Universidade Federal de São Carlos
Centro de Ciências e Tecnologia
Programa de Pós-Graduação em Estatística

Análise de Dados Longitudinais para Variáveis Binárias

por

José Tenylson Gonçalves Rodrigues

sob orientação da

Profa. Dra. Cecilia Candolo

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

UFSCar - São Carlos
Junho/2009

Análise de Dados Longitudinais para Variáveis Binárias

por

José Tenylson Gonçalves Rodrigues

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

Área de Concentração: Estatística

Banca examinadora:

- **Profa. Dra. Mariana Curi - ICMC-USP**
- **Profa. Dra. Maria Aparecida de Paiva Franco - UFSCar**
- **Profa. Dra. Cecilia Candolo (Orientadora) - UFSCar**

UFSCar - São Carlos

Junho/2009

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

R696ad

Rodrigues, José Tenylson Gonçalves.
Análise de dados longitudinais para variáveis binárias / José Tenylson Gonçalves Rodrigues. -- São Carlos : UFSCar, 2009.
89 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2009.

1. Análise de regressão. 2. Regressão logística. 3. Estatística - estudos longitudinais. 4. Modelos lineares generalizados. 5. Modelos lineares (Estatística). I. Título.

CDD: 519.536 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística
Via Washington Luis, Km 235 - C.P.676 - CGC 45358058/0001-40
FONE: (016) 3351-8292/3351-8241 - FAX: (016) 3351-8243
13565-905 - SÃO CARLOS-SP-BRASIL
www.ufscar.br/~des ppgest@power.ufscar.br

DECLARAÇÃO

Declaramos, para os devidos fins, que José Tenylson Gonçalves Rodrigues defendeu sua Dissertação de Mestrado no dia 05/03/2009, tendo sido **aprovado.** O aluno deverá apresentar a versão final da dissertação (com as correções e sugestões da Banca, e a ficha catalográfica anexada), e a Certidão Negativa da Biblioteca Comunitária, para formação do processo de homologação e emissão do Diploma do Título.

Igualmente, o aluno deverá apresentar a documentação da pesquisa (rotinas, arquivos em LaTeX, resultados complementares etc.) ao seu orientador, visando facilitar a confecção de relatórios técnicos que condensarão os resultados obtidos.

Essa declaração é válida pelo período de 30 dias.

UNIVERSIDADE FEDERAL PPG-Es SÃO CARLOS

Prof. Dr. Josemar Rodrigues
Coordenador – PPG-Es / UFSCar

Agradecimentos

Agradeço inicialmente a Deus, por estar ao meu lado e por me guiar nas decisões que tomei para alcançar esta conquista. Agradeço ainda:

À minha família pelo apoio emocional e carinho, aos meus pais (Dos Anjos e Waldir) que sempre me incentivaram a estudar e, ao meu irmão (Temilson) pela força e amizade.

Aos amigos que fiz em São Carlos/SP, Juliano, Rafael e aos do alojamento UFSCar, dos Blocos: K, M, 01, 07, 09, 14, 18, 26. Aos moradores do Bloco 25 (Flávia, Douglas, Luciano) por me acolherem de coração aberto, pelas brincadeiras e tudo mais, pois foram a minha família neste período. E aos demais colegas e amigos que não citei, agradeço de coração o apoio. Às meninas argentinas que vieram de intercâmbio, Florencia pela amizade e a Brenda por ter se tornado uma irmão de coração.

À Alina por fazer parte da minha vida e me fazer descobrir sentimentos que nunca tinha sentido antes, pelas maravilhosas conversas, pelos felizes momentos que passamos juntos, compartilhando brincadeiras, sorrisos e passeios inesquecíveis.

Ao Cursinho Pré-Vestibular da UFSCar pela oportunidade de lecionar e desenvolver minhas práticas de iniciação a docência.

À minha orientadora Professora Dra. Cecília Candolo, pelos valiosos ensinamentos dados, pela confiança, pela amizade, pelo rigoroso acompanhamento e revisão do texto e acima de tudo pela paciência.

Aos professores do programa de pós-graduação da UFSCar que incentivaram e contribuíram para a minha formação acadêmica.

Ao professor Dr. José Rubens Rebellato, do Departamento de Fisioterapia da UFSCar, por ter cedido o conjunto de dados. A Sueli, por esclarecer as dúvidas com relação ao conjunto de dados.

Aos professores de graduação do departamento de estatística da UFC, em especial aos professores Maurício, Rosa, Silvia, João Welliandre, Nelson Braga (*in memoriam*) e também aos demais professores pelos valiosos ensinamentos.

Às funcionárias Luiza (Tia Luiza) e Isabel pelas brincadeiras, amizade e atenção.

Aos membros da banca examinadora pelas correções e sugestões para a dissertação.

A CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) por parte do auxílio concedido.

Resumo

O objetivo deste trabalho é apresentar técnicas de análise de regressão para dados longitudinais quando a variável resposta é binária. Inicialmente, é feita uma revisão sobre modelos lineares generalizados, modelos marginais, modelos de transição, modelos mistos, regressão logística e métodos de estimação, pois serão necessários para o desenvolvimento do trabalho.

Além dos métodos de estimação, algumas estruturas de correlação serão estudadas, na tentativa de captar a dependência serial intra-indivíduo ao longo do tempo. Estes métodos foram aplicados em duas situações; uma quando a variável resposta é contínua, e se assume ter distribuição normal, e a outra quando a variável resposta assume ter distribuição de Bernoulli. Também se procurou pesquisar e apresentar técnicas de seleção de modelos e de diagnósticos para os dois casos.

Ao final, uma aplicação com a metodologia pesquisada será apresentada utilizando um conjunto de dados reais.

Palavras-chave: Dados longitudinais, modelos lineares generalizados, modelos marginais, modelos de transição, modelos mistos, variáveis binárias, regressão logística, equação de estimação generalizada.

Abstract

The objective of this work is to present techniques of regression analysis for longitudinal data when the response variable is binary. Initially, there is a review of generalized linear models, marginal models, transition models, mixed models, and logistic regression methods of estimation, which will be necessary for the development of work.

In addition to the methods of estimation, some structures of correlation will be studied in an attempt to capture the intra-individual serial dependence over time. These methods were applied in two situations, one where the response variable is continuous and normal distribution, and another when the response variable has the Bernoulli distribution. It was also sought to explore and present techniques for selection of models and diagnostics for the two cases.

Finally, an application of the above methodology will be presented using a set of real data.

Keywords: Longitudinal data, generalized linear models, marginal models, transition models, models mixtos, binary variables, logistic regression, generalized estimating equation.

Lista de Figuras

Lista de Tabelas

Conteúdo

Capítulo 1

Introdução

1.1 Estudos longitudinais

Estudos longitudinais compõem uma metodologia que avalia o comportamento de uma ou mais variáveis respostas ao longo de uma dimensão específica, que pode ser, por exemplo, o tempo, a distância ou a profundidade. Esta metodologia procura medir o efeito da dependência entre a variável resposta e variáveis explicativas, como também, medir possíveis efeitos entre e/ou intra-indivíduos. Neste trabalho, optou-se por utilizar o tempo como dimensão de estudo.

Os modelos para estudos longitudinais vêem sendo utilizados desde 1890. Mas, as primeiras referências foram os documentos de Henderson (1975), em que foi apresentado o modelo de componentes de variância e a derivação da equação de Henderson, para predizer conjuntamente os efeitos fixos e aleatórios em modelos observados ao longo do tempo.

Grande parte dos esforços empregados na análise deste tipo de dados estão relacionados com a modelagem da estrutura de correlação intra-indivíduos decorrente de medirmos a mesma variável no mesmo indivíduo em tempos diferentes. Com essa finalidade, Laird e Ware (1982) e Ware (1985) propuseram a utilização de modelos lineares mistos. Liang e Zeger (1986), apresentaram uma extensão dos modelos lineares generalizados para a análise de dados longitudinais e também introduziram uma classe de estimadores consistentes às estimativas dos parâmetros do modelo. Os estimadores

propostos para o modelo de regressão foram deduzidos assumindo determinadas formas de correlação entre as medidas sucessivas dentro de um mesmo indivíduo. Tais correlações são especificadas em uma matriz de correlação de "trabalho". Alguns modelos mais usuais para esta matriz serão definidos adiante. Modelar estas estrutura adequadamente é essencial, pois assim as inferências sobre os parâmetros do modelo tornam-se válidas.

Desta forma, pode-se resumir a modelagem de dados longitudinais como: primeiramente identifica-se a relação funcional entre o valor esperado da resposta e as variáveis explicativas e em seguida, modela-se a estrutura de correlação.

Diggle *et al.* (1996), apresentaram esta metodologia de forma ordenada e completa. O desenvolvimento deste método é, em parte, atribuído às diversas contribuições de vários especialistas que trouxeram para uma discussão mais ampla as dificuldades que estavam encontrando, principalmente quanto ao uso em aplicações com conjuntos de dados reais. Após ganhar um fortalecimento no embasamento teórico, esta técnica passou a ser utilizada por pesquisadores das mais diversas áreas do conhecimento, entre elas, economia, farmacologia, sociologia, biologia, medicina.

É interessante observar que os estudos longitudinais fazem parte de uma classe mais ampla conhecida como estudos com medidas repetidas, Singer *et al.* (2007). Uma característica entre eles é que, nos estudos longitudinais o estudo é observacional e nos estudos de medidas repetidas existe aleatorização nas atribuições nos tratamentos dos indivíduos.

1.2 Motivação

As situações práticas mais usuais que envolvem dados longitudinais são aquelas em que a variável resposta tem distribuição normal. Entretanto, tem crescido muito o interesse em modelar situações em que a variável resposta é binária com distribuição Bernoulli. Assim, o objetivo deste trabalho é apresentar a metodologia para a modelagem de dados longitudinais em que a variável resposta é binária. Além disso, observou-se que são escassos na literatura trabalhos abordando, de maneira organizada, a análise de diagnóstico para regressão logística longitudinal. Neste trabalho, pretende-se então:

- Estudar a modelagem clássica de regressão logística longitudinal, focando as abordagens de estimação existentes na literatura, por exemplo, máxima verossimilhança, equações de estimação generalizadas, entre outras;

- Pesquisar e apresentar técnicas de seleção de modelos e de diagnóstico para este tipo de modelagem;

- Aplicar a metodologia pesquisada em um conjunto de dados reais.

1.3 Estrutura da dissertação

Esta dissertação desenvolve-se ao longo de seis capítulos. O conjunto de metas propostas na seção anterior traduzem, ainda que parcialmente, o modo como o trabalho foi estruturado. Nesta seção, ao apresentar a organização da dissertação, pretende-se orientar o leitor quanto aos capítulos a serem apresentados ao longo do seu desenvolvimento. Desta forma, o segundo capítulo apresenta a estrutura dos dados longitudinais, uma revisão sobre modelos lineares generalizados para o caso geral e algumas extensões, tais como: modelo marginal, modelo de transição e modelos com efeitos aleatórios, para o caso em que os dados seguem distribuição normal, apresentando também métodos de estimação.

No terceiro capítulo é feita uma breve descrição sobre dados binários, uma revisão de modelos de regressão logística, o uso do modelo logístico em dados longitudinais com resposta binária, para o caso onde os dados são binários e também alguns métodos de estimação.

No quarto capítulo são apresentadas técnicas de diagnóstico, sejam elas formais (pontos de alavanca, pontos de influência e pontos *outliers*) ou informais (através de gráficos), além de técnicas de qualidade de ajuste.

No quinto capítulo é apresentado um exemplo usando dados reais abordando a metodologia descrita anteriormente. Buscou-se um conjunto de dados que atendesse às principais características deste trabalho.

Por fim, são apresentadas algumas conclusões gerais sobre o trabalho realizado, e alguns apontamentos para propostas de continuação deste estudo.

Capítulo 2

Análise de dados longitudinais

Dados longitudinais é o termo usado para o conjunto de observações feitas em cada elemento de um conjunto de indivíduos sobre uma variável resposta e algumas variáveis explicativas em sucessivos momentos do tempo. A variável resposta pode ser contínua, binária (dicotômica) ou de contagem. Neste capítulo, será apresentado o caso em que a variável resposta assume distribuição normal, com o intuito de introduzir os conceitos de modelagem para esta situação. Posteriormente será abordado o caso em que a variável resposta é binária. Antes de começar a descrever a metodologia, será apresentada a notação a ser utilizada neste trabalho.

2.1 Notação

Em um estudo longitudinal com n indivíduos , cada um deles é observado em n_i, $i = 1, \ldots, n$ ocasiões do tempo quanto a uma variável resposta y e a um vetor de p variáveis explicativas $\mathbf{x}$. Seja y_{ij} a observação de y no i-ésimo indivíduo no tempo j $(j = 1, 2, \ldots, n_i)$.

Em notação matricial temos:

$\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^T$: vetor de observações de y no i-ésimo indivíduo.

$\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)^T$: vetor ($n_T$ x 1), representando o conjunto completo com as $n_T = \sum_{i=1}^{n} n_i$ medidas.

$\mathbf{x}_i$: matriz de observações, (n_i x p) de p covariáveis do i-ésimo indivíduo.

x: matriz de observações, (n_T x p), que contém as informações das covariáveis de todos os indivíduos.

A Tabela (2.1) mostra um exemplo de uma estrutura de dados longitudinais.

Tabela 2.1: Estrutura dos dados longitudinais

Indivíduo	**Tempo**	**Resposta**	**Variáveis explicativas**		
1	1	y_{11}	x_{111}	...	x_{11p}
1	2	y_{12}	x_{121}	...	x_{12p}
⋮	⋮	⋮	⋮	⋮	⋮
1	n_1	y_{1n_1}	x_{1n_11}	...	x_{1n_1p}
⋮	⋮	⋮	⋮	⋮	⋮
i	1	y_{i1}	x_{i11}	...	x_{i1p}
i	2	y_{i2}	x_{i21}	...	x_{i2p}
⋮	⋮	⋮	⋮	⋮	⋮
i	n_i	y_{in_i}	x_{in_i1}	...	x_{in_ip}
⋮	⋮	⋮	⋮	⋮	⋮
n	1	y_{n1}	x_{n11}	...	x_{n1p}
n	2	y_{n2}	x_{n21}	...	x_{n2p}
⋮	⋮	⋮	⋮	⋮	⋮
n	n_n	y_{nn_n}	x_{nn_n1}	...	x_{nn_np}

2.2 Modelos Lineares Generalizados

Os modelos lineares generalizados (MLG) constituem uma extensão dos modelos lineares clássicos, e foram apresentados por Nelder e Wedderburn (1972). Dada uma variável resposta y e um conjunto de variáveis explicativas $x_1, \ldots, x_p$, um MLG assume uma distribuição da família exponencial para a variável resposta y e especifica uma relação entre uma função da média de y com uma função linear das variáveis x.

Um modelo linear generalizado para a relação da esperança de uma variável aleatória y consiste de três componentes:

- **Componente aleatório** - Supõe-se que observações independentes são feitas sobre n variáveis aleatórias $y_1, \ldots, y_n$ que possuem uma determinada distribuição pertencente à familia exponencial. A função densidade de probabilidade (do tipo

contínu ou discreto) de y_i é dada por:

$$f(y_i;\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right\}, \qquad i = 1,2,\ldots,n. \tag{2.1}$$

Em (2.1) $a_i(.)$, $b(.)$ e $c(.)$ são funções reais conhecidas, θ_i é o parâmetro de locação e ϕ o parâmetro de dispersão. De (2.1) decorre que $\mathrm{E}(y) = \mu = b'(\theta)$ como será provado adiante em (2.2).

- **Componente sistemático** - Refere-se ao conjunto de variáveis explicativas $\mathbf{x}_i$, que produzem um preditor linear η_i, i=1,2,...,n,

$$\eta_i = \mathbf{x}_i^T\boldsymbol{\beta},$$

onde η_i é o preditor linear; $\mathbf{x}_i = (x_1, x_2, \ldots, x_p)^T$ é um vetor (p x 1) de variáveis explicativas para a observação i e $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ é o vetor (p x 1) dos parâmetros a serem estimados.

- **Função de ligação** - É uma função monotônica diferenciável que relaciona o valor esperado da variável resposta com o preditor linear

$$g(\mu_i) = \eta_i.$$

O preditor linear η_i pode assumir qualquer valor real o que não ocorre sempre com μ pois isto depende da distribuição de y_i. Portanto a função $g(\cdot)$ tem que ser definida no conjunto de valores possíveis para μ e tomar valores em $\Re$. Por exemplo, no modelo normal linear a média (μ_i) e o preditor linear (η_i) podem ser idênticos, dado que μ_i e η_i podem assumir qualquer valor na reta real $(-\infty, +\infty)$; sendo assim, uma ligação do tipo $\mu_i = \eta_i$ é plausível para modelar dados que seguem distribuição normal. Se y_i tem distribuição de Poisson, sua média μ é sempre positiva e uma função de ligação adequada $\mathrm{g}(\mu) = \log(\mu) = \eta$ pois a função logarítmica tem domínio no conjunto dos números reais positivos e assume qualquer valor real.

Em um MLG que assume que y_i tenha distribuição binomial com parâmetros n e π, $n = 1, 2, \ldots$ e $0 < \pi < 1$, $\mathrm{E}(y_i) = n\pi$. Uma função da média pode ser dada por $\mathrm{g}(\pi) = \mathrm{g}(n\pi) = n\mathrm{g}^*(\pi)$. O domínio da função $\mathrm{g}^*(\pi)$ é o o intervalo (0,1). Os três principais modelos de função usados para $\mathrm{g}^*(\pi)$ são

(i) Função de ligação "logito"

$$\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right).$$

(ii) Função de ligação "probito"

$$\eta_i = \Phi^{-1}(\pi_i),$$

onde $\Phi^{-1}(\cdot)$ é a função de distribuição acumulada da normal padrão;

(iii) Função de ligação "logaritmo do complemento do logaritmo"

$$\eta_i = \log(\log(1-\pi_i)).$$

Uma discussão com mais detalhes sobre função de ligação e suas propriedades pode ser vista em Firth (1991), como citado em Cordeiro e Neto (2004).

