43	SW	G3	6	2	ADULTOS
43	WS	G3	3	5	ADULTOS
43	SW	G4	6	1	ADULTOS
43	WS	G4	6	2	ADULTOS
44	SW	G2	5	3	ADULTOS
44	WS	G2	4	4	ADULTOS
44	SW	G3	6	2	ADULTOS
44	WS	G3	4	4	ADULTOS
44	SW	G4	5	2	ADULTOS
44	WS	G4	6	2	ADULTOS
45	SW	G2	4	4	ADULTOS
45	WS	G2	5	3	ADULTOS
45	SW	G3	4	4	ADULTOS
45	WS	G3	3	5	ADULTOS
45	SW	G4	4	3	ADULTOS
45	WS	G4	4	4	ADULTOS
46	SW	G2	5	3	ADULTOS
46	WS	G2	6	2	ADULTOS
46	SW	G3	6	2	ADULTOS
46	WS	G3	3	5	ADULTOS
46	SW	G4	7	0	ADULTOS

id	Entonacao	Gate	Sucesso	Fracasso	Idade
46	WS	G4	5	3	ADULTOS
47	SW	G2	7	1	ADULTOS
47	WS	G2	3	5	ADULTOS
47	SW	G3	6	2	ADULTOS
47	WS	G3	1	7	ADULTOS
47	SW	G4	7	0	ADULTOS
47	WS	G4	2	6	ADULTOS
48	SW	G2	4	4	ADULTOS
48	WS	G2	5	3	ADULTOS
48	SW	G3	6	2	ADULTOS
48	WS	G3	4	4	ADULTOS
48	SW	G4	4	3	ADULTOS
48	WS	G4	8	0	ADULTOS
49	SW	G2	7	1	ADULTOS
49	WS	G2	7	1	ADULTOS
49	SW	G3	5	3	ADULTOS
49	WS	G3	7	1	ADULTOS
49	SW	G4	6	1	ADULTOS
49	WS	G4	8	0	ADULTOS
50	SW	G2	8	0	ADULTOS
50	WS	G2	4	4	ADULTOS
50	SW	G3	8	0	ADULTOS
50	WS	G3	6	2	ADULTOS
50	SW	G4	6	1	ADULTOS
50	WS	G4	8	0	ADULTOS
51	SW	G2	5	3	ADULTOS
51	WS	G2	2	6	ADULTOS
51	SW	G3	5	3	ADULTOS
51	WS	G3	4	4	ADULTOS

id	Entonacao	Gate	Sucesso	Fracasso	Idade
51	SW	G4	7	0	ADULTOS
51	WS	G4	4	4	ADULTOS
52	SW	G2	8	0	ADULTOS
52	WS	G2	4	4	ADULTOS
52	SW	G3	8	0	ADULTOS
52	WS	G3	7	1	ADULTOS
52	SW	G4	7	0	ADULTOS
52	WS	G4	7	1	ADULTOS
53	SW	G2	5	1	ADULTOS
53	WS	G2	2	4	ADULTOS
53	SW	G3	5	1	ADULTOS
53	WS	G3	4	2	ADULTOS
53	SW	G4	6	0	ADULTOS
53	WS	G4	3	3	ADULTOS
54	SW	G2	7	1	ADULTOS
54	WS	G2	2	6	ADULTOS
54	SW	G3	8	0	ADULTOS
54	WS	G3	6	2	ADULTOS
54	SW	G4	5	2	ADULTOS
54	WS	G4	7	1	ADULTOS
55	SW	G2	7	1	ADULTOS
55	WS	G2	4	4	ADULTOS
55	SW	G3	8	0	ADULTOS
55	WS	G3	5	3	ADULTOS
55	SW	G4	6	1	ADULTOS
55	WS	G4	8	0	ADULTOS
56	SW	G2	5	3	ADULTOS
56	WS	G2	3	5	ADULTOS
56	SW	G3	6	2	ADULTOS

id	Entonacao	Gate	Sucesso	Fracasso	Idade
56	WS	G3	3	5	ADULTOS
56	SW	G4	5	2	ADULTOS
56	WS	G4	2	6	ADULTOS

A.2 Conjunto de dados do Exemplo 2

Sexo	id	Silabas	Idade	Acertos	Erros
M	1	1	3	10	0
M	1	2	3	10	0
M	1	3	3	9	1
M	1	4	3	6	4
F	2	1	3	10	0
F	2	2	3	10	0
F	2	3	3	10	0
F	2	4	3	5	5
F	3	1	3	10	0
F	3	2	3	10	0
F	3	3	3	8	2
F	3	4	3	9	1
M	4	1	3	10	0
M	4	2	3	10	0
M	4	3	3	10	0
M	4	4	3	10	0
F	5	1	3	10	0
F	5	2	3	9	1
F	5	3	3	7	3
F	5	4	3	7	3
M	6	1	3	10	0
M	6	2	3	9	1

Sexo	id	Silabas	Idade	Acertos	Erros
M	6	3	3	9	1
M	6	4	3	9	1
F	7	1	3	10	0
F	7	2	3	7	3
F	7	3	3	6	4
F	7	4	3	5	5
F	8	1	3	10	0
F	8	2	3	10	0
F	8	3	3	10	0
F	8	4	3	0	10
F	9	1	3	10	0
F	9	2	3	10	0
F	9	3	3	8	2
F	9	4	3	5	5
M	10	1	3	10	0
M	10	2	3	10	0
M	10	3	3	10	0
M	10	4	3	10	0
M	11	1	3	10	0
M	11	2	3	9	1
M	11	3	3	8	2
M	11	4	3	8	2
M	12	1	3	10	0
M	12	2	3	10	0
M	12	3	3	9	1
M	12	4	3	2	8
F	13	1	3	10	0
F	13	2	3	10	0
F	13	3	3	9	1