Uma propriedade atrativa dos modelos lineares generalizados é a possibilidade de ajustar modelos de regressão quando a variável resposta é normal, normal inversa, gama, Poisson, binomial, binomial negativa e geométrica, através da escolha apropriada da função de ligação $g(\cdot)$.

A esperança e a variância da variável y_i é dada por $b'(\theta_i)$ e $a_i(\phi)b"(\theta_i)$, respectivamente. Estes resultados são obtidos através da resolução das equações

$$\mathrm{E}\left(\frac{\partial l(\mathbf{y};\boldsymbol{\theta})}{\partial\theta}\right) = 0, \tag{2.2}$$

e

$$\mathrm{E}\left(\frac{\partial^2 l(\mathbf{y};\boldsymbol{\theta})}{\partial\theta^2}\right) + \mathrm{E}\left(\frac{\partial l(\mathbf{y};\boldsymbol{\theta})}{\partial\theta}\right)^2 = 0. \tag{2.3}$$

Uma forma simples de fazer isto é lembrando que a integral da função densidade, quando for do tipo contínua, dada em 2.1 sobre $\Re$ é 1. Desta forma, tem-se que

$$\int_{\Re}\exp\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right)dy_i = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)}\right). \tag{2.4}$$

Derivando em relação a θ_i e supondo ser possível mudar a ordem de diferenciação e integração, obtém-se

$$\int_{\Re}\frac{y_i}{a_i(\phi)}\exp\left(\frac{y_i\theta_i}{a_i(\phi)} + c(y_i,\phi)\right)dy_i = \frac{b'(\theta_i)}{a_i(\phi)}\exp\left(\frac{b(\theta_i)}{a_i(\phi)}\right). \tag{2.5}$$

Derivando novamente em relação a θ_i, obtém-se

$$\int_{\Re}\left(\frac{y_i}{a_i(\phi)}\right)^2\exp\left(\frac{y_i\theta_i}{a_i(\phi)}+c(y_i,\phi)\right)dy_i=\frac{b''(\theta_i)}{a_i(\phi)}\exp\left(\frac{b(\theta_i)}{a_i(\phi)}\right)+\left(\frac{b'(\theta_i)}{a_i(\phi)}\right)^2\exp\left(\frac{b(\theta_i)}{a_i(\phi)}\right) \tag{2.6}$$

De (2.5) segue que

$$\int_{\Re}\frac{y_i}{a_i(\phi)}a_i(\phi)\exp\left(\frac{y_i\theta_i-b(\theta_i)}{a_i(\phi)}+c(y_i,\phi)\right)dy_i=b'(\theta_i).$$

Logo $\mathrm{E}(y_i)=b'(\theta)$.

De (2.6) segue que

$$\int_{\Re}\left(\frac{y_i}{a_i(\phi)}\right)^2\exp\left(-\frac{b(\theta_i)}{a_i(\phi)}\right)\exp\left(\frac{b(\theta_i)}{a_i(\phi)}+c(y_i,\phi)\right)dy_i=\frac{b''(\theta_i)}{a_i(\phi)}+\left(\frac{b'(\theta_i)}{a_i(\phi)}\right)^2,$$

e,

$$\int_{\Re}(a_i(\phi))^2\left(\frac{y_i}{a_i(\phi)}\right)^2\exp\left(-\frac{b(\theta_i)}{a_i(\phi)}\right)\exp\left(\frac{b(\theta_i)}{a_i(\phi)}+c(y_i,\phi)\right)dy_i=a_i(\phi)b''(\theta_i)+b'(\theta_i)^2. \tag{2.7}$$

De (2.7) segue que $\mathrm{E}(y_i^2)=b''(\theta_i)a_i(\phi)+b'(\theta_i)^2$. Portanto, $\mathrm{V}(y_i)=b''(\theta_i)a_i(\phi)$.

Resumindo, tem-se que a média é dada por $E(y_i)=b'(\theta_i)$ e a variância $V(y_i)=a(\phi)b''(\theta_i)$. Observa-se também que esta variância é resultado do produto das funções $b''(\theta_i)$, que depende apenas do parâmetro canônico θ_i, e $a(\phi)$, que depende de ϕ, Demétrio (2002) e Paula (2004).

2.3 Exemplos de distribuições da família exponencial

Os dois exemplos a seguir apresentam a forma exponencial canônica e as expressões para a média e variância da distribuição normal e Binomial.

2.3.1 Exemplo 1. A Distribuição Normal como um membro da família Exponencial

A função densidade de probabilidade da distribuição Normal de parâmetros μ e σ é dada para $-\infty < y < \infty$ por:

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right\}. \tag{2.8}$$

Na forma canônica, a função densidade da $\mathrm{N}(\mu,\sigma^2)$ é dada por

$$f(y;\mu,\sigma^2) = \exp\left\{\frac{1}{\sigma^2}\left(y\mu - \frac{\mu^2}{2}\right) - \frac{1}{2}\left[\frac{y^2}{\sigma^2} + \log\left(2\pi\sigma^2\right)\right]\right\},$$

Desta forma, a média e variância da distribuição Normal (μ,σ^2), obtidas a partir das relações da seção anterior são:

$$E(y) = b'(\theta) = \mu, \qquad \text{e} \qquad V(y) = a(\phi)b''(\theta) = \sigma^2.$$

2.3.2 Exemplo 2. Distribuição Binomial como membro da família exponencial

A função de probabilidade da Binomial(n,π) é dada por

$$f(y;\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y}I_{(0,\ldots,n)}(y), \tag{2.9}$$

onde $I_{(0,\ldots,n)}(y)$ é a função indicadora de $(0,\ldots,n)$. Para obter a forma canônica da família exponencial para esta função, basta escrevê-la como a exponencial de seu logaritmo e identificar os componentes.

$$f(y;\pi) = \exp\left\{y\log\pi + (n-y)\log(1-\pi) + \log\binom{n}{y}\right\},$$

e,

$$f(y;\pi) = \exp\left\{y\log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi) + \log\binom{n}{y}\right\}.$$

Seja $\theta = \log(\pi/1-\pi)$. Então $\pi = (e^\theta)/(1+e^\theta)$, $\quad b(\theta) = -n\log(1-\pi)$, $c(y;\phi) = \log\binom{n}{y}$ $\quad$ e $\quad$ $a(\phi) = 1$.

Calculando a primeira e segunda derivada de $b(\theta)$ obtemos

$$b'(\theta) = n\frac{e^\theta}{1+e^\theta} = n\pi,$$

e

$$b''(\theta) = n\frac{e^\theta(1+e^\theta) - e^\theta e^\theta}{(1+e^\theta)^2} = n\frac{e^\theta}{(1+e^\theta)^2} = n\pi(1-\pi).$$

2.4 Estimação de parâmetros por máxima verossimilhança através do método de Newton-Raphson

Formulado o modelo MLG, $\mathrm{g}(\mu_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$, para as médias de um conjunto de variáveis aleatórias independentes $y_1, \ldots, y_n$ com o mesmo tipo de função densidade (contínua ou discreta) de probabilidade pertencente à família exponencial, a estimação de seus parâmetros a partir de uma observação de $(y_1, \ldots, y_n)$ pode ser feita pelo método da máxima verossimilhança. Supondo $\phi_i = \phi$, para i= $1, \ldots, n$, e sendo $\mu = b'(\theta_i) = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$, a função de verossimilhança a ser maximizada é dada pela expressão

$$L(\mathbf{y};\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i;\theta_i,\phi) = \prod_{i=1}^{n} \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right\}. \tag{2.10}$$

A maximização da função (2.10) ocorre em pontos onde a derivada da função (2.10) se anula. Logo a busca da estimativa de máxima verossimilhança dos parâmetros se inicia pela procura de soluções do sistema de equações

$$\frac{d\mathrm{L}(\mathbf{y};\boldsymbol{\theta})}{d\boldsymbol{\theta}} = 0 \qquad \text{ou} \qquad \frac{d\mathrm{logL}(\mathbf{y};\boldsymbol{\theta})}{d\boldsymbol{\theta}} = 0. \tag{2.11}$$

Em vista do modelo para $\mu_i = \mathbf{x}_i^T\boldsymbol{\beta} = \mathrm{g}^{-1}(\theta)$, é possível reescrever (2.11) em termos dos parâmetros $\boldsymbol{\beta}$ e buscar as estimativas dos parâmetros $\boldsymbol{\beta}$ como soluções do sistema de equações

$$\frac{d\mathrm{L}(\mathbf{y};\boldsymbol{\beta})}{d\boldsymbol{\beta}} = 0 \qquad \text{ou} \qquad \frac{d\mathrm{logL}(\mathbf{y};\boldsymbol{\beta})}{d\boldsymbol{\beta}} = 0. \tag{2.12}$$

A equação obtida em, (2.11) não é linear e sua solução tem que ser buscada através de métodos numéricos. Na literatura podem ser encontrados alguns métodos de otimização, por exemplo, o método de Newton-Raphson, o método escore de Fisher, o método Simplex proposto por Nelder e Mead (1965), o método EM proposto por Laird

e Ware (1982) ou o método baseado em espaço de estados usando filtro de Kalman proposto por Jones (1993), citados em Rocha (2004).

O método de Newton-Raphson será utilizado na resolução da equação (2.11), por apresentar um tempo de convergência menor em relação aos demais métodos. Este método encontra a raiz v^r de uma equação $h(v) = 0$ usando iterativamente uma aproximação de Taylor para $h(v)$ quando v se encontra na vizinhança de um ponto v^m. No caso em que $h(\cdot)$ seja função de apenas uma variável real, unidimensional, sendo v^0 uma tentativa inicial para o valor da raiz v^r,

$$h(v) \approx h(v^0) + (v - v^0)f'(v^0) = 0,$$

obtendo-se

$$v \approx v^0 - \frac{h(v^0)}{h'(v^0)},$$

ou, de uma forma mais geral,

$$v^{(m+1)} \approx v^{(m)} - \frac{h(v^{(m)})}{h'(v^{(m)})}. \tag{2.13}$$

A seqüência de pontos (v^m) converge para a raiz da equação $h(v) = 0$. No caso em que $h(\cdot)$ é função de p variáveis, $p > 1$, isto é, $h(v_1, \ldots, v_p)$, a busca da raiz da equação $h(v) = 0$ por um processo iterativo de consiste em, a partir de um vetor inicial v^0, obter sucessivamente os valores v_m dados por

$$h(\boldsymbol{v}^{(m+1)}) \approx h(\boldsymbol{v}^{(m)}) + \frac{\partial h(\boldsymbol{v}^{(m)})}{\partial \boldsymbol{v}}[\boldsymbol{v}^{(m+1)} - \boldsymbol{v}^{(m)}]$$

ou

$$\boldsymbol{v}^{(m+1)} \approx \boldsymbol{v}^{(m)} - \left[\frac{\partial h(\boldsymbol{v}^{(m)})}{\partial \boldsymbol{v}}\right]^{-1} h(\boldsymbol{v}^{(m)}). \tag{2.14}$$

Usando o método de Newton-Raphson para a solução de (2.10), para o caso em que $v = \theta$ e $\mathrm{h}(v) = \mathrm{h}(\theta) = d\log(\mathrm{L}(y, \theta))/d\theta = \mathrm{U}(\theta)$ e usando ainda as restrições $\mathbf{x}_i^T\boldsymbol{\beta} = \mathrm{g}^{-1}(\theta)$,

$$U(\boldsymbol{\theta}^{(m+1)}) \approx U(\boldsymbol{\theta}^{(m)}) + \frac{\partial U(\boldsymbol{\theta}^{(m)})}{\partial \boldsymbol{\theta}}[\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}]$$

ou

$$\boldsymbol{\theta}^{(m+1)} \approx \boldsymbol{\theta}^{(m)} - \left[\frac{\partial U(\boldsymbol{\theta}^{(m)})}{\partial \boldsymbol{\theta}}\right]^{-1} U(\boldsymbol{\theta}^{(m)}). \tag{2.15}$$

Pode-se demonstrar que θ_m converge para a solução do sistema (2.10 ou 2.11) se $\left[\partial U(\boldsymbol{\theta}^{(m)})/\partial\boldsymbol{\theta}\right]^{-1}$ for substituído pela matriz de informação de Fisher, e, assim, o método é chamado de escore de Fisher. Neste caso, na mésima iteração,

$$\boldsymbol{\theta}^{(m+1)} \approx \boldsymbol{\theta}^{(m)} - (\boldsymbol{x}^T\mathbf{W}^{(m)}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{W}^{(m)}\mathbf{z}^{(m)}, \tag{2.16}$$

onde $z_i = (y_i - \mu_i)(\partial g(\mu_i)/\partial\mu_i)$ é o i-ésimo elemento do vetor $\mathbf{z}$ e $\mathbf{W}$ =diag$(\omega_1, \ldots, \omega_n)$ é uma matriz de pesos, com $\omega_i = (\partial\mu_i/\partial\eta_i)^2/v_i$, sendo que v_i é a função de variância para o i-ésimo indivíduo.

Colocando $(\mathbf{x}^T\mathbf{W}^{(m)}\mathbf{x})^{-1}$ em evidência tem-se,

$$\boldsymbol{\theta}^{(m+1)} \approx (\mathbf{x}^T\mathbf{W}^{(m)}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{W}^{(m)}\mathbf{y}^{*(m)}, \tag{2.17}$$

onde $\mathbf{y}^{*(m)}$ é uma variável resposta modificada denotada por

$$\mathbf{y}^{*(m)} = \boldsymbol{x}\boldsymbol{\theta}^{(m)} + \mathbf{z}^{(m)}.$$

Observa-se que cada iteração do método de Newton-Raphson corresponde a uma regressão ponderada da variável dependente modificada $\mathbf{y}^*$ sobre a matriz $\mathbf{x}$, com matriz de pesos $\mathbf{W}$, Cordeiro e Neto (2004) e Paula (2004).

A formulação de um MLG depende da escolha de uma distribuição de probabilidade ou densidade para a variável resposta. Para uma escolha adequada desta distribuição, é aconselhável examinar os dados de maneira a encontrar algumas características, tais como: assimetria, natureza discreta ou contínua, intervalo de variação, etc. Esta distribuição deve ser conhecida e pertencer à família exponencial. Associados à variável resposta y, há um conjunto de variáveis explicativas $x_1, \ldots, x_p$, que, podem influenciar a resposta através de um preditor linear.

Em 1974, Wedderburn propôs a metodologia de métodos de Quase-verossimilhança, que pode ser interpretada como uma generalização dos MLGs no sentido de assumir uma função de variância para a variável resposta bem como uma relação funcional entre a média e o vetor paramétrico $\boldsymbol{\beta}$. Com isso, não requerem mais o conhecimento da distribuição da resposta. Esta metodologia é aplicada para dados correlacionados, o que não era possível com os MLGs, que assumem que as respostas são independentes, Cordeiro e Neto (2004) e Paula (2004).

2.5 Métodos de Quase-verossimilhança

Os métodos de quase-verossimilhança necessitam apenas da existência dos dois primeiros momentos da distribuição da variável resposta y, sem ser necessário conhecer a forma da sua distribuição.

Suponha que y_i, $i = 1, 2, \ldots, n$, seja um conjunto de observações com $\mathrm{E}(y_i) = \mu_i$ e $\mathrm{V}(y_i) \propto \mathrm{V}(\mu_i)$, em que $\mathrm{V}(\mu_i)$ é alguma função conhecida de μ_i. Também suponha que μ_i seja uma função de um conjunto de parâmetros de interesse $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ e $\mathbf{x}_i = (x_1, x_2, \ldots, x_p)^T$. A função de quase-verossimilhança $\mathrm{Q}(y_i, \mu_i)$ é definida pela relação

$$\frac{\partial Q(y_i, \mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{V(\mu_i)}. \tag{2.18}$$

Weddeburn (1974) mostrou que se pode usar qualquer função $\mathrm{Q}(y_i, \mu_i)$ que satisfaça (2.18) como uma base para definir um modelo linear generalizado e obter estimativas de $\boldsymbol{\beta}$ pelo uso iterativo das equações de mínimos quadrados ponderados. Mais informações sobre estimação dos parâmetros deste método pode ser consultada em Paula (2004).

2.6 Função de estimação

Uma função de estimação é, de uma maneira simplificada, uma função $\psi_i(\mathbf{y}; \boldsymbol{\theta})$ de um vetor aleatório $\mathbf{y}$ e dos parâmetros de interesse $\boldsymbol{\theta}$. Em termos práticos, elas são construídas de modo que raízes $\boldsymbol{\theta}$ de $\psi_i(\mathbf{y}; \boldsymbol{\theta}) = \mathbf{0}$, quando existem, sejam estimativas dos parâmetros em estudo. Em geral, deseja-se a construção de estimadores consistentes e com distribuição conhecida, ao menos assintoticamente. Um ponto importante na construção dessas funções é o estabelecimento de condições que garantam que os estimadores obtidos possuam boas propriedades, Artes e Botter (2005).

Seja $y_1, y_2, \ldots, y_n$, uma amostra de uma variável aleatória y, para que cada y_i esteja associada uma função de estimação $\psi_i(y_i; \theta)$, $i = 1, 2, \ldots, n$. A função de estimação para a amostra é definida através de (2.19):

$$\boldsymbol{\Psi}(\mathbf{y}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{\psi}_i(y_i; \theta). \tag{2.19}$$

O estudo das propriedades de uma função de estimação requer algumas definições que são apresentadas a seguir.