Sexo	id	Silabas	Idade	Acertos	Erros
F	13	4	3	4	6
F	14	1	3	10	0
F	14	2	3	9	1
F	14	3	3	6	4
F	14	4	3	4	6
F	15	1	3	10	0
F	15	2	3	9	1
F	15	3	3	8	2
F	15	4	3	4	6
F	16	1	3	10	0
F	16	2	3	10	0
F	16	3	3	7	3
F	16	4	3	3	7
F	17	1	3	10	0
F	17	2	3	9	1
F	17	3	3	8	2
F	17	4	3	8	2
M	18	1	3	10	0
M	18	2	3	8	2
M	18	3	3	8	2
M	18	4	3	7	3
M	19	1	3	10	0
M	19	2	3	10	0
M	19	3	3	9	1
M	19	4	3	9	1
M	20	1	3	9	1
M	20	2	3	7	3
M	20	3	3	6	4
M	20	4	3	9	1

Sexo	id	Silabas	Idade	Acertos	Erros
M	21	1	3	10	0
M	21	2	3	9	1
M	21	3	3	9	1
M	21	4	3	9	1
F	22	1	3	9	1
F	22	2	3	8	2
F	22	3	3	7	3
F	22	4	3	7	3
M	23	1	3	9	1
M	23	2	3	8	2
M	23	3	3	7	3
M	23	4	3	7	3
M	24	1	3	10	0
M	24	2	3	10	0
M	24	3	3	10	0
M	24	4	3	10	0
F	25	1	3	10	0
F	25	2	3	10	0
F	25	3	3	9	1
F	25	4	3	6	4
M	26	1	3	10	0
M	26	2	3	7	3
M	26	3	3	6	4
M	26	4	3	9	1
M	27	1	3	10	0
M	27	2	3	9	1
M	27	3	3	3	7
M	27	4	3	5	5
M	28	1	3	10	0

Sexo	id	Silabas	Idade	Acertos	Erros
M	28	2	3	9	1
M	28	3	3	9	1
M	28	4	3	3	7
F	29	1	3	10	0
F	29	2	3	10	0
F	29	3	3	9	1
F	29	4	3	6	4
F	30	1	3	10	0
F	30	2	3	8	2
F	30	3	3	6	4
F	30	4	3	7	3
M	31	1	3	9	1
M	31	2	3	9	1
M	31	3	3	6	4
M	31	4	3	9	1
F	32	1	3	10	0
F	32	2	3	8	2
F	32	3	3	10	0
F	32	4	3	10	0
F	33	1	3	10	0
F	33	2	3	9	1
F	33	3	3	7	3
F	33	4	3	8	2
M	34	1	3	10	0
M	34	2	3	6	4
M	34	3	3	2	8
M	34	4	3	0	10
M	35	1	4	10	0
M	35	2	4	10	0

Sexo	id	Silabas	Idade	Acertos	Erros
M	35	3	4	7	3
M	35	4	4	7	3
F	36	1	4	9	1
F	36	2	4	8	2
F	36	3	4	7	3
F	36	4	4	8	2
M	37	1	4	10	0
M	37	2	4	9	1
M	37	3	4	9	1
M	37	4	4	8	2
F	38	1	4	10	0
F	38	2	4	8	2
F	38	3	4	6	4
F	38	4	4	3	7
F	39	1	4	10	0
F	39	2	4	10	0
F	39	3	4	10	0
F	39	4	4	5	5
M	40	1	4	8	2
M	40	2	4	6	4
M	40	3	4	7	3
M	40	4	4	3	7
F	41	1	4	10	0
F	41	2	4	8	2
F	41	3	4	9	1
F	41	4	4	8	2
F	42	1	4	10	0
F	42	2	4	8	2
F	42	3	4	9	1

Sexo	id	Silabas	Idade	Acertos	Erros
F	42	4	4	6	4
F	43	1	4	10	0
F	43	2	4	10	0
F	43	3	4	9	1
F	43	4	4	6	4
M	44	1	4	10	0
M	44	2	4	8	2
M	44	3	4	6	4
M	44	4	4	3	7
F	45	1	4	10	0
F	45	2	4	10	0
F	45	3	4	10	0
F	45	4	4	9	1
M	46	1	4	10	0
M	46	2	4	10	0
M	46	3	4	8	2
M	46	4	4	5	5
M	47	1	4	10	0
M	47	2	4	9	1
M	47	3	4	6	4
M	47	4	4	8	2
F	48	1	4	9	1
F	48	2	4	10	0
F	48	3	4	8	2
F	48	4	4	4	6
F	49	1	4	10	0
F	49	2	4	10	0
F	49	3	4	9	1
F	49	4	4	9	1

Sexo	id	Silabas	Idade	Acertos	Erros
M	50	1	4	9	1
M	50	2	4	9	1
M	50	3	4	9	1
M	50	4	4	5	5
F	51	1	4	10	0
F	51	2	4	8	2
F	51	3	4	8	2
F	51	4	4	2	8
M	52	1	4	10	0
M	52	2	4	10	0
M	52	3	4	5	5
M	52	4	4	1	9
F	53	1	4	10	0
F	53	2	4	9	1
F	53	3	4	8	2
F	53	4	4	8	2
M	54	1	4	10	0
M	54	2	4	8	2
M	54	3	4	8	2
M	54	4	4	8	2
M	55	1	4	9	1
M	55	2	4	8	2
M	55	3	4	7	3
M	55	4	4	4	6
F	56	1	4	10	0
F	56	2	4	10	0
F	56	3	4	9	1
F	56	4	4	9	1
F	57	1	4	10	0

Sexo	id	Silabas	Idade	Acertos	Erros
F	57	2	4	8	2
F	57	3	4	5	5
F	57	4	4	2	8
F	58	1	4	9	1
F	58	2	4	10	0
F	58	3	4	8	2
F	58	4	4	6	4
F	59	1	4	9	1
F	59	2	4	9	1
F	59	3	4	9	1
F	59	4	4	4	6
F	60	1	4	10	0
F	60	2	4	9	1
F	60	3	4	6	4
F	60	4	4	8	2
M	61	1	4	10	0
M	61	2	4	10	0
M	61	3	4	9	1
M	61	4	4	8	2
M	62	1	4	9	1
M	62	2	4	9	1
M	62	3	4	10	0
M	62	4	4	8	2
M	63	1	4	9	1
M	63	2	4	8	2
M	63	3	4	8	2
M	63	4	4	7	3
M	64	1	4	9	1
M	64	2	4	8	2

Sexo	id	Silabas	Idade	Acertos	Erros
M	64	3	4	9	1
M	64	4	4	10	0
M	65	1	4	9	1
M	65	2	4	8	2
M	65	3	4	6	4
M	65	4	4	10	0
M	66	1	4	10	0
M	66	2	4	10	0
M	66	3	4	10	0
M	66	4	4	10	0
M	67	1	4	10	0
M	67	2	4	9	1
M	67	3	4	8	2
M	67	4	4	2	8
M	68	1	4	9	1
M	68	2	4	6	4
M	68	3	4	4	6
M	68	4	4	4	6
M	69	1	5	10	0
M	69	2	5	10	0
M	69	3	5	10	0
M	69	4	5	6	4
F	70	1	5	10	0
F	70	2	5	10	0
F	70	3	5	9	1
F	70	4	5	3	7
M	71	1	5	10	0
M	71	2	5	10	0
M	71	3	5	10	0

Sexo	id	Silabas	Idade	Acertos	Erros
M	71	4	5	10	0
M	72	1	5	10	0
M	72	2	5	9	1
M	72	3	5	3	7
M	72	4	5	7	3
M	73	1	5	9	1
M	73	2	5	7	3
M	73	3	5	9	1
M	73	4	5	10	0
M	74	1	5	10	0
M	74	2	5	10	0
M	74	3	5	5	5
M	74	4	5	9	1
F	75	1	5	10	0
F	75	2	5	10	0
F	75	3	5	10	0
F	75	4	5	10	0
F	76	1	5	10	0
F	76	2	5	10	0
F	76	3	5	10	0
F	76	4	5	8	2
M	77	1	5	10	0
M	77	2	5	10	0
M	77	3	5	10	0
M	77	4	5	8	2
F	78	1	5	10	0
F	78	2	5	9	1
F	78	3	5	6	4
F	78	4	5	1	9

Sexo	id	Silabas	Idade	Acertos	Erros
M	79	1	5	10	0
M	79	2	5	10	0
M	79	3	5	10	0
M	79	4	5	10	0
F	80	1	5	10	0
F	80	2	5	10	0
F	80	3	5	10	0
F	80	4	5	5	5
F	81	1	5	10	0
F	81	2	5	10	0
F	81	3	5	10	0
F	81	4	5	10	0
F	82	1	5	10	0
F	82	2	5	10	0
F	82	3	5	10	0
F	82	4	5	9	1
F	83	1	5	10	0
F	83	2	5	10	0
F	83	3	5	4	6
F	83	4	5	4	6
M	84	1	5	10	0
M	84	2	5	10	0
M	84	3	5	10	0
M	84	4	5	10	0
F	85	1	5	10	0
F	85	2	5	9	1
F	85	3	5	10	0
F	85	4	5	10	0
M	86	1	5	10	0

Sexo	id	Silabas	Idade	Acertos	Erros
M	86	2	5	10	0
M	86	3	5	10	0
M	86	4	5	10	0
M	87	1	5	10	0
M	87	2	5	10	0
M	87	3	5	10	0
M	87	4	5	10	0
F	88	1	5	10	0
F	88	2	5	10	0
F	88	3	5	10	0
F	88	4	5	8	2
F	89	1	5	10	0
F	89	2	5	10	0
F	89	3	5	9	1
F	89	4	5	10	0
M	90	1	5	10	0
M	90	2	5	10	0
M	90	3	5	7	3
M	90	4	5	9	1
F	91	1	5	10	0
F	91	2	5	10	0
F	91	3	5	8	2
F	91	4	5	6	4
F	92	1	5	10	0
F	92	2	5	10	0
F	92	3	5	10	0
F	92	4	5	10	0
M	93	1	5	10	0
M	93	2	5	10	0

Sexo	id	Silabas	Idade	Acertos	Erros
M	93	3	5	10	0
M	93	4	5	9	1
M	94	1	5	10	0
M	94	2	5	10	0
M	94	3	5	7	3
M	94	4	5	5	5
F	95	1	5	10	0
F	95	2	5	10	0
F	95	3	5	9	1
F	95	4	5	9	1
M	96	1	5	9	1
M	96	2	5	8	2
M	96	3	5	9	1
M	96	4	5	10	0
M	97	1	5	10	0
M	97	2	5	10	0
M	97	3	5	9	1
M	97	4	5	10	0
M	98	1	5	9	1
M	98	2	5	10	0
M	98	3	5	9	1
M	98	4	5	10	0
M	99	1	5	10	0
M	99	2	5	10	0
M	99	3	5	10	0
M	99	4	5	8	2
F	100	1	5	10	0
F	100	2	5	10	0
F	100	3	5	10	0

Sexo	id	Silabas	Idade	Acertos	Erros
F	100	4	5	9	1
F	101	1	5	10	0
F	101	2	5	10	0
F	101	3	5	9	1
F	101	4	5	9	1
F	102	1	5	10	0
F	102	2	5	10	0
F	102	3	5	9	1
F	102	4	5	3	7
F	103	1	6	10	0
F	103	2	6	10	0
F	103	3	6	10	0
F	103	4	6	10	0
F	104	1	6	10	0
F	104	2	6	10	0
F	104	3	6	10	0
F	104	4	6	9	1
F	105	1	6	10	0
F	105	2	6	10	0
F	105	3	6	10	0
F	105	4	6	9	1
M	106	1	6	10	0
M	106	2	6	10	0
M	106	3	6	10	0
M	106	4	6	10	0
F	107	1	6	10	0
F	107	2	6	10	0
F	107	3	6	8	2
F	107	4	6	7	3

Sexo	id	Silabas	Idade	Acertos	Erros
F	108	1	6	10	0
F	108	2	6	10	0
F	108	3	6	10	0
F	108	4	6	10	0
F	109	1	6	10	0
F	109	2	6	10	0
F	109	3	6	10	0
F	109	4	6	10	0
M	110	1	6	10	0
M	110	2	6	10	0
M	110	3	6	10	0
M	110	4	6	10	0
F	111	1	6	10	0
F	111	2	6	10	0
F	111	3	6	10	0
F	111	4	6	10	0
M	112	1	6	10	0
M	112	2	6	10	0
M	112	3	6	10	0
M	112	4	6	10	0
F	113	1	6	10	0
F	113	2	6	10	0
F	113	3	6	10	0
F	113	4	6	7	3
F	114	1	6	10	0
F	114	2	6	10	0
F	114	3	6	10	0
F	114	4	6	10	0
M	115	1	6	10	0