- Uma função de estimação $\boldsymbol{\Psi}(\mathbf{y};\boldsymbol{\theta})$ é dita não viciada quando

$$\mathrm{E}_\theta(\boldsymbol{\Psi}(\mathbf{y};\boldsymbol{\theta})) = 0. \tag{2.20}$$

- Neste caso a matriz de variância é dada por

$$\mathrm{V}_\theta(\boldsymbol{\Psi}(\boldsymbol{y};\boldsymbol{\theta})) = \mathrm{E}_\theta(\boldsymbol{\Psi}(\boldsymbol{y};\boldsymbol{\theta})\boldsymbol{\Psi}^T(\boldsymbol{y};\boldsymbol{\theta})). \tag{2.21}$$

2.7 Equações de estimação generalizadas

O método de equações de estimação generalizadas (EEG), proposto por Liang e Zeger (1986), pode ser utilizado para analisar conjunto de dados onde a variável resposta é contínua ou discreta. As EEG são uma técnica de estimação que leva em consideração a correlação entre as variáveis, e que produzem estimadores consistentes e assintoticamente normais dos parâmetros sob a especificação correta da função de ligação e da variância em função da média, sem a necessidade de se conhecer totalmente a distribuição multivariada dos dados Baia (1997).

As EEG são uma extensão multivariada da função de quase-verossimilhança, apresentada por Weddeburn (1974), que não exige conhecimento da distribuição paramétrica da variável resposta, mas apenas especificar a relação entre a média e a variância das observações, supondo alguma estrutura de correlação para os dados.

Seja $y_1, y_2, \ldots, y_n$, onde y_i $(i = 1, 2, \ldots, n)$, tem distribuição de probabilidade pertencente à família exponencial e $\mathbf{x}_i$ uma matriz $(n \times p)$ de observações com p variáveis explicativas associadas ao i-ésimo indivíduo, para $i = 1, 2, \ldots, n$. Admite-se também que $\mathrm{E}(y_i) = \mu_i$, $\mathrm{V}(y_i) = \phi\upsilon(\mu_i)$ e $\mathrm{cor}(y_i) = \Gamma(\mu_i)$.

Para a modelagem de μ_i serão utilizadas as mesmas convenções usadas nos modelos lineares generalizados, isto é,

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta}.$$

A função de estimação para $\boldsymbol{\beta}$ é dada por $\boldsymbol{\Psi}(\boldsymbol{y};\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{\psi}_i(y_i;\beta)$, na qual

$$\psi_i(\beta) = \mathbf{D}^T\mathbf{W}^{-1}\mathbf{u}_i, \tag{2.22}$$

onde $\mathbf{D}^T = \mathbf{X}^T\mathbf{H}$, $\mathbf{H}$ =diag$(\partial\mu_1/\partial\eta_1, \ldots, \partial\mu_n/\partial\eta_n)$, $\mathbf{W}$ é uma matriz de pesos, com $\omega_i = (\partial\mu_i/\partial\eta_i)^2/v_i$, sendo que v_i é a função de variância para o i-ésimo indivíduo, e $\mathbf{u}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$.

Aproximando $\boldsymbol{\Psi}(\boldsymbol{\beta}) \approx 0$, obtém-se um estimador para o vetor $\boldsymbol{\beta}$, que sob condições gerais de regularidade, conforme apresentado na seção (2.6), é considerado ótimo e consistente, Artes e Botter (2005). Observe que agora $\boldsymbol{\Psi}(\boldsymbol{\beta})$, traz na sua estrutura a correlação serial proposta por Liang e Zeger (1986), e que será visto na seção (2.8).

2.8 Extensões dos MLG para dados longitudinais com distribuição normal

Diggle *et al.* (1996) apresentam três extensões dos modelos lineares generalizados para dados longitudinais, incorporando a dependência entre as observações ao longo do tempo. São eles: modelos marginais, modelos de transição (ou condicionais) e modelos de efeitos aleatórios.

2.8.1 Modelos marginais

As respostas são modeladas marginalmente em relação às demais respostas, observando os efeitos no conjunto e, associado a este modelo, há uma estrutura de correlação envolvida, pois para um mesmo indivíduo são feitas várias medidas. Segundo Diggle *et al.* (1996) este modelo é capaz de modelar separadamente o efeito das variáveis explicativas na esperança da variável resposta, ou seja, esta esperança individual, E(y_{ij}), é expressa em função de $\mathbf{x}_{ij}^T\boldsymbol{\beta}$, onde $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$, $p < n$, é o vetor dos parâmetros da regressão a serem estimados.

A equação que representa o modelo é

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \epsilon_{ij}, \qquad i = 1, 2, \ldots, n, \qquad j = 1, 2, \ldots, n_i, \tag{2.23}$$

onde y_{ij}, representa a resposta para o i-ésimo individuo observado no j-ésimo tempo, $\mathbf{x}_{ij}$ representa um vetor (p x 1) de variáveis explicativas, $\boldsymbol{\beta}$ representa o vetor (p x

1) de coeficientes e ϵ_{ij} o erro aleatório responsável pela natureza estocástica da variável resposta. Os coeficientes deste modelo apresentam interpretação similar aos dos coeficientes em um modelo de regressão linear, ou seja, o quanto varia a média de y_{ij} para um aumento de uma unidade da variável $\mathbf{x}_{ij}$. Além disso, tem-se o interesse nestes coeficientes considerando uma estrutura de correlação para o vetor de respostas individual. Sendo assim, as suposições necessárias para o modelo marginal são:

(i) A esperança marginal da variável resposta, $\mathrm{E}(y_{ij}) = \mu_{ij}$, depende das variáveis explicativas $\mathbf{x}_{ij}$ através da relação $g(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta}$, onde $g(\cdot)$ é a função de ligação definida por

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

onde $\eta_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta}$ é o preditor linear e $g(\cdot)$ uma função monótona e diferenciável.

(ii) A variância marginal, depende da média marginal através da relação

$$\mathrm{V}(y_{ij}) = \upsilon(\mu_{ij})\phi,$$

onde $\upsilon(\cdot)$ é a função de variância conhecida e ϕ o parâmetro de dispersão;

(iii) A correlação entre y_{ij} e y_{ik} pode ser dada, às vezes, por parâmetros adicionais α, isto é,

$$\mathrm{corr}(y_{ij}, y_{ik}) = \rho(\mu_{ij}, \mu_{ik}; \alpha),$$

onde $\rho(\cdot)$ é uma função conhecida e α é uma matriz adicional utilizada para medir a dependência intra-indivíduos, Diggle *et al.* (1996).

Porém, quando as suposições de independências não são satisfeitas, uma alternativa, é aplicar o método de estimação de equações generalizadas (EEG), visto na seção (2.7). Originalmente, a teoria das EEG corrige os problemas de ordem prática associados à quase-verossimilhança, por exemplo, correlação serial intra-indivíduos, Artes e Botter (2005).

Com o método EEG, a dependência das observações de cada indivíduo é modelada através de uma matriz de covariância de $\mathbf{y}_i$, denotada por $\mathbf{V}_i$, dada por

$$\mathrm{cov}(\mathbf{y}_i) = \mathbf{V}_i = \mathbf{A}_i^{\frac{1}{2}}\mathbf{R}_i(\alpha)\mathbf{A}_i^{\frac{1}{2}}, \tag{2.24}$$

onde $\mathbf{A}_i$ = matriz diag $(v(\mu_{i1}), v(\mu_{i2}), \ldots, v(\mu_{ij}))$, que define a variância de y_{ij} como função da média marginal μ_{ij}, $\mathbf{R}_i(\boldsymbol{\alpha})$ é chamada matriz de correlação de "trabalho", que depende do vetor de parâmetros α, e permite incorporar aos modelos marginais diferentes estruturas de correlação. Quando a estrutura de correlação definida pela matriz de "correlação de trabalho", coincide com a verdadeira estrutura, os estimadores de $\boldsymbol{\beta}$ apresentam propriedades ótimas, Costa (2003).

As estruturas de correlação mais comuns são apresentadas a seguir.

Estruturas de covariâncias

Liang e Zeger (1986), sugeriram diferentes modelos para a estrutura de correlação entre as observações de um mesmo indivíduo. Isto implicou numa grande facilidade de interpretação das covariâncias dos modelos de regressão para este segmento.

A seguir são mostradas algumas das principais estruturas de correlação utilizadas nestes modelos para $n_i = 4$.

- **Independência:** Quando a matriz de correlação $\mathbf{R}_i(\boldsymbol{\alpha})$ é a identidade, isto é,

$$\mathbf{R}_i(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

  as observações são não correlacionadas.

- **Uniforme (Exchangeable - EX)**: Essa estrutura assume que a $\text{corr}(y_{ik}, y_{ik'}) = \begin{cases} ^{1, \text{ se } k=k'} \\ \alpha, \text{ se } k \neq k' \end{cases}$, ou seja,

$$\mathbf{R}_i(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{bmatrix}.$$

- **Não-estruturada:** Uma matriz de correlação é dita não-estruturada, quando apresenta todos os valores de α completamente não especificados, e, como conseqüência haverá $(k(k-1))/2$ parâmetros de correlação a serem estimados, ou

seja,

$$\mathbf{R}_i(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & \alpha_{21} & \alpha_{31} & \alpha_{41} \\ \alpha_{21} & 1 & \alpha_{32} & \alpha_{42} \\ \alpha_{31} & \alpha_{32} & 1 & \alpha_{43} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & 1 \end{bmatrix}.$$

Essa estrutura é útil somente quando há poucos tempos de observação e muitas unidades de corte transversal.

- **Auto-regressiva (AR-1):** seja $\text{Corr}(y_{ik}, y_{ik'}) = \alpha^{|k-k'|}$, logo

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

Em uma estrutura de correlação auto-regressiva, as correlações dependem das distâncias entre os tempos em que são tomado as as medidas, diminuindo com o aumento das distâncias.

A Tabela (2.2) mostra os estimadores para α utilizados nas matrizes de correlação de trabalho descritas anteriormente e sugeridas por Saavedra (2006).

Tabela 2.2: Estimadores α para a matriz correlação de trabalho.

Estrutura de R	α_{jk} =**corr**(y_{ij}, y_{ik})	**Estimativas de corr**(y_{ij}, y_{ik})
Independente	0	
Uniforme (EX)	α	$\widehat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{n_i(n_i-1)} \sum_{j \neq k} \varepsilon_{ij}\varepsilon_{ik}$
AR(1)	$\alpha^{\lvert j-k \rvert}$	$\widehat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(n_i-1)} \sum_{j \leq n_i-1} \varepsilon_{ij}\varepsilon_{i,j+1}$
Não-estruturada	α_{jk}	$\widehat{\alpha}_{jk} = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{ij}\varepsilon_{ik}$

As estruturas de correlações apresentadas na Tabela (2.2) podem ser utilizadas tanto para dados contínuos quanto para dados discretos. O termo $\epsilon_{ij} = (y_{ij} - \mu_{ij})/(\sqrt{y_{ij}})$ representa o resíduo padronizado, Saavedra (2006).

Outros tipos de estruturas de variâncias e covariâncias podem ser encontrados com mais detalhes em Diggle *et al.* (1996).

2.8.2 Modelos de transição

Os modelos marginais, apesar de permitirem incorporar uma estrutura de correlação para os dados através da matriz de trabalho $\mathbf{R}_i(\boldsymbol{\alpha})$, não são suficientes para descrever toda a informação relacionada a um estudo longitudinal, por não captam todos os efeitos intra-indivíduos. Assim, pretende se mostrar uma relação de dependência da distribuição condicional da resposta atual no tempo j, y_{ij}, sobre as respostas anteriores $(y_{ij-1}, \ldots, y_{i1})$ e as variáveis explicativas $\mathbf{x}_{ij}$. Uma das vantagens deste modelo é que modela as mudanças individuais no tempo e avalia como essas mudanças são influenciadas pelas variáveis explicativas consideradas, Lara (2007). Este modelo é uma extensão dos modelos lineares generalizados e é utilizado em situações onde a variável resposta atual tem forte ligação com as variáveis respostas anteriores. A função de densidade para o modelo de transição é dada por

$$f(y_{ij}|y_{ij-1}, \ldots, y_{i1}) = \exp\left\{\frac{y_{ij}\gamma_{ij} - \Phi(\gamma_{ij})}{\phi} + c(y_{ij}, \phi)\right\}, \qquad i = 1, 2, \ldots, n, \quad (2.25)$$

onde $\Phi(\gamma_{ij})$ é uma função semelhante à $b(\gamma_{ij})$ do modelo linear generalizado, adicionando a influência das respostas anteriores e $c(y_{ij}, \phi)$ é uma função de depende de y_{ij} e do parâmetro de dispersão ϕ.

A média e variância condicionais de y_{ij} são dadas por

$$\mu_{ij}^c = \mathrm{E}(y_{ij}|y_{ij-1}, \ldots, y_{i1}) = \Phi'(\gamma_{ij})$$

e

$$v_{ij}^c = \mathrm{V}(y_{it}|y_{ij}|y_{ij-1}, \ldots, y_{i1}) = \Phi''(\gamma_{ij})\phi.$$

Um modelo de regressão condicional é definido por

$$h(\mu_{ij}^c) = \mathbf{x}_{ij}'\boldsymbol{\beta} + \sum_{r=1}^{s} f_r(y_{ij-1}, \ldots, y_{i1}; \boldsymbol{\alpha}), \quad (2.26)$$

onde $f_r(\cdot)$ é uma função conhecida, e $\mathrm{h}(\mu_{ij}^c)$ é uma função de ligação. Note que quando $\sum_{r=1}^{s} f_r(y_{ij-1}, \ldots, y_{i1}; \alpha) = 0$, $\mathrm{h}(\mu_{ij}^c) = \mathrm{g}(\mu_{ij})$ pode-se aplicar a metodologia de modelos lineares generalizados.

Para se estimar os parâmetros do modelo de transição, utiliza-se o método de máxima verossimilhança, Diggle *et al.* (1996). Inicialmente, escreve-se a distribuição

conjunta das respostas $y_{i1}, y_{i2}, \ldots, y_{ij}$ na forma

$$f(y_{i1}, y_{i2}, \ldots, y_{ij}) = f(y_{ij}|y_{ij-1}, \ldots, y_{i1}) f(y_{ij-1}|y_{ij-2}, \ldots, y_{i1}) \ldots f(y_{i2}|y_{i1}) f(y_{i1}). \tag{2.27}$$

Em seguida, encontra-se a função de verossimilhança

$$\mathrm{L}_i(\mathbf{y}; \boldsymbol{\beta}) = f(y_{i1}) \prod_{j=2}^{n_i} f(y_{ij}|y_{ij-1}). \tag{2.28}$$

Maximizando a função (2.28) via método iterativo, por exemplo, escore de Fisher, pode-se encontrar as estimativas para os parâmetros do modelo de transição.

O estudo dos modelos de transição está baseado na teoria dos processos estocásticos e, freqüentemente a propriedade markoviana é a mais utilizada. Em alguns livros este modelo é encontrado como modelo condicional.

2.8.3 Modelos mistos

O modelo de regressão com efeitos aleatórios costuma ser conhecido como modelos misto, porque traz em sua estrutura coeficientes de regressão compostos de uma parte fixa (entre-indivíduo) e outra aleatória (variação no intercepto e inclinação individual, por exemplos). É um modelo que incorpora a dependência e a estrutura de correlação dos erros e supõe que os coeficientes da regressão variem entre os indivíduos. Seu uso é especialmente adequado para dados em que a variabilidade entre os indivíduos é maior do que a variabilidade dentro do indivíduo, Rocha (2004).

Tais modelos têm sido freqüentemente usados na análise de medidas repetidas, dados agrupados e dados longitudinais. Apresentam uma grande aplicabilidade em diversas áreas de pesquisa como agricultura, biologia e economia, Diggle *et al.* (1996). No contexto deste trabalho, a importância destes modelos é explicada pela flexibilidade que eles oferecem para modelar a correlação entre e/ou intra-indivíduos, freqüentemente presente em estudos longitudinais, Laird e Ware (1982).

Os modelos mistos apresentam alguns casos particulares, tais como: o modelo linear clássico, o modelo de componentes de variância e os modelos hierárquicos (multiníveis), Natis (2002). O modelo misto é encontrado na literatura, para respostas contínuas, na forma

$$\mathbf{y}_i = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{Z}_i^T \mathbf{b}_i + \epsilon_i, \qquad i = 1, 2, ..., n, \tag{2.29}$$

onde $\mathbf{y}_i$ (n_i x 1) é a resposta do i-ésimo indivíduo, $\mathbf{x}_i$ é a matriz de dimensão (n_i x p) de variáveis explicativas associados aos efeitos fixos $\boldsymbol{\beta}$, $\mathbf{Z}_i$ é a matriz de dimensão (n_i x q) de variáveis explicativas associados aos efeitos aleatórios $\mathbf{b}_i$ (q x 1) e $\boldsymbol{\epsilon}_i$ (n_i x 1) é o vetor de erros aleatórios. Geralmente $\mathbf{Z}_i$ é uma sub-matriz de $\mathbf{x}_i$.

Em geral, supõe-se que tanto os erros aleatórios como os efeitos aleatórios são normalmente distribuídos, ou seja,

$$\mathbf{b}_i \sim \mathbf{N}(\mathbf{0}, \mathbf{D}), \tag{2.30}$$

e

$$\boldsymbol{\epsilon}_i \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}_i), \tag{2.31}$$

onde $\boldsymbol{\Sigma}_i = \sigma^2 I$.

Um caso particular importante de (2.29), ocorre quando o modelo apresenta apenas o intercepto aleatório, ou seja,

$$\mathbf{y}_i = \mathbf{x}_i^T \boldsymbol{\beta} + \boldsymbol{\zeta}_i + \epsilon_i, \qquad i = 1, 2, ..., n, \tag{2.32}$$

onde $\boldsymbol{\zeta}_i$ representa o intercepto aleatório.

Os efeitos aleatórios no intercepto representam a heterogeneidade natural entre os indivíduos decorrente de fatores não medidos.

No modelo (2.29) a

$$\mathrm{E}(\mathbf{y}_i) = \mathbf{x}_i^T \boldsymbol{\beta} \qquad \text{e} \qquad \mathrm{V}(\mathbf{y}_i) = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \boldsymbol{\Sigma}_i.$$

onde $\boldsymbol{D}$ e $\boldsymbol{\Sigma}_i$ são desconhecidos e, que podem ser substituídos pelas matrizes estimadas $\mathbf{G}$, referente aos efeitos aleatórios e $\mathbf{R}_i$, correspondente a correlação serial. Estas matrizes são obtidas do processo de ajuste do modelo.

No ajuste do modelo misto é preciso avaliar qual é a estrutura de covariância que melhor se adapta aos dados, pois nesta estrutura serão incorporados os efeitos fixos e efeitos aleatórios associados aos indivíduos. A seguir mostra-se apenas dois exemplos destas estruturas que serão utilizadas neste trabalho, a saber:

- **Uniforme**: apresentam homogeneidade tanto nas variâncias quanto nas covariâncias.

$$\mathrm{V}(\mathbf{y}_i) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \mathbf{\Sigma}_i = \begin{bmatrix} \sigma^2+\tau & \tau & \tau & \tau \\ \tau & \sigma^2+\tau & \tau & \tau \\ \tau & \tau & \sigma^2+\tau & \tau \\ \tau & \tau & \tau & \sigma^2+\tau \end{bmatrix}.$$

Neste caso, temos que $\mathbf{D} = \tau$ com $\tau > 0$, e $\mathbf{Z}_i = \mathbf{1}_{n_i}$ é um vetor (n_i x n_i) com todos os elementos iguais a 1 e $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}_{n_i}$.

- **Auto-regressiva (AR-1):**

$$\mathrm{V}_i(\mathbf{y}_i) = \begin{bmatrix} 1 & \phi & \phi^2 & \phi^3 \\ \phi & 1 & \phi & \phi^2 \\ \phi^2 & \phi & 1 & \phi \\ \phi^3 & \phi^2 & \phi & 1 \end{bmatrix}.$$

Nesta estrutura tem-se $\mathbf{D} = 0$ e $\mathbf{\Sigma}_i$ é uma matriz gerada por um modelo em que os erros aleatórios das medidas realizadas no i-ésimo indivíduo têm a seguinte estrutura,

$$e_{ij} = \phi e_{ij-1} + \delta_{ij},$$

em que $\delta_{ij} \sim \mathrm{N}(0, \tau^2)$ são não correlacionados com $e_{il}, l = 1, 2, \ldots, j-1$. Esta estrutura é função de $\theta = (\phi, \sigma^2)$, com $\sigma^2 = \tau^2/(1-\phi^2)$ e $|\phi| < 1$ para garantir a estacionaridade.

Outros tipos de estruturas de covariâncias podem ser vistas com mais detalhes em Rocha (2004).

Para ajustar um modelo misto é comum encontrar na literatura algumas alternativas: aproximações por métodos Bayesianos, estimação por máxima verossimilhança (EMV) e estimação por máxima verossimilhança restrita, Verbeke e Molenberghs (2000).

A função de verossimilhança para o modelo misto para o i-ésimo indivíduo é dada por

$$L_i(\mathbf{y}_i, \boldsymbol{\beta}, \mathbf{b}_i) = \int_{\Re} \prod_{j=1}^{n_i} f(\mathbf{y}_{ij}|\mathbf{b}_i) f(\mathbf{b}_i) d\mathbf{b}_i, \tag{2.33}$$

onde $f(\mathbf{y}_{ij}|\mathbf{b}_i)$ é a função densidade de probabilidade associada ao i-ésimo indivíduo no tempo j condicionada aos efeitos aleatórios $\mathbf{b}_i$, e $f(\mathbf{b}_i)$ é a função densidade de probabilidade dos efeitos aleatórios, Saavedra (2006).

A função de verossimilhança para o conjunto de todos os n indivíduos é

$$\mathrm{L}(\mathbf{y}, \boldsymbol{\beta}, \mathbf{b}) = \prod_{i=1}^{n} \mathrm{L}_i(\mathbf{y}_i, \boldsymbol{\beta}, \mathbf{b}_i) = \prod_{i=1}^{n} \int_{\Re} \prod_{j=1}^{n_i} f(y_{ij}|\mathbf{b}) f(\mathbf{b}) d\mathbf{b}. \tag{2.34}$$

Uma solução para isto é utilizar métodos de iteração numérica, por exemplo, algoritmo EM e/ou soluções através do escore de Fisher. As equações a seguir mostram as estimativas iniciais dos parâmetros do modelo quando se usa o algoritmo EM.

$$\hat{\boldsymbol{\beta}} = [\sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i]^{-1} [\sum_{i=1}^{n} \mathbf{x}_i^T (\mathbf{y}_i - \mathbf{1}_i \tilde{\mathbf{v}}_\mathbf{i})], \tag{2.35}$$

$$\hat{\boldsymbol{\sigma}}^2 = \frac{1}{n} \sum_{i=1}^{n} \tilde{\mathbf{v}}_\mathbf{i} + \boldsymbol{\sigma}^2_{v|y_i}, \tag{2.36}$$

$$\boldsymbol{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}} - \mathbf{1}_i \tilde{\mathbf{v}}_\mathbf{i})^T (\mathbf{y}_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}} - \mathbf{1}_i \tilde{\mathbf{v}}_\mathbf{i}) + n_i \boldsymbol{\sigma}^2_{v|y_i}, \tag{2.37}$$

em que,

$$\tilde{\boldsymbol{\sigma}}^2 = \rho_{n_i n_i} \frac{1}{n_i} \mathbf{1}_i^T (\mathbf{y}_i - \mathbf{x}_i^T \boldsymbol{\beta}) = \rho_{n_i n_i} \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}); \boldsymbol{\sigma}^2_{v|y_i} = \sigma_v^2 (1 - \rho_{n_i n_i}),$$

onde $\mathbf{x}_i$ é um vetor de covariáveis para o indivíduo i no tempo j, $\rho_{n_i n_i} = n_i r/[1 + (n-1)r]$ e r é igual a correlação intraclasse. O processo iterativo termina quando ocorre a convergência no algoritmo EM. Mais detalhes sobre o algoritmo EM podem ser encontradas em Liu & Rubin (1994), McLachlan & Krishnan (1997) e Meng & Van Dyk (1998) citados em Nobre (2004).

Capítulo 3

Dados Binários

É bastante comum que os estatísticos se deparem, nas suas análises, com conjuntos de dados onde a variável resposta assume dois estados, fracasso ou sucesso. Algumas das áreas de aplicação onde se pode encontrar tal situação são: a medicina, as finanças, as ciências sociais, a indústria, entre outras, Mills (2000). A seguir são citados alguns exemplos:

- Um estudo foi realizado para avaliar a relação entre o aparecimento de doenças cardíacas e outras variáveis, tais como: idade, sexo, hábito de fumar, nível de colesterol, peso e pressão sanguínea. A variável resposta foi definida com dois possíveis resultados: o indivíduo desenvolveu ou não desenvolveu a doença cardíaca durante o estudo. Estes resultados podem ser codificados por 1 e 0 respectivamente, ou vice-versa.

- Para analisar a relação entre a ocorrência de infecção hospitalar e outras variáveis, foi feito um estudo com vários hospitais com o levantamento de dados como o tempo de internação dos pacientes, idade média dos pacientes, número de camas no hospital e sua região geográfica. A variável resposta y, pode ser definida como:

$$y=\begin{cases} 1, \text{ se o hospital apresenta risco de infecção hospitalar,} \\ 0, \text{ se o hospital não apresenta risco de infecção hospitalar.} \end{cases}$$

- Num estudo sobre a participação de mulheres no mercado de trabalho, como função da idade, número de filhos e renda, pode-se definir a variável resposta y como:

$$y=\begin{cases} 1, \text{ a mulher participa do mercado de trabalho,} \\ 0, \text{ a mulher não participa do mercado de trabalho.} \end{cases}$$

- Em marketing, deseja-se saber se alguém comprará ou não um carro na chegada de um novo ano. Aqui os preditores tais como renda anual, número de dependentes, valor da prestação do financiamento da casa, e assim por diante, são preditores relevantes, David (1999).

Estes exemplos dão uma idéia da grande variedade de aplicações onde a variável resposta tem dois resultados possíveis e que pode ser representada por uma variável binária.

Antes de descrever a metodologia para dados binários em estudos longitudinais, uma revisão sobre regressão logística será apresentada, pois estes conceitos serão necessários mais adiante.

3.1 Regressão logística

A regressão logística é uma ferramenta de análise estatística que vem se tornando muito utilizada pelos estatísticos na modelagem de dados com resposta binária, quanto a relação com uma ou mais variáveis explicativas, sendo que estas podem ser qualitativas ou quantitativas. A regressão logística é um caso particular de modelos lineares generalizados, McCullagh e Nelder (1989).

Seja uma amostra aleatória de n indivíduos, para cada um dos quais existe uma resposta associada a y_i dada por

$$y_i=\begin{cases} 1, \text{ se a resposta do } i\text{-ésimo indivíduo é "sucesso",} \\ 0, \text{ se a resposta do } i\text{-ésimo indivíduo é "fracasso".} \end{cases}$$

Supondo que y_i tenha distribuição Bernoulli com probabilidade de sucesso π_i e que para cada um dos n indivíduos haja observações sobre p variáveis explicativas, $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})$. O modelo de regressão logística é dado por

$$\mathrm{E}(y_i|\mathbf{x}_i) = P(y_i = 1) = \pi_i = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \tag{3.1}$$

e, assim

$$\mathrm{P}(y_i = 0) = 1 - \pi_i = \frac{1}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}}, \tag{3.2}$$

sendo $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ o vetor de parâmetros do modelo. Das equações (3.1) e (3.2), aplicando o logaritmo na razão de π_i por $(1 - \pi_i)$, tem-se a seguinte função de ligação logito

$$g(\pi_i) = \log(\frac{\pi_i}{1 - \pi_i}). \tag{3.3}$$

Uma medida muito utilizada em diversas áreas do conhecimento é a denominada razão de chances (*Odds Ratio* (OR)). Se em (3.1) fizerem $\pi_i = \pi_i(x)$ e x associar apenas os valores 0 e 1, tem-se

$$\mathrm{OR} = \frac{\pi_i(1)/[1 - \pi_i(1)]}{\pi_i(0)/[1 - \pi_i(0)]}.$$

Por exemplo, se y_i representa a presença $y_i = 1$ ou ausência $y_i = 0$ de câncer no pulmão e $x_i = 1(0)$ representa se a pessoa é (não) fumante, um valor OR= 2 pode ser interpretado como: a chance de uma pessoa que fuma adquirir câncer no pulmão é duas vezes maior que a de uma pessoa que não fuma.

Outra medida utilizada em estudos prospectivos fornecendo o risco de desenvolvimento de uma determinada condição para um grupo quando comparado a outro é denominada risco relativo (RR) que é expresso por

$$\mathrm{RR} = \frac{\pi_i(1)}{\pi_i(0)}.$$

Os modelos de regressão logística podem ser usados para:

- Quantificar a importância da relação existente entre cada uma das covariáveis e a variável resposta, como também mostrar a existência de interação e efeito de confudimento com respeito à variável resposta.

- Classificar indivíduos dentro das categorias (presente/ausente) da variável resposta, segundo a probabilidade que tenha de pertencer a uma delas, dada a presença de determinadas covariáveis.

- Este modelo se diferencia dos modelos lineares clássicos quanto à sua apresentação gráfica da relação entre a experância da variável resposta e cada variável explicativa, pois tem a aparência de "S". A Figura (3.1) mostra o gráfico da função logística, que representa a forma funcional da relação entre a probabilidade de sucesso, $\mathrm{E}(y_i|x = x_i)$, e, uma variável explicativa x.

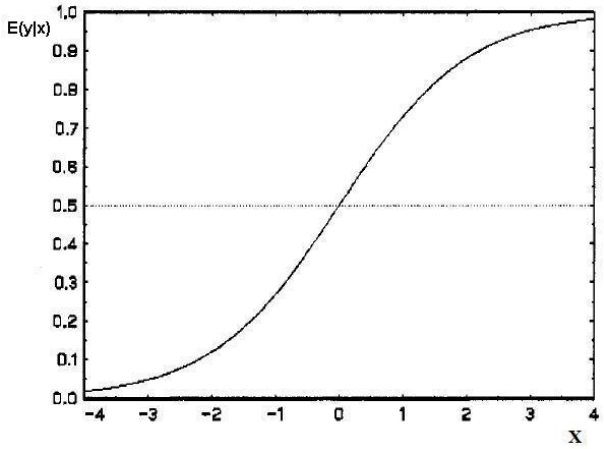

Figura 3.1: Função logística $\mathrm{E}(y_i|x = x_i)$.

- Na figura (3.2) é apresentado o gráfico da transformação *logit* (3.3) que lineariza a relação entre a esperança condicional da variável resposta $\mathrm{E}(y_i|x = x_i)$ e a variável explicativa x_i.

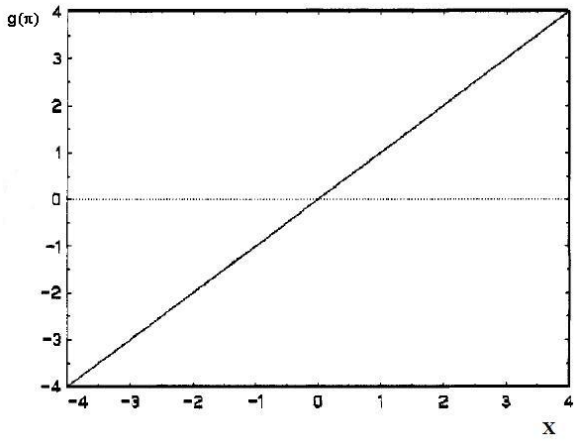

Figura 3.2: Transformação $\mathrm{g}(\pi_i)$.

Os parâmetros da regressão logística são geralmente estimados por máxima verossimilhança. A verossimilhança de uma amostra aleatória de n observações de variáveis binárias independentes de média π_i, $i = 1, 2, \ldots, n$ é dada por

$$\mathrm{L}(\mathbf{y}; \boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}. \tag{3.4}$$

Em (3.4), $\pi_i = e^{\mathbf{x}_i^T \boldsymbol{\beta}} / 1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}$, é a probabilidade de ocorrência de $y_i = 1$ com os valores amostrais das variáveis explicativas $\mathbf{x}_1, \ldots, \mathbf{x}_p$, para o i-ésimo indivíduo.

Aplicando uma transformação logarítma na função (3.4) obtém-se,

$$\mathrm{l}(\mathbf{y}; \boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \ln(\pi_i) + (n - \sum_{i=1}^{n} y_i) \ln(1 - \pi_i). \tag{3.5}$$

A seguir, calcula-se a função *score* obtida a partir da derivada primeira da função $\mathrm{l}(\mathbf{y}; \boldsymbol{\beta})$ com relação a $\boldsymbol{\beta}$. Assim, a função *score* é dada, na forma matricial, por

$$\mathrm{U}(\boldsymbol{\beta}) = \frac{\partial l(\mathbf{y}; \boldsymbol{\beta})}{\partial(\boldsymbol{\beta})} = \mathbf{x}^T(\mathbf{y} - \boldsymbol{\pi}). \tag{3.6}$$

A matriz de segundas derivadas parciais é conhecida como matriz de informação ou Hessiana, e é dada por

$$\mathrm{H}(\boldsymbol{\beta}) = \frac{\partial^2 l(\boldsymbol{\beta})}{\partial(\boldsymbol{\beta}) \partial(\boldsymbol{\beta})^T} = -\mathbf{x}^T \mathbf{W} \mathbf{x}, \tag{3.7}$$

sendo $\mathbf{W}$ uma matriz diagonal, (n x n), cujos elementos da diagonal principal são dados pelos produtos $\pi_i(1 - \pi_i)$,

$$\mathbf{W} = \begin{bmatrix} \pi_1(1-\pi_1) & 0 & \ldots & \ldots & 0 \\ 0 & \pi_2(1-\pi_2) & 0 & \ldots & 0 \\ \vdots & 0 & \ddots & 0 & \vdots \\ 0 & \ldots & 0 & \pi_{n-1}(1-\pi_{n-1}) & 0 \\ 0 & \ldots & \ldots & 0 & \pi_n(1-\pi_n) \end{bmatrix}.$$

O modelo de regressão logística é um caso particular dos modelos lineares generalizados e segue a mesma metodologia de estimação por máxima verossimilhança apresentada na seção (2.2).

3.2 Regressão logística para dados longitudinais

Neste seção, é abordada a modelagem de dados que apresentam estrutura longitudinal, quando a variável resposta é binária, usando a função de ligação logística, como descrito nos objetivos deste trabalho. Assim, pode-se considerar diversos modelos para explicar a relação dos dados.