Sexo	id	Silabas	Idade	Acertos	Erros
M	115	2	6	10	0
M	115	3	6	10	0
M	115	4	6	10	0
M	116	1	6	10	0
M	116	2	6	10	0
M	116	3	6	9	1
M	116	4	6	9	1
M	117	1	6	10	0
M	117	2	6	10	0
M	117	3	6	10	0
M	117	4	6	10	0
M	118	1	6	10	0
M	118	2	6	10	0
M	118	3	6	10	0
M	118	4	6	10	0
M	119	1	6	10	0
M	119	2	6	10	0
M	119	3	6	7	3
M	119	4	6	5	5
F	120	1	6	9	1
F	120	2	6	10	0
F	120	3	6	10	0
F	120	4	6	9	1
M	121	1	6	10	0
M	121	2	6	10	0
M	121	3	6	10	0
M	121	4	6	10	0
M	122	1	6	10	0
M	122	2	6	10	0

Sexo	id	Silabas	Idade	Acertos	Erros
M	122	3	6	10	0
M	122	4	6	10	0
F	123	1	6	10	0
F	123	2	6	9	1
F	123	3	6	10	0
F	123	4	6	9	1
F	124	1	6	10	0
F	124	2	6	10	0
F	124	3	6	10	0
F	124	4	6	8	2
F	125	1	6	10	0
F	125	2	6	10	0
F	125	3	6	10	0
F	125	4	6	10	0
M	126	1	6	10	0
M	126	2	6	10	0
M	126	3	6	10	0
M	126	4	6	10	0
M	127	1	6	10	0
M	127	2	6	9	1
M	127	3	6	10	0
M	127	4	6	9	1
M	128	1	6	10	0
M	128	2	6	10	0
M	128	3	6	10	0
M	128	4	6	8	2
M	129	1	6	10	0
M	129	2	6	8	2
M	129	3	6	9	1

Sexo	id	Silabas	Idade	Acertos	Erros
M	129	4	6	9	1
F	130	1	6	9	1
F	130	2	6	10	0
F	130	3	6	10	0
F	130	4	6	9	1
M	131	1	6	10	0
M	131	2	6	10	0
M	131	3	6	10	0
M	131	4	6	10	0
F	132	1	6	10	0
F	132	2	6	10	0
F	132	3	6	10	0
F	132	4	6	10	0
F	133	1	6	10	0
F	133	2	6	10	0
F	133	3	6	10	0
F	133	4	6	8	2
F	134	1	6	9	1
F	134	2	6	9	1
F	134	3	6	10	0
F	134	4	6	7	3
F	135	1	6	10	0
F	135	2	6	9	1
F	135	3	6	7	3
F	135	4	6	6	4
M	136	1	6	10	0
M	136	2	6	8	2
M	136	3	6	10	0
M	136	4	6	2	8

A.3 Conjunto de dados do Exemplo 3

Rato	Droga	NGB	NGV	NCC	Tempo
1	1	15	2	14	1
1	1	18	3	14	2
1	1	19	2	12	3
1	1	24	5	11	4
2	1	8	2	17	1
2	1	11	4	18(NA)	2
2	1	14	4	18	3
2	1	14	5	16	4
3	1	4	7	23(NA)	1
3	1	5	5	20	2
3	1	6	4	19	3
3	1	4	4	19	4
4	1	16	3	13	1
4	1	14	4	12	2
4	1	14	4	12	3
4	1	12	2	11(NA)	4
5	1	6	7	24	1
5	1	4	6	20	2
5	1	4	5	20	3
5	1	4	2	19	4
6	1	22	4	12	1
6	1	20	3	12	2
6	1	21	3	10	3
6	1	18	2	9	4
7	1	18	5	16(NA)	1
7	1	17	3	16	2
7	1	17	5	14	3

Rato	Droga	NGB	NGV	NCC	Tempo
7	1	16	2	12	4
8	1	4	8	28	1
8	1	7	7	26(NA)	2
8	1	4	4	26	3
8	1	4	4	26	4
9	1	14	3	14	1
9	1	12	4	13	2
9	1	12	4	12	3
9	1	10	5	10	4
10	1	10	3	16	1
10	1	10	4	15(NA)	2
10	1	10	5	15	3
10	1	10	2	14	4
11	2	14	6	16	1
11	2	14	6	15	2
11	2	16	7	15	3
11	2	17	6	14	4
12	2	7	4	36	1
12	2	7	4	32	2
12	2	6	4	30	3
12	2	5	2	29	4
13	2	9	8	18	1
13	2	8	8	16(NA)	2
13	2	9	7	17	3
13	2	11	4	15	4
14	2	21	3	14	1
14	2	20	3	13	2
14	2	20	4	13	3
14	2	20	3	12	4

Rato	Droga	NGB	NGV	NCC	Tempo
15	2	18	4	19	1
15	2	17	4	19	2
15	2	17	2	18(NA)	3
15	2	17	2	17	4
16	2	3	10	38	1
16	2	6	10	38	2
16	2	6	8	37	3
16	2	2	7	37	4
17	2	8	3	18	1
17	2	9	3	18	2
17	2	9	2	17	3
17	2	8	2	16	4
18	2	29	6	8	1
18	2	30	6	8	2
18	2	29	5	7	3
18	2	29	4	6	4
19	2	8	9	19(NA)	1
19	2	8	9	19	2
19	2	8	8	18	3
19	2	7	8	17	4
20	2	5	8	36	1
20	2	4	7	35	2
20	2	4	7	30	3
20	2	3	7	29	4
21	3	16	2	15(NA)	1
21	3	17	3	16	2
21	3	17	4	17	3
21	3	18	2	15(NA)	4
22	3	13	6	17	1