3.2.1 Modelo marginal

Suponha que n indivíduos são selecionados aleatoriamente de uma população, e que sobre cada indivíduo são colhidos observações em j momentos do tempo sobre uma variável resposta binária y, representando sucesso ($y = 1$) ou fracasso ($y = 0$), e sobre um vetor de variáveis explicativas $\mathbf{x}$. Supondo também que cada y_i tenha distribuição de Bernoulli de parâmetro (π_i) em M.L.G., uma função de ligação logito é dada por

$$\mathrm{E}(y_{ij}|\mathbf{x}_{ij}) = \pi_{ij} = \frac{e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}}, \tag{3.8}$$

onde $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos da regressão.

A variância de y_{ij} para o i-ésimo indivíduo no j-ésimo tempo é dada por

$$\mathrm{V}(y_{ij}) = \pi_{ij}(1 - \pi_{ij}),$$

e a correlação é

$$\mathrm{corr}(y_{ij}, y_{ij-1}) = \rho_{ij}.$$

Um estimador para o vetor de parâmetros $\boldsymbol{\beta}$, é obtida através da solução das equações de estimação generalizadas (EEG) apresentadas por Liang e Zeger (1986). A estimativa do vetor $\boldsymbol{\beta}$ é solução do sistema de equações escores, isto é,

$$\mathrm{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} D_i' V_i^{-1}(y_i - \pi_i) = 0, \tag{3.9}$$

onde $D_i' = \partial\pi_i/\partial\beta$ e V_i é a matriz diagonal de variâncias para o i-ésimo indivíduo. Note que quando V_i for uma matriz identidade, volta-se ao caso de MLG, ou seja, poderá se aplicar todas as suposições de independência nestas observações, Liang e Zeger (1986).

Para obter as estimativas destes parâmetros, faz-se necessário o uso de métodos iterativos, e, o processo finaliza quando a precisão atribuída ao processo iterativo é atingida.

3.2.2 Modelos de transição

Como visto na seção (2.3.2), modelos de transição, ou condicionais, a dependência da variável resposta atual (y_{ij}) no tempo j, $j = 1, 2, \ldots, n_i$, em relação às respostas $j-1$ tempos anteriores. No caso de variáveis binárias, ao invés de se estimar todas as probabilidades de transição em separado, procura-se modelar estas probabilidade de tal forma que as estimativas dos parâmetros obtidas possam ser interpretadas como os pesos que cada uma das variáveis explicativas exercem na estimação da probabilidade de transição Lara (2007).

A probabilidade condicional $P(y_j = b|y_{j-1} = a)$ denotada por π_{ba}, que é a probabilidade de ir do estado a para o estado b. Por exemplo, considere um processo estacionário markoviano $y_{i1}, y_{i2}, \ldots, y_{ij}$, em que os indivíduos são observados nos tempos definidos para o estudo, quanto a possuir ou não a característica de interesse Diggle *et al.*(1996). A matriz de probabilidade de transição de $y_j|y_{j-1}$ é denotada por

$$\mathrm{P}_i = \begin{bmatrix} 1-\pi_a & \pi_a \\ 1-\pi_b & \pi_b \end{bmatrix}$$

onde $\pi_b = P(y_j = b|y_{j-1} = a)$ para $a \in 0, 1$, que é a probabilidade de mudança de estado. Note que cada linha da matriz de transição tem soma igual a 1, ou seja, $\mathrm{P}(y_j = 0|y_{j-1} = b) + \mathrm{P}(y_j = 1|y_{j-1} = b) = 1$, Lara (2007).

Para se estimar os parâmetros do modelo de transição, utiliza-se o método de máxima verossimilhança. Este processo é análogo ao visto para o caso onde a variável resposta segue distribuição normal. Assim, a função de verossimilhança é

$$\mathrm{L}(\mathbf{y}_i; \boldsymbol{\pi}_i) = f(y_{i1}) \prod_{j=2}^{n_i} f(y_{ij}|y_{ij-1}). \tag{3.10}$$

Após maximizar a função (3.10) via processo iterativo, encontra-se as estimativas para $\boldsymbol{\pi}_i$, Saavedra (2006).

3.2.3 Modelos mistos

Para dados com resposta binária, há modelos que apresentam estas características, por exemplo, o modelo logístico pode conter dois efeitos aleatórios. Na literatura, por exemplo, pode-se encontrar o trabalho de Snijders e Bosker (1999), em que é

apresentado o modelo de regressão logística com efeitos aleatórios, com um resumo utilizando vários métodos de estimação para os parâmetros do modelo. Assim, conforme visto na seção (2.3.3), onde foi abordado o caso em que a variável resposta é contínua, uma solução foi a generalização do modelo marginal, combinando efeitos fixos e aleatórios.

Seja $y_1, y_2, \ldots, y_{n_i}$, uma amostra aleatória, onde cada y_i tem distribuição Bernoulli com probabilidade de sucesso π_i. O modelo misto é

$$\Phi(\pi_i) = \mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{Z}_i^T\mathbf{b}_i, \tag{3.11}$$

onde $\Phi(\cdot)$ é a função de ligação que engloba efeitos individuais fixos e aleatórios em $\mathbf{b}_i$ para o i-ésimo indivíduo.

Assim, as suposições do modelo misto com função de ligação logística e distribuição de Bernoulli são:

(i) A Esperança condicional é obtida de

$$\pi_i = \mathrm{E}(y_i|\mathbf{b}_i) = \frac{\exp(\Phi(\pi_i))}{1+\exp(\Phi(\pi_i))}. \tag{3.12}$$

Observe que a média condicional é uma função dos efeitos individuais, $\pi_i = f(\mathbf{b}_i)$, e que o valor de π_i, é obtido por

$$logit(\boldsymbol{\pi}_i) = \log\frac{\pi_i}{1-\pi_i} = \Phi(\pi_i) = \mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{Z}_i^T\mathbf{b}_i, i = 1, 2, \ldots, n, \tag{3.13}$$

(ii) O efeito aleatório $\mathbf{b}_i$ é normalmente distribuído:

$$\mathbf{b}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{D}).$$

Caso particular

Um caso particular do modelo misto é o que considera apenas o intercepto aleatório $b_i = \begin{pmatrix} b_{i0} \\ 0 \end{pmatrix}$. Desta forma, o modelo visto em (3.11), ficará com $Z_1^T = \mathbf{1}$:

$$\Phi(\pi_i) = \mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{b}_{i0}, \tag{3.14}$$

onde $\mathbf{b}_{i0} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

Outro ponto é encontrar uma estrutura de variância que seja adequada ao modelo. As estruturas apresentadas para o caso de variável resposta contínua, visto na seção

(2.3.3), serão utilizadas aqui e verificada seu comportamento quando a variável resposta é binária.

Para estimar os parâmetros do modelo, utiliza-se o método de máxima verossimilhança condicionado ao efeito aleatório, que é dada pela expressão

$$\mathrm{L}(\mathbf{y}_i|\mathbf{b}_i) = \prod_{j=1}^{n_i} \pi_{ij}^{y_{ij}} (1-\pi_{ij})^{1-y_{ij}} \tag{3.15}$$

A seguir apresentam-se as expressões utilizadas na estimação destes efeitos, e que são implementadas em algoritmos numéricos.

Observe a similaridade entre esta função de verossimilhança e a função de verossimilhança apresentada em (3.4). Lá tínhamos n observações e a função de verossimilhança era calculada sobre todos indivíduos. Agora, esta função é calculada para cada indivíduos nos n_i tempos.

Na tentativa de encontrar uma expressão que não dependa dos efeitos aleatórios, faz-se necessário uma nova função que seja obtida da integração em relação dos erros. Isto gerará uma probabilidade marginal para $\mathbf{y}_i$,

$$\mathrm{h}(\mathbf{y}_i) = \int_{b_i} L(\mathbf{y}_i|\boldsymbol{b}_i) g(b_i) db_i, \tag{3.16}$$

onde $\mathrm{g}(b_i) \sim \mathrm{N}(0, \sigma_v^2)$ representa a distribuição populacional dos efeitos aleatórios. Assim, pode-se escrever a verossimilhança para todo o conjunto de dados:

$$\mathrm{L} = \prod_{i=1}^{n} h(\mathbf{y}_i),$$

A derivada parcial do logaritmo de L em relação a um conjunto de parâmetros em $\eta = (\beta, \sigma_v^2)$ é:

$$\frac{\partial \log \mathrm{L}}{\partial \boldsymbol{\eta}} = \sum_{i=1}^{n} h^{-1}(\mathbf{y}_i) \frac{\partial h(\mathbf{y}_i)}{\partial \boldsymbol{\eta}}, \tag{3.17}$$

onde $\boldsymbol{\eta}$ representa ou efeito $\boldsymbol{\beta}$ ou parâmetro de variância $\boldsymbol{\sigma_v^2}$.

Encontradas as derivadas para a equação (3.17) é possível utilizar processos iterativos existentes na literatura, por exemplo, escore de Fisher, para encontrar as estimativas para o vetor de parâmetros $\boldsymbol{\beta}$, através da expressão

$$\boldsymbol{\beta}_{i+1} = \boldsymbol{\beta}_i + I(\boldsymbol{\beta}_i)^{-1} \frac{\partial \log \mathrm{L}}{\partial \boldsymbol{\beta}_i}. \tag{3.18}$$

O processo para quando ocorre a convergência no algoritmo. Mais detalhes sobre o processo de estimação pode ser visto em Hedeker e Gibbons (2006).

3.2.4 Comparação entre modelos marginais, modelo de transição e modelo com efeitos aleatórios

As três metodologias diferem quanto a maneira de levar em conta a dependência entre as observações pelo fato de não ser independentes no decorrer do tempo.

No modelo marginal, é comum estimar o vetor de parâmetros, $\boldsymbol{\beta}$, usando as Equações de Estimação Generalizadas (EEG) propostas por Liang e Zeger (1986), ou seja, usando uma matriz de correlação de trabalho, $R_i(\boldsymbol{\alpha})$, especificada pelo vetor de parâmetros, $\boldsymbol{\alpha}$, assumindo que esta correlação de trabalho seja a mesma para todos os indivíduos. O procedimento de estimação EEG para modelos marginais não é difícil de ser implementado, devido o fato de estar disponível nos principais pacotes de análise estatística. Quanto ao modelo misto com variável resposta binária existem poucos pacotes com algumas limitações, porém, vêem sendo implementados nos principais softwares.

Em contraste, no modelo misto, a dependência das observações no mesmo indivíduo é levado em conta na verossimilhança marginal, pela integração da função de densidade condicional ao efeito individual em relação à distribuição deste efeito Molenberghs e Verbeke (2000).

Já o modelo de transição apresenta uma característica que o distigue dos citados anteriormente, ou seja, poder modelar as mudanças individuias (transições) no tempo e, avaliar, como estas mudanças são influenciadas pelas variáveis explicativas no estudo. Neste trabalho, este método é apenas citado mas seu estudo não é aprofundado. Maiores informações podem ser obtidas em Lara (2007).

Capítulo 4

Técnicas de diagnóstico

A análise de diagnóstico é uma etapa importante no ajuste de um modelo de regressão, pois auxilia na verificação de possíveis afastamentos das suposições feitas para o modelo e permite detectar observações extremas que podem vir a interferir nos resultados do ajuste.

Quando se está ajustando um modelo a um conjunto de dados, é importante que as estimativas obtidas a partir do modelo proposto sejam resistentes a pequenas perturbações, tanto no modelo como nos dados. Se o modelo ajustado não apresentar uma boa descrição dos dados que foram observados, o mesmo pode conduzir a inferências errôneas, Souza (2006).

As análises de diagnóstico e de resíduos são utilizadas para detectar problemas, tais como:

- Presença de observações discrepantes (pontos aberrantes);
- Inadequação das pressuposições para os erros aleatórios;
- Colinearidade entre as colunas.

Paula (2004) descreveu algumas técnicas de diagnóstico para modelos lineares generalizados, tais como, os elementos da diagonal principal da matriz de projeção (matriz chapéu), a distância de Cook e os resíduos do modelo ajustado usados para detectar observações influentes na matriz de variáveis explicativas (pontos de alavanca) ou no vetor de respostas para detectar pontos discrepantes ("*outliers*").

Venezuela (2003) apresenta, baseando-se no trabalho de Tan, Qu e Kutner (1997), uma proposta para modelos com medidas repetidas, da qual será feita aqui uma adaptação para o caso longitudinal com variável resposta dicotômica, abordando as mesmas técnicas de diagnóstico utilizadas em modelos lineares generalizados e levando em consideração a estrutura de dependência entre observações intra-indivíduo.

Na literatura, são apresentados alguns trabalhos que tratam de técnicas de diagnóstico, usando como método de estimação as equações de estimação generalizadas, tais como Pan (2001) que apresenta medidas para a escolha da matriz de correlação de trabalho e para a seleção de variáveis explicativas, baseando-se no critério de informação de Akaike (AIC), Preisser e Qaqish (1996) que apresenta medidas para detectar observações influentes em modelos lineares generalizados com medidas repetidas, Venezuela (2003).

A seguir serão apresentadas de forma resumida as técnicas de diagnóstico para o modelo marginal, tais como: detecção de pontos de alavanca, pontos influentes, pontos *outliers*, análise gráfica e seleção de modelos.

4.1 Pontos de alavanca, influentes e *outliers*

Utilizando um processo iterativo para a obtenção das estimativas do vetor de parâmetros $\boldsymbol{\beta}$ no modelo marginal, para variáveis contínuas, conforme visto na seção (2.4), obtém-se a equação

$$\boldsymbol{\beta}^{(m+1)} \approx \boldsymbol{\beta}^{(m)} + (\mathbf{x}^T\mathbf{W}^{(m)}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{W}^{(m)}\mathbf{z}^{(m)}, \tag{4.1}$$

em que $m = 0, 1, 2, \ldots$ indica o número de iterações, $\mathbf{x} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)$ é uma matriz (n x p), $\mathbf{W}^m$ são matrizes (n x n) de pesos associadas às observações correlacionadas e que mudam a cada iteração, $\mathbf{z}^m = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T$ é um vetor (n x 1) de variáveis dependentes ajustadas cujos elementos são dados por

$$\mathbf{z}_1 = \mathbf{x}_i^T\beta_i + (\mathbf{y}_i - \boldsymbol{\mu}_i)(\frac{\partial \boldsymbol{\eta_i}}{\partial \boldsymbol{\mu_i}}).$$

Neste contexto, o vetor $\widehat{\boldsymbol{\beta}}$ pode ser interpretado como a solução de mínimos quadrados da regressão normal linear de $\mathbf{W}^{1/2}\mathbf{z}$ sobre $\mathbf{W}^{1/2}\mathbf{x}$, Artes e Botter (2005). Nessa perspectiva, o resíduo ordinário, que é a diferença entre os valores observados e

ajustados, fica sendo

$$\mathbf{r}^* = \mathbf{W}^{1/2}(\mathbf{z} - \boldsymbol{\eta}) = \mathbf{W}^{1/2}\mathbf{A}^{-1}(\mathbf{y} - \widehat{\boldsymbol{\mu}}), \tag{4.2}$$

em que $\mathbf{A} = \text{diag}(\mathbf{A}_1, \ldots, \mathbf{A}_n)$ é uma matriz diagonal (n x n) dos valores observados ajustados e $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ e $\mu = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n)$ com dimensões (n x 1).

Assumindo que $\text{Cov}(\mathbf{z}) = \mathbf{A}^{-1}\text{Cov}(\mathbf{y})\mathbf{A}^{-1} \cong \mathbf{W}^{-1}$, tem-se que

$$\text{cov}(\mathbf{r}^*) = (\mathbf{I} - \mathbf{H})\mathbf{W}^{1/2}\text{Cov}(\mathbf{z})\mathbf{W}^{1/2}(\mathbf{I} - \mathbf{H}) \cong (\mathbf{I} - \mathbf{H}), \tag{4.3}$$

sendo $\mathbf{I}$ a matriz identidade e $\mathbf{H}$ uma matriz diagonal simétrica e idempotente dada por $\mathbf{H} = \text{diag}(\mathbf{H}_1, \ldots, \mathbf{H}_n)$, com

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{x}(\mathbf{x}^T\mathbf{W}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{W}^{1/2}, \tag{4.4}$$

onde o posto de $\mathbf{H}$ é igual ao traço de $\mathbf{H}$ que é igual a p.

Observe que alguns elementos da matriz $\mathbf{W}$ são negativos, dificultando o cálculo da raiz quadrada desta matriz. Uma alternativa apresentada por Banerjee e Frees (1997), citado em Nobre (2004), sugere utilizar como matriz de alavancagem

$$\mathbf{H}^* = \mathbf{W}^{-1/2}\mathbf{x}(\mathbf{x}^T\mathbf{W}^{-1}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{W}^{-1/2}, \tag{4.5}$$

com $\mathbf{W}^{-1} = (\mathbf{W}^{-1/2})^T\mathbf{W}^{-1/2}$.

Como os elementos de $\mathbf{r}^*$ possuem variâncias diferentes, o que dificulta compará-los entre si, define-se o resíduo padronizado associado à observação y_{ij} por

$$(r_{SD})_{ij} = \frac{\mathbf{e}_{ij}^T\mathbf{W}_i^{1/2}\mathbf{H}_i^{-1}(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_i)}{\sqrt{1 - h_{ii}}}, \tag{4.6}$$

sendo $\mathbf{e}_{ij}$ um vetor de tamanho (n_i x 1) com a posição referente à observação y_{ij} contendo o valor 1 e as demais posições contendo o valor zero e h_{ii} o i-ésimo elemento da diagonal principal de $\mathbf{H}_i$, $i = 1, \ldots, n$ e $j = 1, \ldots, n_i$.

O resíduo estudentizado também pode ser escrito na forma $\mathbf{r}^* = (\mathbf{I} - \mathbf{H})\mathbf{W}^{-1/2}\mathbf{z}$. Assim, considerando que $\mathbf{W}^{-1/2}\mathbf{z}$ faz o papel do vetor resposta, $\mathbf{H}$ é chamada de matriz de projeção ortogonal (ou matriz chapéu), como na regressão normal linear em que $\mathbf{W}$ é uma matriz identidade. Isto, sugere a utilização dos elementos da diagonal principal de $\mathbf{H}$ para se detectar a presença de pontos alavanca, conforme Paula (2004) fez para

os MLGs, e, Tan, Qu e Kutner (1997) propuseram para o modelo de regressão logística com medidas repetidas, e, que será aplicado no caso longitudinal.

Um ponto de alavanca ocorre quando este possui uma característica diferente dos demais, quando este ponto está distante do centro do espaço gerado pelas variáveis explicativas. Assim, um valor alto de h_{ii} indica a influência de $\mathbf{x}_{ij}$ sobre o correspondente valor ajustado, y_{ij}.

Supondo que todos os pontos exercem a mesma influência sobre os valores ajustados, pode-se esperar que cada valor da diagonal principal de $\mathbf{H_i}$ esteja próximo de $\text{tr}(\mathbf{H}_i)/n = p/n$. Dessa forma, os pontos para os quais $h_{ij} \geq 2p/n$ podem ser considerados de alta *leverage*, Artes e Botter (2001).

Analogamente, o i-ésimo indivíduo pode ser um ponto *leverage*, se

$$h_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} h_{ij} = \frac{\text{tr}(\mathbf{H}_i)}{n_i} \geq \frac{2p}{n}. \tag{4.7}$$

Esses resultados podem ser vistos pelo gráfico dos valores da diagonal principal da matriz de projeção, h_{ii}, versus i, em que este índice indica a ordem em que cada indivíduo aparece no conjunto de dados, visualizando assim se o h_{ii} para o indivíduo i é considerado ou não um ponto de *leverage*.

Para detectar um ponto discrepante na análise gráfica, podemos utilizar o resíduo padronizado, $(r_{SD})_{ij}$, com $i = 1, \ldots, n$ e $j = 1, \ldots, n_i$, versus o índice i. Um ponto discrepante ("*outlier*") ocorre quando este apresenta um perfil diferente dos demais no que tange aos valores da variável resposta e também apresenta um valor baixo na matriz de projeção $\mathbf{H}_i$. Desta forma, um mesmo ponto dificilmente é um ponto de *leverage* e/ou um ponto discrepante.

Finalmente, um ponto influente ocorre quando este apresenta um valor diferente dos demais no que se refere aos valores da variável resposta, porém apresenta valor alto na matriz de projeção $\mathbf{H}_i$. Este tipo de ponto tem grande peso na estimação dos parâmentros do modelo e para detectá-lo, a medida mais conhecida é distância de Cook. Esta mede o afastamento entre a estimativa do vetor paramétrico utilizando todas as observações ($\widehat{\beta}$) e sem a observação y_{ij} ($\widehat{\beta}_{ij}$), Venezuela (2003). Assim, a distância de Cook, quando se exclui a observação y_{ij}, é definida por

$$\text{DC}_{ij} = \frac{1}{p}\left(\widehat{\beta} - \widehat{\beta}_{ij}\right)^T \mathbf{x}_i^T \mathbf{W}_i \mathbf{x}_i \left(\widehat{\beta} - \widehat{\beta}_{ij}\right) = (r_{SD})_{ij}^2 \frac{h_{ii}}{p(1-h_{ii})}, \tag{4.8}$$

indicando como ponto influente aquele que possui um valor alto de DC_{ij} quando comparado aos demais pontos.

Para modelos mistos em que a variável resposta é contínua, Christensen e Pearson (1992), citados em Nobre (2004), sugerem avaliar os pontos de alavanca do i-ésimo indivíduo através do valor $\mathbf{h}_i^* = \mathbf{h}_i/\mathbf{s}_i$, em que

$$\mathbf{h}_i = \mathbf{x}_i^T(\mathbf{x}^T\mathbf{V}^{-1}\mathbf{x})^{-1}\mathbf{x}_i,$$

$$\mathbf{x}_i = \mathbf{x}_i - \mathbf{x}_{(I)}^T\mathbf{V}_{(I)}^{-1}\mathbf{v}_i,$$

$$\mathbf{s}_i = v_{ii} - \mathbf{v}_i^T\mathbf{V}_{(I)}^{-1}\mathbf{v}_i,$$

com $\mathbf{x}_i$ a i-ésima coluna da matriz $\mathbf{x}$ e $\mathbf{v}_i$ a i-ésima coluna da matriz $\mathbf{V}$, conforme definido na equação (2.24), enquanto $\mathbf{x}_{(I)}$ e $\mathbf{V}_{(I)}$ representam, respectivamente, as matrizes $\mathbf{x}$ e $\mathbf{V}$ sem a i-ésima coluna e v_{ii} refere-se ao i-ésimo elemento da diagonal principal de $\mathbf{V}$.

Para modelos lineares mistos, foi proposta por Chatterjee e Hadi (1986, 1988), citados em Nobre (2004), a seguinte expressão para a distância de Cook

$$\mathrm{D}_I = \frac{(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(I)})_T\mathbf{y}^{-1}(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(I)})}{c}, \tag{4.9}$$

onde c representa um parâmetro de escala, I representa o conjunto de observações eliminadas e $\widehat{\mathbf{y}}$ é um vetor de observações estimadas.

Além destas estatísticas de diagnósticos, utiliza-se também técnicas gráficas para ajudar na detecção de possíveis anomalias no ajuste. A seguir apresenta-se alguns destes métodos gráficos.

4.2 Análise gráfica de diagnóstico

O uso de gráficos de diagnóstico é comum na análise de regressão para variáveis respostas com distribuição normal, uma vez que estes servem para detectar observações discrepantes ou comportamentos diferentes nos dados ou ainda verificar suposições feitas na modelagem. Esta etapa envolve a construção de vários tipos de gráficos de diagnóstico para verificar aspectos do modelo estimado, pois cada tipo de gráfico tenta identificar determinados desvios do modelo.

Para regressão logística, deve-se atentar aos mesmos cuidados, pois se deve examinar as relações entre a resposta e as variáveis explicativas para verificar se possíveis melhorias podem vir a surgir como conseqüência dos gráficos de diagnósticos. Neste trabalho, procurou-se adequar estas análises gráficas à situação em que os dados são observados ao longo do tempo, e que tendem a ser correlacionados.

Na literatura pode-se encontrar gráficos para este tipo de análise, porém será dada ênfase aos mais significativos no que se diz respeito à análise de diagnóstico no modelo de regressão logística, tais como, os gráficos de resíduos padronizados, distância de Cook e envelope de simulação. Estes gráficos fornecem uma avaliação da contribuição de cada ponto nos valores das estatísticas de diagnóstico em função das probabilidades estimadas. Por exemplo, grandes valores do resíduo padronizado sugerem que, entre as observações analisadas, existem candidatos a pontos aberrantes. Com o gráfico da distância de Cook mostra-se a influência de cada observação nas estimativas dos coeficientes.

Landwehr, Pregibon e Shoemaker (1984), citado em Farhat (2003), propuseram e discutiram três métodos gráficos que auxiliam na avaliação do ajuste do modelo de regressão logística. Tais métodos são generalizações de gráficos já existentes adaptados para levar em conta o aspecto binário da variável resposta. Sendo assim, será observado o comportamento destes gráficos na situação longitudinal.

O gráfico Q-Q e o de probabilidades simuladas são utilizados para detectar *outliers* e para avaliar a adequabilidade do modelo, respectivamente. Outro gráfico comumente utilizado é o gráfico de resíduos parciais com a finalidade de avaliar a linearidade do modelo. Estes gráficos são utilizados para o caso onde as observações são independentes. Será mostrado aqui uma adaptação onde os dados são correlacionados.

Pode-se resumir que o resíduo usado no gráfico envelope de simulação, é a diferença entre a observação y_i e o valor ajustada $\widehat{y}_i$. Sendo assim, o gráfico envelope de simulação, para o caso onde a variável resposta é contínua, com distribuição normal, pode ser obtido pelos seguintes passos:

(i) Para cada observação i, $i = 1, 2, \ldots, n$, simula-se um vetor de respostas de tamanho j, $j = 1, 2, \ldots, n_i$, levando em consideração a distribuição dos dados, que nesta situação supõe-se que seja uma distribuição normal, em relação aos

dados originais ajustados, o vetor de médias e a matriz de covariâncias;

(ii) Ajusta-se às respostas simuladas no passo anterior o mesmo modelo ajustado para y;

(iii) Calculam-se os resíduos padronizados conforme expressão dada pela equação (4.6) e, depois ordenam-se seus valores absolutos;

(iv) Repetem-se os passos $(i) - (iii)$ mais 24 vezes. Define-se o $(r_{SD})_{lm}$ como sendo o l-ésimo valor absoluto ordenado do resíduo padronizado pertencente à m-ésima simulação, $l = 1, 2, \ldots, n$ e $m = 1, 2, \ldots, M$, com $M = 25$. O valor $M = 25$ simulações é sugerido por Tan, Qu e Kutner (1997), citado em Venezuela (2003);

(v) Determina-se o mínimo, a mediana e o máximo dos menores valores absolutos dos resíduos padronizados de todas as simulações;

(vi) Repete-se o passo anterior para os segundos menores valores absolutos dos resíduos das simulações, $(r_{SD})_{2m}$, em seguida, os terceiros $(r_{SD})_{3m}$, e assim sucessivamente, até os maiores valores absolutos dos resíduos simulados. Ao final haverá três vetores de tamanho n contendo os mínimos, as medianas e os máximos dos resíduos padronizados, em valores absolutos;

(vii) Por fim faz-se um gráfico contendo os valores mínimos, medianas e máximos dos resíduos padronizados, como visto em Venezuela (2003).

Assim, pode-se concluir a partir do gráfico de envelope simulado quando apresenta grandes desvios dos pontos em torno da mediana dos valores simulados ou pontos próximos dos limites ou fora destes, que o modelo não está bem ajustado.

Já com relação ao gráfico de envelope simulado para o modelo logístico, o procedimento utilizado anteriormente sofre uma modificação no item (i), pois nesta situação a distribuição usada é de Bernoulli.

Quando se utiliza o gráfico de resíduos parciais, por exemplo, no caso de regressão linear normal, tem-se a finalidade de avaliar a necessidade de introduzir funções não lineares das variáveis explicativas ou não no modelo. Todavia, devido à natureza binária da variável y_{ij}, o gráfico de resíduos parciais consistirá de duas nuvens de pontos separadas, uma correspondente a $y_{ij} = 0$ e a outra $y_{ij} = 1$. Por esse motivo Landwehr,

Pregibon e Shoemaker (1984) usaram o método de suavização proposto por Cleveland (1979) com o intuito de facilitar a determinação da tendência exibida por esse gráfico. Neste trabalho não será explorado este tipo de gráfico, deixando como sugestão para estudos futuros.

Após o ajuste do modelo e de ter aplicado algumas técnicas gráficas, cabe ao estatístico escolher o modelo que melhor representa o comportamento dos dados. Um dos critérios de seleção de modelos, o critério de informação de Akaike (AIC), é comumente utilizado. A expressão que define este critério é

$$\text{AIC} = -2l(\widehat{\boldsymbol{\beta}}; \mathbf{y}) + 2p,$$

em que $l(\widehat{\boldsymbol{\beta}}; \mathbf{y})$ é a função de verossimilhança, $\widehat{\boldsymbol{\beta}}$ é o EMV de $\boldsymbol{\beta}$ sob o modelo candidato. Assim de uma classe de modelos candidatos, em que cada um é indexado por $\boldsymbol{\beta}$, é escolhido o modelo que minimiza o AIC. Entretanto, este critério não se aplica quando se utiliza o método EEG, pois o AIC é baseado na função de verossimilhança e nas propriedades assintóticas destes estimadores, ao contrário do que ocorre no método EEG, que está fundado no princípio de quase-verossimilhança.

Pan (2001) propôs uma modificação do AIC, substituindo a função de verossimilhança ($\text{l}(\widehat{\boldsymbol{\beta}}; \mathbf{y})$) pela função de quase-verossimilhança ($\text{Q}(\widehat{\boldsymbol{\beta}}; \mathbf{y})$) e também fez uma alteração no segundo termo da expressão do AIC. Este critério ficou conhecido como QIC, *Quasi-likelihood Information Criterion*, e é dado pela expressão

$$\text{QIC}_{(R)} = -2Q(\widehat{\boldsymbol{\beta}}; \mathbf{y}) + 2tr(\mathbf{W}, \mathbf{V}), \tag{4.10}$$

onde $\text{Q}(\widehat{\boldsymbol{\beta}}; \mathbf{y})$ é a função de quase-verossimilhança para o vetor de respostas $\mathbf{y}$, $\mathbf{W}$ é obtido pelo estimador $\mathbf{V} = -\partial^2 Q(\boldsymbol{\beta}; \mathbf{y})/\partial\beta\partial\beta^T$ e $\mathbf{W} = \mathbf{A}^{\frac{1}{2}}\mathbf{R}(\alpha)\mathbf{A}^{\frac{1}{2}}$, como visto na seção (2.8.1). Este critério é usado quando é usado o modelo marginal.

Capítulo 5

Aplicação

Neste capítulo, procuramos modelar os dados segundo as técnicas vistas nos capitulos 2 e 3 e, depois fazer uma análise de diagnóstico como apresentado no capítulo 4.

Os dados utilizados aqui, foram gentilmente fornecido pelo professor Dr. José Rubens Rebellato, do Departamento de Fisioterapia da UFSCar. Este conjunto consta de uma avaliação longitudinal em idosos no município de São Carlos/SP com o objetivo de observar a melhoria da qualidade de vida destes idosos através de determinadas atividades físicas. O procedimento para a coleta dos dados ocorreu da seguinte maneira: todos os idosos foram submetidos a quatro avaliações ao ano, uma inicial, ou seja, antes do início do programa de atividade física, e as outras foram realizadas a cada três meses, totalizando dez medições ao longo do estudo. Também, é importante lembrar que os idosos foram submetidos à avaliação médica que considerava características físicas e histórico de enfermidades pregressas que impediam a participação em qualquer das atividades previstas no programa. A seguir descreveremos as variáveis que foram medidas nesta avaliação:

- **Equilíbrio dinâmico (y)** - É a capacidade física que permite manter o corpo em equilíbrio durante o movimento. Para esta medida foi demarcada no chão (com fita adesiva) uma faixa com largura de 33,3 centímetros e comprimento de 3,33 metros. O idoso permaneceu em pé ao lado externo da borda, com os pés juntos, olhando para frente e depois, orientado a percorrer o trajeto demarcado,

na máxima velocidade que conseguia andar, mas sem correr. Ao final do percurso foi anotado o tempo gasto na travessia.

- **Idade (x1)** e **sexo (x2)**.

- **Pressão Arterial** - Foi medida por um esfigmomanômetro e um estetoscópio, onde foram coletadas as pressões arteriais sistólica **(x3)** (corresponde à pressão da artéria) no momento em que o sangue foi bombeado pelo coração, que, é representado pelo maior valor, e a pressão arterial diastólica **(x4)** (corresponde à pressão na mesma artéria, no momento em que o coração está relaxado após uma contração e, é representado pelo menor valor).

- **Peso (x5)** e **altura (x6)** - Foram medidos por meio de uma balança do tipo plataforma, que continha um estadiômetro para verificação da estatura. Nesta medição os idosos foram posicionados de costas para a balança e sem sapatos.

- **Freqüência cardíaca (x7)** - Para medição da freqüência cardíaca o paciente permaneceu posicionado da mesma forma, e em seguida o avaliador colocou seus dedos (2° e 3° dedos), sobre a artéria radial localizada na parte lateral do punho, tomando os batimentos cardíacos do indivíduo durante quinze segundos.

- **Força muscular (x8)** - É uma capacidade física que se utiliza quando se realiza movimentos musculares para vencer algum tipo resistência. Foi medida por meio da dinamômetria manual (os músculos responsáveis pelo movimento de pressão da mão).

- **Flexibilidade corporal (x9)** - É a capacidade física que permite a realização de movimentos com amplitude máxima, sem causar lesão. Foi avaliada por meio de um equipamento denominado Banco de Wells, que identificava a flexibilidade anterior do tronco (cadeia muscular posterior).

- **Equilibro estático perna esquerda (x10)** e **Equilibro estático perna direita (x11)** - É a capacidade física que permite manter o corpo equilibrado em posição estacionária. Foram realizados testes para a perna direita e a perna esquerda. O idoso ficou em pé com as mãos na cintura e foi orientado a olhar um ponto fixo (a uma distância de aproximadamente dois metros) e a flexionar na

altura do joelho uma das pernas, dizendo o idoso se manter nessa posição por pelo menos trinta segundos ou até ter se desequilibrado.

A variável Equilíbrio dinâmico foi escolhida como variável resposta neste estudo. Desta forma, procurou-se modelar a resposta média do equilíbrio dinâmico com relação às seguintes variáveis explicativas: tempo, idade dos pacientes, sexo, pressão arterial sistólica, pressão arterial diastólica, peso, altura, freqüência cardíaca, força muscular, flexibilidade corporal, equilibro estático perna esquerda e direita. A Tabela (5) mostra parte dos dados organizados na forma longitudinal.