Rato	Droga	NGB	NGV	NCC	Tempo
22	3	11	4	16(NA)	2
22	3	12	5	16	3
22	3	12	4	18	4
23	3	7	3	28	1
23	3	8	2	25	2
23	3	6	2	27	3
23	3	5	3	31	4
24	3	9	4	29	1
24	3	8	5	30	2
24	3	9	3	32(NA)	3
24	3	9	3	30	4
25	3	18	3	11	1
25	3	19	2	12	2
25	3	21	5	12	3
25	3	20	4	13	4
26	3	23	5	8	1
26	3	25	5	10(NA)	2
26	3	24	4	9	3
26	3	24	4	8	4
27	3	27	7	7	1
27	3	28	6	8	2
27	3	27	6	8	3
27	3	30	4	7	4
28	3	30	6	4	1
28	3	32	7	5	2
28	3	33	8	5	3
28	3	35	7	4	4
29	3	17	4	14(NA)	1
29	3	19	3	13	2

Rato	Droga	NGB	NGV	NCC	Tempo
29	3	20	3	13	3
29	3	21	2	12	4
30	3	12	3	17	1
30	3	12	5	15	2
30	3	13	4	16	3
30	3	11	5	16	4

Apêndice B

Algoritmos

B.1 Pontos de alavanca, influentes e aberrantes

```
diag_gee <- function(fit.model, dados){
  n <- length(levels(as.factor(fit.model$id)))
  N <- fit.model$nobs
  R <- fit.model$working.correlation #Matriz de corr. de trabalho
  pr <- ncol(R) #dimensao de R. Numero de repeticoes.
  mu <- fit.model$fitted.values #valores ajustados
  X <- model.matrix(lm(formula=as.formula(fit.model$call$formula),
                                                                dados))

  p <- ncol(X) #numero de parametros de beta.
  y <- fit.model$y
  fit.family <- fit.model$family
  beta <- fit.model$coef
  fit.id <- as.character(fit.model$call$id)
  fit.resid <- fit.model$resid

  ifelse((fit.family$family=="binomial"),
```

```

match_resp <-
substring(as.character(fit.model$call$formula[2]),
  regexpr("(", as.character(fit.model$call$formula[2]),
    fixed = TRUE)[1] + 1,
  ifelse(regexpr(",",
    as.character(fit.model$call$formula[2]))[1] - 1 > 0,
    regexpr(",",
      as.character(fit.model$call$formula[2]))[1] - 1,
      nchar(as.character(fit.model$call$formula[2]))
    )
  )
),
match_resp <- as.character(fit.model$call$formula[2]))

resp <- dados[, !is.na(match(colnames(dados), match_resp))]

N_bin <- NA
fracasso <- NA
if(fit.family$family=="binomial"){
  match_fracasso <-
  substring(as.character(fit.model$call$formula[2]),
    regexpr(",",
      as.character(fit.model$call$formula[2]))[1] + 2,
    ifelse(regexpr(")",
      as.character(fit.model$call$formula[2]),
      fixed = TRUE)[1] - 1 > 0,
      regexpr(")",
        as.character(fit.model$call$formula[2]),
        fixed = TRUE)[1] - 1,

```



```

      nchar(as.character(fit.model$call$formula[2])))
  ifelse(
    regexpr(",",
      as.character(fit.model$call$formula[2]))[1]==-1,
    fracasso <- -1*(resp-1),
    fracasso <- dados[,!is.na(match(colnames(dados),
                                     match_fracasso))])

  N_bin <- resp + fracasso
}

if(fit.family$family=="binomial") {
  fit.resid <- fit.model$y - fit.model$fitted*N_bin[!is.na(N_bin)]
}
cluster <- dados[,colnames(dados)==fit.id]

#Conta o numero de repeticoes validas para cada unidade amostral.
Valid_rep <- as.vector(table(cluster) - tapply(is.na(resp),
                                                cluster, sum))

invphi <- fit.model$scale
phi <- 1/invphi
if(fit.family$family=="binomial"){
  A <- diag(N_bin[!is.na(N_bin)]*mu*(1-mu))
  Delta <- diag(N)
}
if(fit.family$family=="gaussian"){
  A <- diag(N)
  Delta <- diag(N)
}
if(fit.family$family=="poisson"){

```

```

A <- diag(mu,N)
Delta <- diag(N)
}
if (fit.family$family=="Gamma"){
  A <- diag(mu^2,N)
  Delta <- diag(1/mu,N)
}

#Ligação canônica -> Delta=Identidade
C <- Delta%*%A

#Matriz Omega - variancia e covariancia de y
Omega <- matrix(0,N,N)
invOmega <- matrix(0,N,N)
l <- 1
selec <- matrix(FALSE, nrow=n, ncol=pr)
for(i in 1:n)
{
  INT <- 1:(l+Valid_rep[i]-1) # Intervalo da submatriz.
  INT_R <- !is.na(resp[as.factor(cluster)==
                        levels(as.factor(cluster))[i]])
  if(length(INT_R) < ncol(R)) {
    INT_R <- c(INT_R, rep(FALSE, ncol(R) - length(INT_R)))
  }
  A_box <- A[INT,INT]
  R_box <- R[INT_R,INT_R]
  Omega[INT,INT] <- sqrt(A_box)%*%R_box%*%sqrt(A_box)
  invOmega[INT,INT] <- solve(Omega[INT,INT])
  l <- l+Valid_rep[i]
  selec[i,] <- INT_R # Usado para selecionar o haux.