Tabela 5.1: Dados referentes a avaliação de idosos para melhoria da qualidade de vida.

id	**o**	**t**	**x1**	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**	**x8**	**x9**	**x10**	**x11**	**y**	**status**
1	1	1	65	F	140	90	83,7	1,69	60	30	340	5	19	2,02	1
2	1	2	65	F	130	90	86,5	1,69	56	29	355	10	27	2,06	1
3	1	3	65	F	140	90	86,5	1,69	64	31	357	21	25	1,89	1
4	1	4	65	F	130	80	87,3	1,70	64	30	373	24	20	1,97	1
5	1	5	65	F	140	100	84,9	1,70	60	28	345	24	12	2,11	1
6	1	6	65	F	130	90	87,6	1,69	56	30	346	23	28	1,76	1
7	1	7	66	F	150	90	87,0	1,69	76	31	364	24	23	1,86	1
8	1	8	66	F	130	90	88,5	1,69	72	30	365	28	22	1,75	1
9	1	9	66	F	120	80	86,2	1,69	84	31	336	30	30	1,69	1
10	1	10	66	F	130	80	86,6	1,69	72	33	370	29	30	1,63	1
⋮	⋮							⋮							⋮
361	37	1	51	F	130	90	70,2	1,52	76	40	355	30	30	2,28	1
362	37	2	51	F	120	70	69,8	1,52	88	39	360	30	30	2,15	1
363	37	3	52	F	110	80	71,1	1,54	88	38	375	30	30	2,06	1
364	37	4	52	F	110	90	69,8	1,52	84	40	363	30	30	1,89	1
365	37	5	52	F	125	80	70,2	1,51	88	37	386	30	30	2,26	1
366	37	6	52	F	120	80	70,9	1,52	60	42	365	30	30	1,76	1
367	37	7	53	F	130	80	72,8	1,52	72	40	352	30	30	1,69	1
368	37	8	53	F	120	80	72,2	1,52	60	40	355	30	30	1,64	1
369	37	9	53	F	120	90	73,0	1,52	88	39	360	30	30	1,57	1
370	37	10	53	F	120	90	70,8	1,52	60	40	360	30	30	1,54	1

Fonte: Projeto de reavitalização de adultos/DFisio - UFSCar

5.1 Análise Exploratória

Inicialmente, foi realizada uma análise exploratória no conjunto de dados, com intuito de detectar algum tipo de anomalia, por exemplo, pontos discrepantes ("*outliers*"). Também foram utilizados gráficos *boxplot* e de perfis individuais, na tentativa de observar como a variabilidade dos dados se comporta ao longo do tempo e de identificar padrões individuais que podem ocorrer. O pacote estatístico utilizado nesta etapa da análise foi o *software* R, de domínio livre.

- **Resumo estatístico dos dados**

x1 - Idade

Mínimo	1º.Quartil	Mediana	Média	3º.Quartil	Máximo
47,0	57,0	61,0	61,4	66,0	79,0

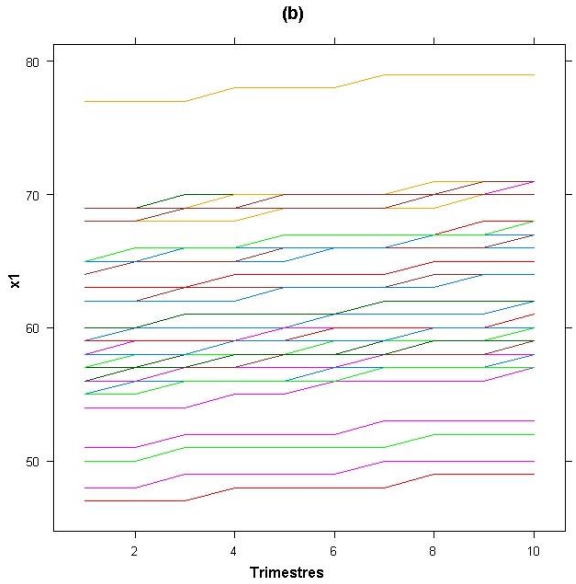

Figura 5.1: (b) Gráfico de perfis individuais da variável x1.

Observando as Figuras (5.1), nota-se que a idade média dos idosos envolvidos neste estudo é de 61 anos, havendo um idoso com idade superior a 75 anos e outros três com idade inferir a 50 anos. O conjunto de dados é composto por 30 idosos do sexo feminino e 7 do sexo masculino.

x3 - Pressão arterial sistólica

Mínimo	1º.Quartil	Mediana	Média	3º.Quartil	Máximo
90	110	120	123	130	170

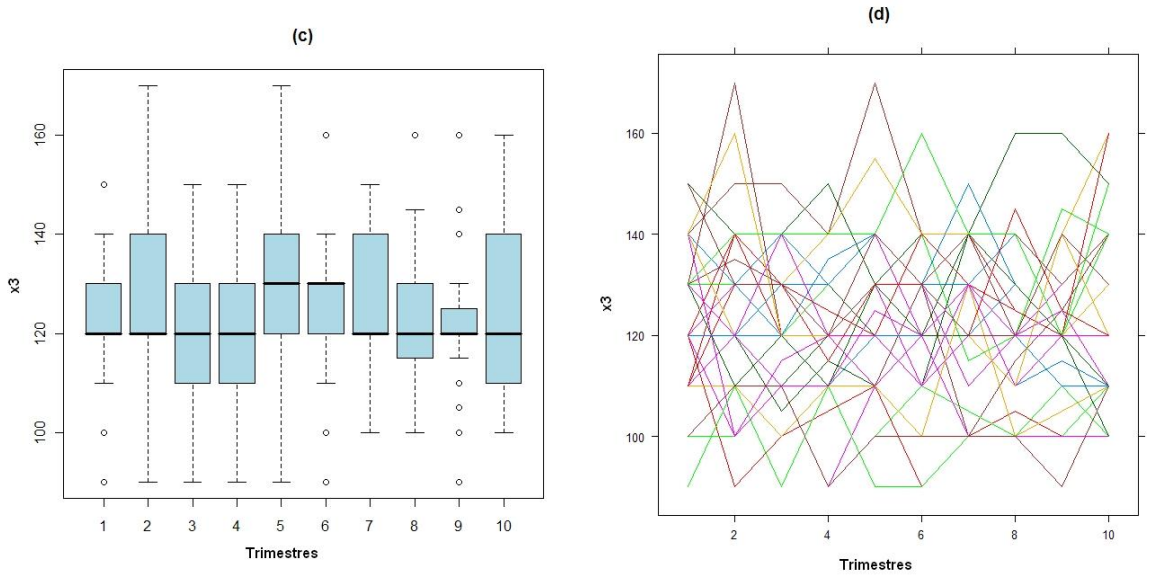

Figura 5.2: (c) *Boxplot* da variável x3 e (d) Gráfico de perfis individuais.

x4 - Pressão arterial diastólica

Mínimo	1º.Quartil	Mediana	Média	3º.Quartil	Máximo
50,0	70,0	80,0	78,4	80,0	100,0

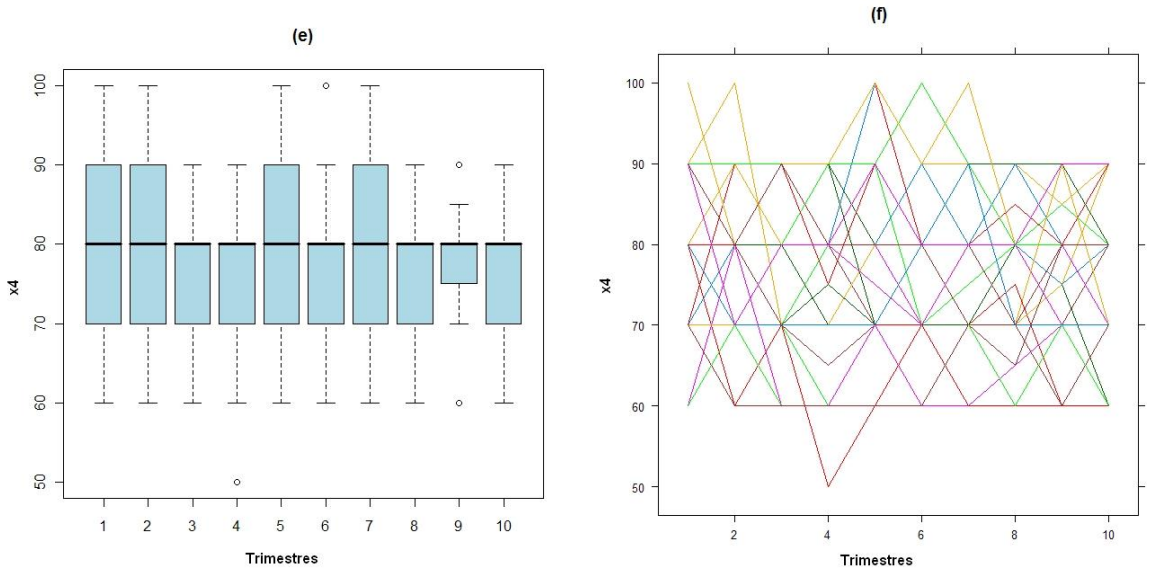

Figura 5.3: (e) *Boxplot* da variável x4 e (f) Gráfico de perfis individuais.

As Figuras (5.2) e (5.3), referentes a pressão arterial sistólica e diastólica, respectivamente, mostram uma pressão média em torno de 12,3(mmHg) / 7,8(mmHg). Porém, há uma observação de uma pressão 17/10 (mmHg), indicando indício de uma

hipotensão moderada.

x5 - Peso

Mínimo	1º.Quartil	Mediana	Média	3º.Quartil	Máximo
44,6	60,5	69,7	68,9	74,3	94,8

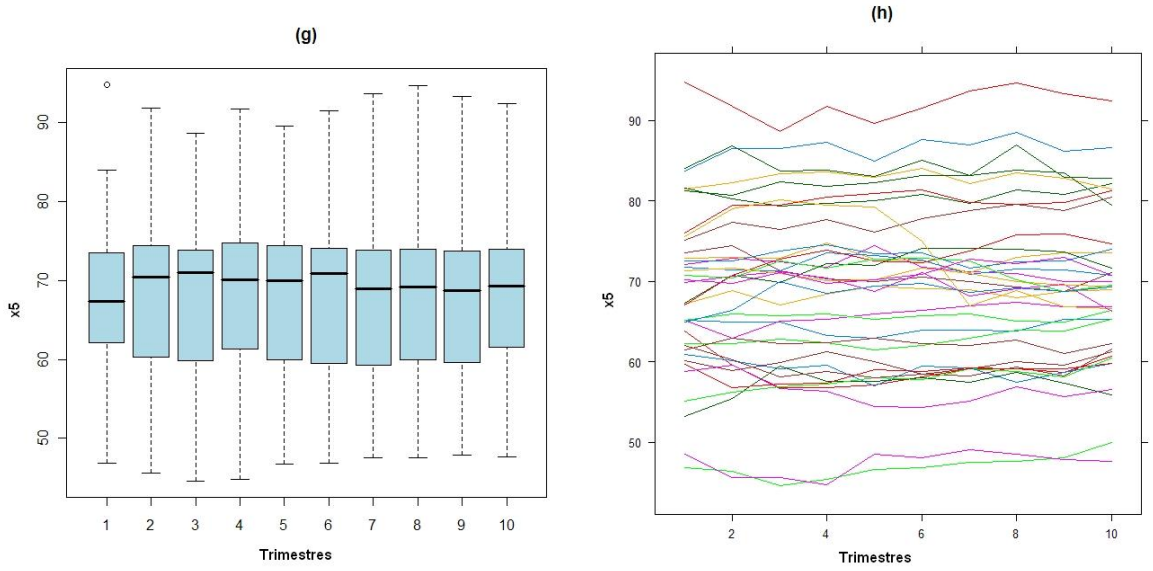

Figura 5.4: (g) *Boxplot* da variável x5 e (h) Gráfico de perfis individuais.

x6 - Altura

Mínimo	1º.Quartil	Mediana	Média	3º.Quartil	Máximo
1,44	1,53	1,58	1,59	1,64	1,81

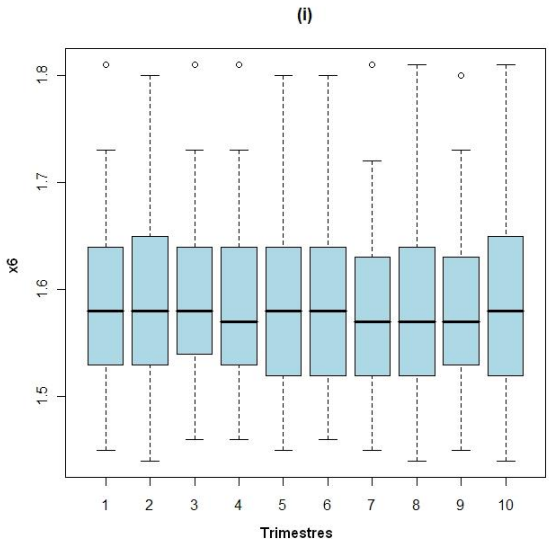

Figura 5.5: (i) *Boxplot* da variável x6.

Observando a Figura (5.4), referente a variável peso, verifica-se que o peso médio ao longo do tempo está em torno de 68,86 Kg. Já com relação à Figura (5.5), eferente à variável altura, observa-se que a maior parte dos idosos mede entre 1,5 e 1,7m.

Uma relação entre estas duas medidas é conhecida com IMC (Índice de Massa Corpórea), dada pela relação peso por altura ao quadrado. Segundo a Organização Mundial de Saúde, uma pessoa com um IMC acima de 25 é considerada levemente obesa, podendo ter complicações futuras com a saúde.

x7 - Freqüência cardíaca

Mínimo	1º.Quartil	Mediana	Média	3º.Quartil	Máximo
44,0	64,0	72,0	72,4	80,0	100,0

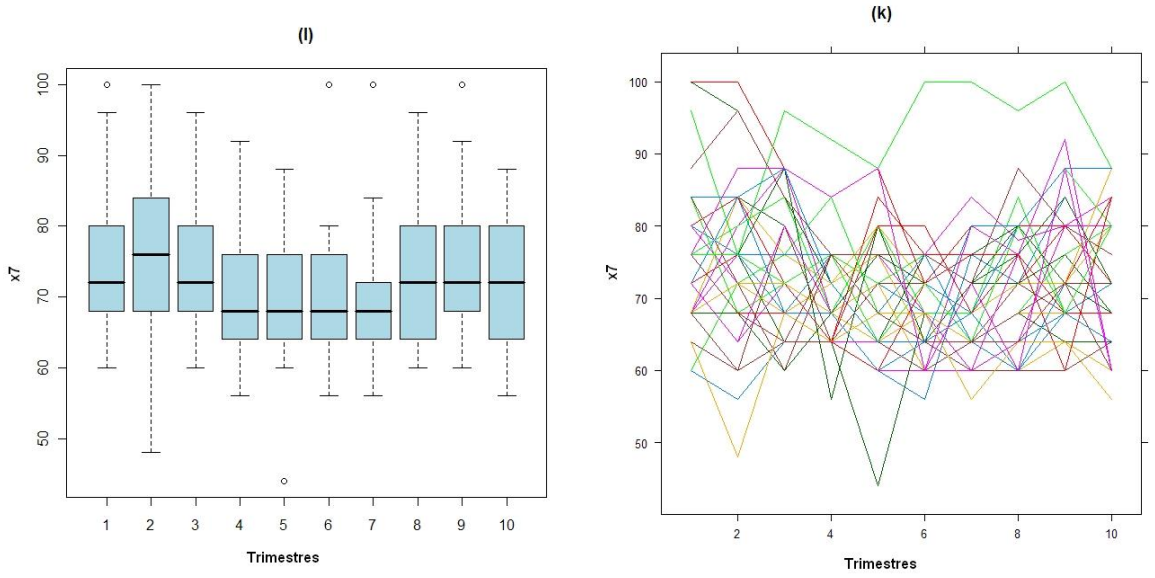

Figura 5.6: (l) *Boxplot* da variável x7 e (k) Gráfico de perfis individuais.