```

```

    }
    Omega <- invphi*Omega
    invOmega <- phi*invOmega

#Matriz W e H
W <- C%*%invOmega%*%C
H <- solve(t(X)%*%W%*%X)
raizW <- matrix(0,N,N)
l <- 1
for(i in 1:n)
{
  INT <- 1:(l+Valid_rep[i]-1)
  auto <- eigen(W[INT,INT], symmetric=T)
  ifelse(length(INT)==1,
          raizW[INT,INT] <- sqrt(auto$values),
          raizW[INT,INT] <- auto$vectors%*%
                                sqrt(diag(auto$values))%*%
                                t(auto$vectors)
  )
  l <- l+Valid_rep[i]
}
H <- raizW%*%X%*%H%*%t(X)%*%raizW
h <- diag(H)

#Ponto Alavanca por UE
cut <- 2*p/N
hue<-as.vector(rep(0,n))
haux <- matrix(nrow=n, ncol=pr)
haux[selec] <- resp[!is.na(resp)]
haux <- t(haux)

```

```

haux[!is.na(haux)] <- h
hue <- apply(haux, 2, sum, na.rm=TRUE)/Valid_rep

#Residuo Padronizado
rsd <- (raizW%%solve(C)%%(fit.resid))/sqrt(1-h)

#Distancia de Cook
cd <- (rsd^2*h)/((1-h)*p)

#Pontos de corte
val.out.h <- as.matrix(cbind(dados[!is.na(resp)],
  [h>cut,1:2],h[h>cut]))
val.out.cd <- as.matrix(cbind(dados[!is.na(resp)],
  [cd>.01,1:2],cd[cd>0.01]))
val.out.rsd <- as.matrix(cbind(dados[!is.na(resp)],
  [abs(rsd)>3,1:2],rsd[abs(rsd)>3]))
Ptos <- list("Diag_H"=val.out.h, "Cook_dist"=val.out.cd,
  "Resid_pad"=val.out.rsd)

par(mfrow=c(2,2))
plot(cluster[!is.na(resp)],h,xlab="Unidade Experimental",
  ylab="Diagonal de H", pch=16, ylim=c(0,max(h,cut+0.1*cut)))
abline(cut,0,lty=2)

plot(hue,xlab="Unidade Experimental", ylab="H por unidade
  experimental", pch=16, ylim=c(0,max(h,cut+0.1*cut)))
abline(cut,0,lty=2)

plot(cluster[!is.na(resp)],cd,xlab="Unidade Experimental",
  ylab="Distancia de Cook", pch=16)

```

```

plot(cluster[!is.na(resp)], rsd, xlab="Unidade Experimental",
ylab="Resíduo Padronizado", pch=16)

print(Ptos)
RETURN <- list(title="Diagnostico GEE", Ptos=Ptos,
invOmega=invOmega, Omega=Omega, A=A, C=C, raizW=raizW, phi=phi,
invphi=invphi, N_bin=N_bin, resp=resp, fracasso=fracasso,
selec=selec, h=h, hue=hue, cook=cd, rsd=rsd)
return(invisible(RETURN))
}

```

B.2 Envelope

```

envel_gee <- function(fit.model, dados, diaggee){
  library(MASS) #para a funcao mvnrm.
  n <- length(levels(as.factor(fit.model$id))) #numero de u.e.
  N <- fit.model$nobs # ncol(X)
  R <- fit.model$working.correlation # Matriz de corr. de trabalho
  pr <- ncol(R) # numero de repeticoes maxima das u.e.
  mu <- fit.model$fitted.values # valores ajustados
  X <- model.matrix(lm(formula=as.formula(fit.model$call$formula),
                        dados))

  p <- ncol(X) # número de parametros do modelo
  y <- fit.model$y
  beta <- fit.model$coef
  fit.constr <- fit.model$call$constr
  fit.family <- fit.model$family
  fit.formula <- as.formula(paste("random.y[,k]",

```

```

fit.model$call$formula[1],
fit.model$call$formula[3]))
fit.id <- as.character(fit.model$call$id)
phi <- diaggee$phi
invphi <- diaggee$invphi
resp <- diaggee$resp
if(fit.family$family=="binomial"){
  fracasso <- diaggee$fracasso
  N_bin <- resp + fracasso
  fit.formula <- as.formula(paste("cbind(random.y[,k],
                                   N_bin-random.y[,k])",
                                   fit.model$call$formula[1],
                                   fit.model$call$formula[3]))
}

cluster <- dados[,colnames(dados)==fit.id]
Valid_rep <- as.vector(table(cluster) - tapply(is.na(resp),
                                                cluster, sum))

repl <- 25 #Numero de replicas.
random.y <- array(dim=c(N,repl))
mu.fit <- diaggee$selec
mu.fit[mu.fit==FALSE] <- NA
mu.fit[1,!is.na(mu.fit[1,])] <- mu[1:Valid_rep[1]]
for(i in 2:n){
  mu.fit[i,!is.na(mu.fit[i,])] <-
    mu[(sum(Valid_rep[1:(i-1)])+1):sum(Valid_rep[1:i])]
}

#ResÃduo padronizado do modelo original:

```

```

rsd <- diaggee$rsd

if (fit.family$family=="gaussian")
{
  random.y <- array(dim=c(N, repl)) #Matriz com resp. simuladas
  for (i in 1:repl)
  {
    aux <- NA
    for (j in 1:n)
    {
      covy <- diaggee$Omega[1:Valid_rep[j], 1:Valid_rep[j]]
      media <- mu.fit[j,]
      aux <- c(aux, mvrnorm(1, mu=media[!is.na(media)],
                           Sigma=covy))
    }
    random.y[, i] <- aux[-1]
  }
}

#####
# Gera variaveis binarias, Gamma ou Poisson correlacionadas. #
#####
if (fit.family$family=="binomial" ||
    fit.family$family=="poisson" || fit.family$family=="Gamma")
{
  cont<-1
  for (i in 1:n)
  {
    flag<-T
    k<-0

```

```

TS<-matrix(c(1:Valid_rep[i]))
lambda<-0

media<-mu.fit[i,]
selec_media <- !is.na(media)
R_box <- R[selec_media,selec_media]
if(fit.family$family=="binomial")
{
  ifelse(sum(!is.na(media))==1,
    diag_media <- sqrt((1-media[!is.na(media))]/
      media[!is.na(media)]),
    diag_media <- sqrt(diag((1-media[!is.na(media))]/
      media[!is.na(media)])))
  )
  lambda_aux <- log(matrix(1,Valid_rep[i],Valid_rep[i])
    + diag_media%%R_box%%diag_media)
}
if(fit.family$family=="poisson" || fit.family$family=="Gamma")
{
  ifelse(sum(!is.na(media))==1,
    diag_media <- sqrt(media[!is.na(media)]),
    diag_media <- sqrt(diag(media[!is.na(media)])))
  )
  lambda_aux <- diag_media%%R_box%%diag_media
}
lambda_aux <- round(lambda_aux,6)

while (flag) #Estima os parametros das Poisson
{
  k<-k+1

```



```

lambda.min<-min(lambda_aux[lambda_aux>0])
indice  <- matrix(c(1:Valid_rep[i]), Valid_rep[i],
                    Valid_rep[i], byrow=TRUE)
rs <- indice[lambda_aux==lambda.min]
S <- unique(rs)

for (j in 1:length(S))
  for (l in j:length(S))
    if (lambda_aux[S[j],S[l]]==lambda.min) rs<-c(S[j],S[l])

S <- sort(unique(rs))

for (j in 1:Valid_rep[i])
  if (all(lambda_aux[j,S]>0)) S <- c(S,j)
S <- sort(unique(S))

lambda_aux[S,S] <- lambda_aux[S,S]-lambda.min

ts <- rep(0,Valid_rep[i])
for (j in 1:Valid_rep[i])
  if (is.element(j,S)) ts[j] <- 1

TS <- cbind(TS,ts)
lambda <- c(lambda,lambda.min)

flag <- sum(lambda_aux)!=0
}

TS <- TS[,-1]
lambda <- lambda[-1]

```

```

m <- k

for(j in 1:repl) #Gera as variaveis correlacionadas.
{
  x <- rpois(1,lambda[1]) #Gera a variavel Poisson
                           #Serve para Binomial e Poisson
  if(fit.family$family=="binomial")
  {
    aux <- rep(0,Valid_rep[i])
    for(l in 1:max(N_bin, na.rm=TRUE))
    {
      if(m!=1)
      {
        for(k in 2:m) x <- cbind(x,rpois(1,lambda[k]))
      }
      U <- TS%%t(x) #Soma Poisson independentes.
                     #Criando as Poisson dependentes.
      Z <- as.vector(rep(0,Valid_rep[i]))
      Z[U==0] <- 1 #Gera as Bernoulli.
      Z[N_bin[!is.na(N_bin)][cont:(cont+Valid_rep[i]-1)]<1]
        <- 0 #Trunca as Bernoullis quando necessario

      aux <- aux + Z #Soma Bernoullis
      x <- rpois(1,lambda[1]) #Recomeca o vetor de Poisson.
    }
    random.y[cont:(cont+Valid_rep[i]-1),j] <- aux
  }

  if(fit.family$family=="poisson")
  {

```

```

        if(m!=1) for(k in 2:m) x <- cbind(x,rpois(1,lambda[k]))
        U <- TS%%t(x)
        random.y[cont:(cont+Valid_rep[i]-1),j] <- U
      }
    if(fit.family$family=="Gamma")
    {
      x <- rgamma(1,phi,phi*1/lambda[1])
      if(m!=1) for(k in 2:m)
      {
        x <- cbind(x,rgamma(1,phi,phi*1/lambda[k]))
      }
      U <- TS%%t(x)
      random.y[cont:(cont+Valid_rep[i]-1),j] <- U
    }
  }
  cont<-cont+Valid_rep[i]
}
}

#####
#### Residuo dos modelos simulados ####
#####
dados_aux<-cbind(dados[!is.na(resp),],random.y)
random.resid<-array(dim=c(N,repl))
if(fit.family$family=="binomial") N_bin <- N_bin[!is.na(N_bin)]
for(k in 1:repl)
{
  cat("k:",k,"\\n")

  temp.gee.all<-gee(formula=fit.formula,

```

```

id=dados_aux[,colnames(dados_aux)==fit.id],
data = dados_aux, family = fit.family,
corstr = fit.corstr,
R=R, scale.value = phi,
scale.fix = ifelse(fit.family %in%
                    c("binomial", "poisson"),
                    TRUE, FALSE))

y_aux <- temp.gee.all$y
beta_aux <- temp.gee.all$coef
cat("beta:", beta_aux, "\n")
R_aux <- temp.gee.all$work
cat("R:", R_aux, "\n")
mu_aux <- temp.gee.all$fitted.values

#Matriz Omega – variancia e covariancia de y
Omega <- matrix(0,N,N)
invOmega <- matrix(0,N,N)
i <- 1
l <- 1
for(i in 1:n)
{
  INT <- 1:(l+Valid_rep[i]-1) # Intervalo da submatriz.
  INT_R <- !is.na(resp[as.factor(cluster)==
                      levels(as.factor(cluster))[i]])
  A_box <- diaggee$A[INT,INT]
  R_box <- R_aux[INT_R,INT_R]
  Omega[INT,INT] <- sqrt(A_box)%*%R_box*sqrt(A_box)
  invOmega[INT,INT] <- solve(Omega[INT,INT])
  l <- l+Valid_rep[i]
}

```

```

    }
    Omega <- invphi*Omega
    invOmega <- phi*invOmega

if ( fit.family$family=="binomial" ){
  A <- diag ( N_bin [! is.na ( N_bin )] * mu_aux * ( 1 - mu_aux ) )
  Delta <- diag ( N )
}
if ( fit.family$family=="gaussian" ){
  A <- diag ( N )
  Delta <- diag ( N )
}
if ( fit.family$family=="poisson" ){
  A <- diag ( mu_aux , N )
  Delta <- diag ( N )
}
if ( fit.family$family=="Gamma" ){
  A <- diag ( mu_aux ^ 2 , N )
  Delta <- diag ( 1 / mu_aux , N )
}
#Matriz C <- A * Delta
C <- Delta%*%A

#Matriz H e W
W <- C%*%invOmega%*%C
H <- solve ( t ( X ) %*% W %*% X )
raizW <- matrix ( 0 , N , N )
l <- 1
for ( i in 1:n )
{

```

```

INT  <- 1:(1+Valid_rep[i]-1)
auto <- eigen(W[INT,INT], symmetric=T)
ifelse(length(INT)==1,
        raizW[INT,INT] <- sqrt(auto$values),
        raizW[INT,INT] <- auto$vectors%*%
                                sqrt(diag(auto$values))%*%
                                t(auto$vectors)
        )
l <- 1+Valid_rep[i]
}
H <- raizW%*%X%*%H%*%t(X)%*%raizW
h <- diag(H)

#Residuo Padronizado
fit.resid_aux <- temp.gee.all$resid
if(fit.family$family=="binomial")
{fit.resid_aux <- y_aux - mu_aux*N_bin[!is.na(N_bin)]}
random.resid[,k] <- (raizW%*%solve(diaggee$C)%*%
                    fit.resid_aux )/sqrt(1-h)
}

ABSrandom.resid <- abs(random.resid)
SORTrandom.resid <- array(dim=c(N,repl))
for(k in 1:repl)
{SORTrandom.resid[,k] <- sort(ABSrandom.resid[,k])}

#####
### Gera linhas do grafico: ###
#####
descritiva <- cbind(apply(SORTrandom.resid,1,min),

```

```

                                apply(SORTrandom.resid,1,median),
                                apply(SORTrandom.resid,1,max))
Z <- qnorm(((1:N)+N-1/8)/(2*N+1/2))

final <- cbind(Z, descritiva, sort(abs(rsd)))
limite <- range(final[,c(2,4,5)])

#####
### Faz os graficos ###
#####
par(mfrow=c(1,1), pty="s")
Ptos_fora <- ((final[,2] > final[,5]) | (final[,4] < final[,5]))
cores <- rep("black",N)
cores[Ptos_fora] <- "red"
plot(final[,1], final[,5],
      xlab="Valor Esperado da Estatistica de Ordem Meio-Normal",
      ylab="Valor Absoluto Ordenado do ResÃduo Padronizado",
      ylim=limite, pch=16, col=cores)

lines(final[,1], final[,2])
lines(final[,1], final[,3], lty=2)
lines(final[,1], final[,4])

#porcentagem de pontos fora da banda de confianca:
text(min(final[,1]), max(final[,5]),
     paste("fora da banda:\n",round(sum(Ptos_fora)/N*100,1),"%"),
     cex=.8, adj=c(0,1))
}

```


Referências Bibliográficas

- Afiune, J., Singer, J. e Leone, C. (2005). Evolução ecocardiográfica de recém-nascidos com persistência do canal arterial, *Jornal de Pediatria* **81**: 454–460.
- Artes, R. e Botter, D. A. (2005). *Funções de Estimação em Modelos de Regressão*, São Paulo: Associação Brasileira de Estatística.
- Atkinson, A. (1985). *Plots, Transformations and Regressions*, Oxford Statistical Science Series.
- Botter, D. A., Sandoval, M. C., Brunelli, R. T. e Santos, R. P. (2008). *Relatório de análise estatística sobre o projeto: 'A estrutura prosódica das primeiras palavras: comparação entre o português brasileiro e o holandês'*, Universidade de São Paulo. Relatório de Análise Estatística - CEA - 08P28.
- Cook, D. (1979). Influential observations in linear regression, *Journal of the American Statistical Association* **74**: 169–174.
- Cook, R. (1977). Detection of influential observations in linear regression, *Technometrics* **19**: 15–18.

- Elian, S. N., Freitas, A. G. V. M. e Luz, B. A. (2006). *Relatório de análise estatística sobre o projeto: 'Memória operacional e compreensão de orações em crianças normais quanto ao desenvolvimento de linguagem'*, Universidade de São Paulo. Relatório de Análise Estatística - CEA - 06P02.
- Genz, A. (1992). Numerical computation of multivariate normal probabilities, *Journal of Computational and Graphical Statistics* **1**: 141–149.
- Godambe, V. P. (1991). *Estimating Function (Oxford Statistical Science Series - 7)*, Oxford: Oxford University Press.
- Graybill, F. (1976). *Theory and application of the linear model*, California: Wadsworth.
- Hardin, J. e Hilbe, J. (2003). *Generalized Estimating Equations*, Chapman & Hall/CRC.
- Johnston, G. (1996). *Repeated measures analysis with discrete data using the SAS Symtem.*, Cary: SAS Institute Inc.
- Liang, K.-Y. e Zeger, S. L. (1986). Longitudinal analysis using generalized linear model, *Biometrika* **73**: 13–22.
- Montgomery, D., Myers, R. e Vining, G. (2001). *Generalized Linear Models: with Applications in Engineering and the Sciences*, New York: John Wiley & Sons.
- Nelder, J. e Wedderburn, R. (1972). Generalized linear models, *Jornal of the Royal Statistical Society* **135**: 370–384.
- Neter, J., Kutner, M. H., Nachstheim, C. J. e Wasserman, W. (1996). *Applied Linear Statistical Models*, 4^a edn, Chigado: IE McGraw Hill.

- Park, C., Park, T. e Shin, D. (1996). A simple method for generating correlated binary variates, *The American Statistician* **50**: 306–310.
- Park, C. e Shin, D. (1998). An algorithm for generating correlated random variables in a class of infinitely divisible distributions, *Journal of Statistical Computation and Simulation* **61**: 127–139.
- Paula, G. (2010). *Modelos de Regressão com Apoio Computacional*, São Paulo: IME - Universidade de São Paulo.
- Pregibon, D. (1981). Logistic regression diagnostics, *The Annals of Statistics* **9**: 705–724.
- Prentice, R. (1988). Correlated binary regression with covariates specific to each binary observation, *Biometrics* **44**: 1033–1048.
- Prentice, R. e Zhao, L. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses, *Biometrika* **47**: 835–839.
- Ripley, B. D. (1987). *Stochastic simulation*, John Wiley & Sons.
- Silva, P. V. (2009). *Equações de estimação para dados com medidas repetidas em mais de um fator*, São Paulo: IME-USP. Dissertação de Mestrado, 90p.
- Tan, M., Qu, Y. e M.H., K. (1997). Model diagnostics for marginal regression analysis of correlated binary data, *Commun. Statist. - Simula.* **26**: 539–558.
- Venezuela, M., Botter, D. e Sandoval, M. (2007). Diagnostic techniques in generalized estimating equations, *Journal of Statistical Computation and Simulation* **77**: 879–888.

- Venezuela, M. K. (2003). *Modelos Lineares Generalizados para Análise de Dados com Medidas Repetidas*, São Paulo: IME-USP. Dissertação de Mestrado, 95p.