A freqüência cardíaca, visualizada nas Figuras (5.6) (l) e (k), apresenta pequena variação ao longo do tempo, o que é esperado, para esta faixa etária.

x8 - Força muscular

Mínimo	1º.Quartil	Mediana	Média	3º.Quartil	Máximo
16,0	27,0	32,0	33,8	40,0	67,0

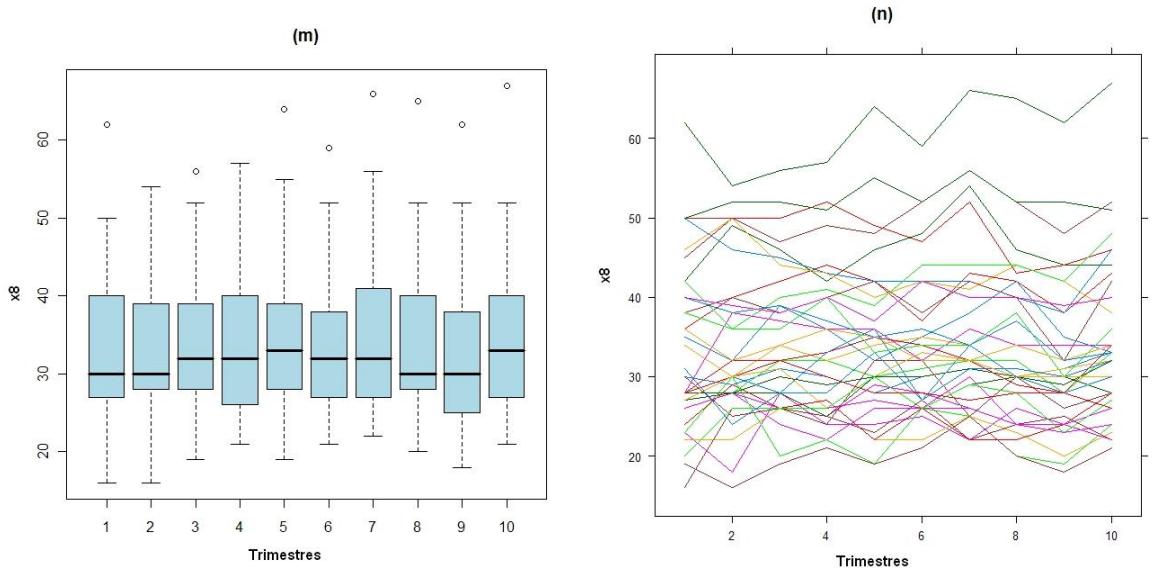

Figura 5.7: (m) *Boxplot* da variável x8 e (n) Gráfico de perfis individuais.

x9 - Flexibilidade

Mínimo	1º.Quartil	Mediana	Média	3º.Quartil	Máximo
90	214	290	283	355	447

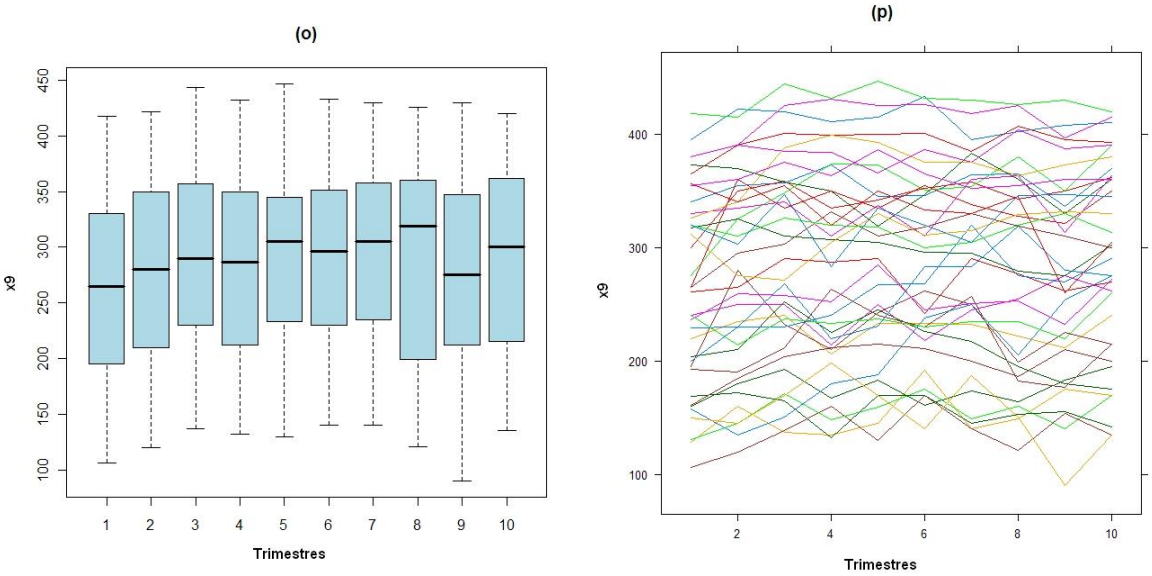

Figura 5.8: (o) *Boxplot* da variável x9 e (p) Gráfico de perfis individuais.

A força muscular média obtida ao longo do estudo foi de 33,8 libras. Nas Figuras (5.7) (m) e (n), pode-se observar que não houve muita variação, pois a maior parte dos idosos tem a medida da força muscular inferior a 50 libras, e apenas um idoso apresentou força superior a 50 libras.

Outra medida analisada neste estudo foi a flexibilidade corporal. Seu comportamento pode ser visto nas figuras (5.8) (o) e (p). Esta variável mediu a capacidade física que o idoso tem de realizar certos movimentos, por exemplo, sentar, levantar ou locomover-se com agilidade, sem causar lesões. Observa-se uma certa variabilidade entre os idosos no decorrer do estudo.

x10 - Equilíbrio estático P.E.

Mínimo	1º.Quartil	Mediana	Média	3º.Quartil	Máximo
3,0	23,0	30,0	25,2	30,0	30,0

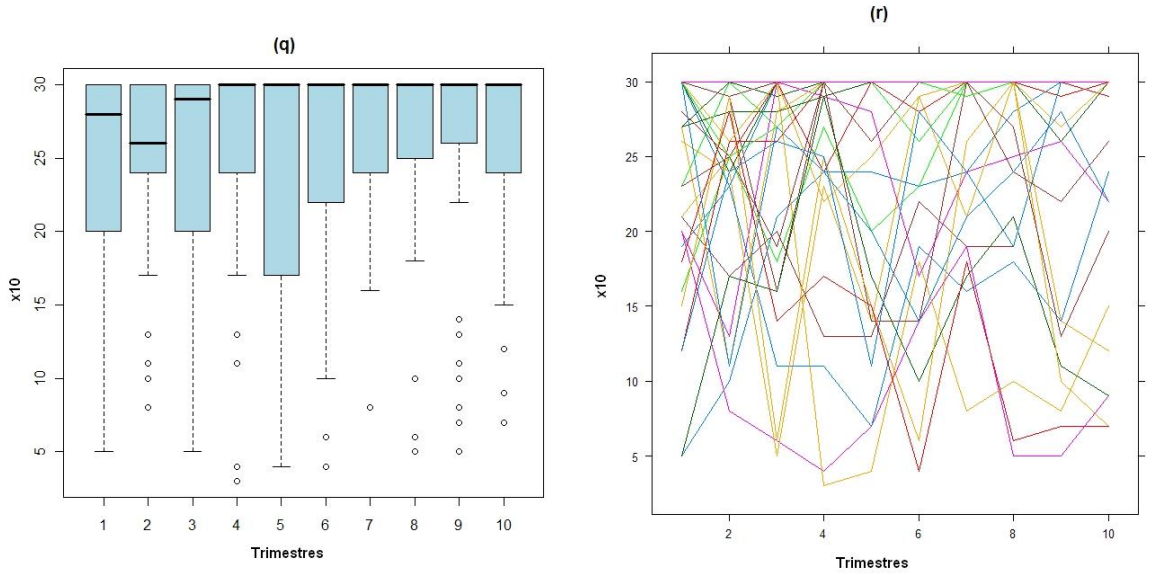

Figura 5.9: (q) *Boxplot* da variável x10 e (r) Gráfico de perfis individuais.

x11 - Equilíbrio estático P.D.

Mínimo	1º.Quartil	Mediana	Média	3º.Quartil	Máximo
0,0	21,0	30,0	24,6	30,0	30,0

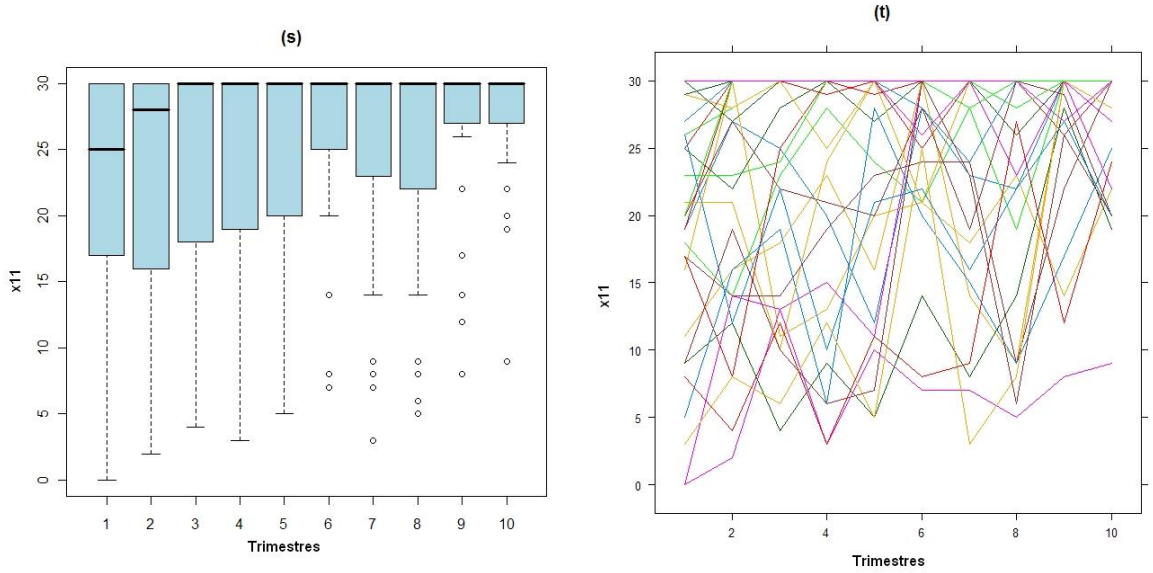

Figura 5.10: (s) *Boxplot* da variável x11 e (t) Gráfico de perfis individuais.

y - Equilíbrio dinâmico

Mínimo	1º.Quartil	Mediana	Média	3º.Quartil	Máximo
1,13	1,71	1,93	1,94	2,15	3,13

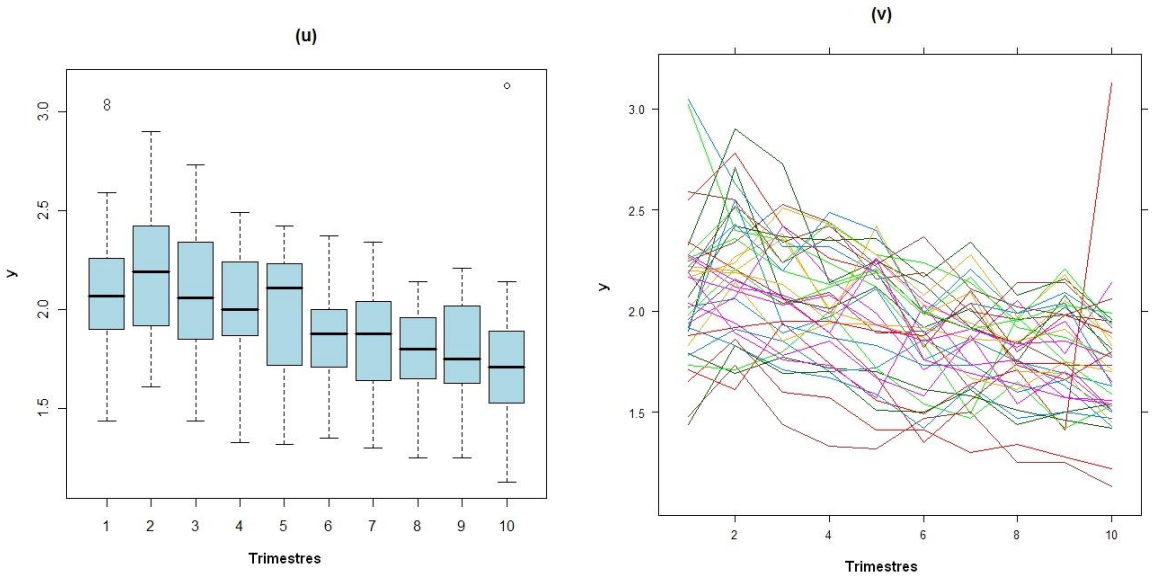

Figura 5.11: (u) *Boxplot* da variável y e (v) Gráfico de perfis individuais.

As Figuras (5.9) e (5.10), referentes ao equilíbrio estático PDA e PEA, respectivamente, mostram um leve aumento no tempo do idoso de manter seu corpo em equilíbrio ao executar determinada atividade.

Já o equilíbrio dinâmico, representado na Figura (5.11), mostra que houve uma melhora nos idosos, ao longo do tempo, na capacidade de manter o corpo equilibrado durante o movimento, sem sofrer alguma lesão.

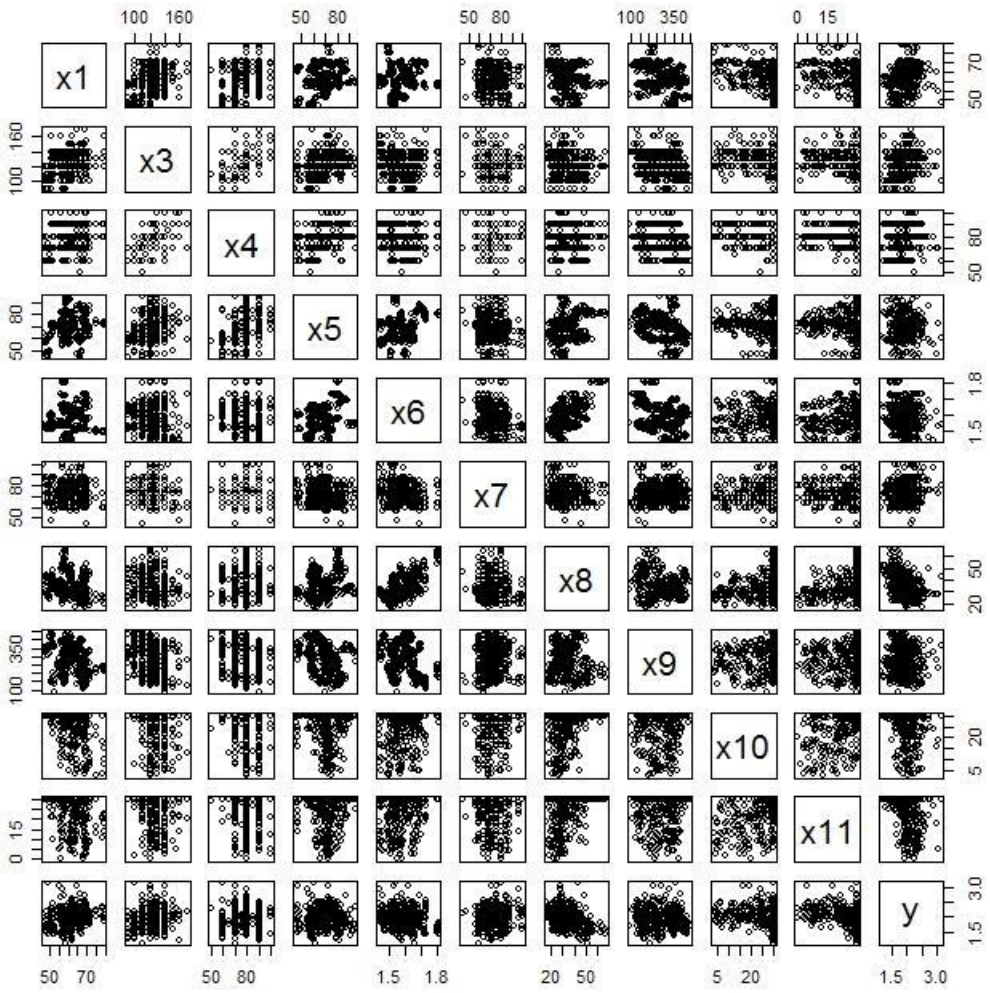

Figura 5.12: Gráfico de disersão de pares.

As correlações vista na Figura (5.12), mostram que há uma correlação moderada entre as variáveis x3 (Pressão Arterial Sistólica) e x4(Pressão Arterial Diastólica), ($\rho = 0,6744$), e, também entre as variáveis x6(Altura) e x8(Força muscular), ($\rho = 0,6306$).

5.2 Modelagem com variável resposta contínua

Após uma análise exploratória nos dados, foram ajustados modelos de regressão, conforme descrito no capítulo 2, nas seções (2.3.1) e (2.3.3), referentes a modelagem marginal e modelagem com efeitos mistos para variável resposta contínua. Como os indivíduos foram observados ao longo do tempo, uma estrutura de correlação foi utilizada, como visto na seção (2.3.1).

No ajuste deste modelo, foram utilizados os pacotes *gee* e *nlme*, do *software* R. Depois, verificou-se através de técnicas gráficas, vistas no capítulo 4, como os resíduos se comportaram e calcularam-se, a distância de Cook, para averiguar a influência das variáveis explicativas nas estimativas dos parâmetros. Também utilizou-se o envelope de simulação, para verificar a adequabilibade do modelo.

5.2.1 Modelo marginal

Neste ajuste foi utilizado o modelo marginal, conforme expressão dada na equação (2.23), sendo y_i a variável resposta (Equilíbrio dinâmico) e utilizada uma estrutura de correlação uniforme, vista na seção (2.3.1). A Tabela (5.2) apresenta os resultados deste ajuste.

Tabela 5.2: Estimativas dos parâmetros e P-valores do modelo marginal com resposta contínua ajustado com estrutura de correlação uniforme (EX).

Coeficientes	**Estimativas(EX)**	**P-valor(EX)**	**Estimativas(EX*)**	**P-valor(EX*)**
(Intercepto)	2,16392	0,00111	2,27104	0,00104
x1	0,01154	0,00795	0,01170	0,01154
x2	-0,27173	0,06013	-0,24464	0,08908
x3	0,00121	0,22953	0,00080	0,42492
x4	-0,00198	0,24686	-0,00071	0,63244
x5	0,00086	0,71575	0,00087	0,71871
x6	-0,30401	0,52392	-0,41236	0,39245
x7	0,00191	0,17919	0,00253	0,05410
x8	0,00048	0,88711	-0,00055	0,86324
x9	-0,00050	0,11579	-0,00055	0,09791
x10	-0,00330	0,07450	-0,00322	0,08122
x11	-0,00462	0,02708	-0,00373	0,04764
t	-0,04916	0,00000	-0,05251	0,00000

(*) ajuste sem o quarto indivíduo

A Figura (5.13), apresenta os gráficos distância de Cook e resíduos padronizados do modelo de regressão normal quando ajustado com estrutura de correlação uniforme. Observa-se um valor discrepante, referente a observação 40, que pode ser um possível ponto de influência e, está influenciando na estimativa dos parâmetros. Este ponto

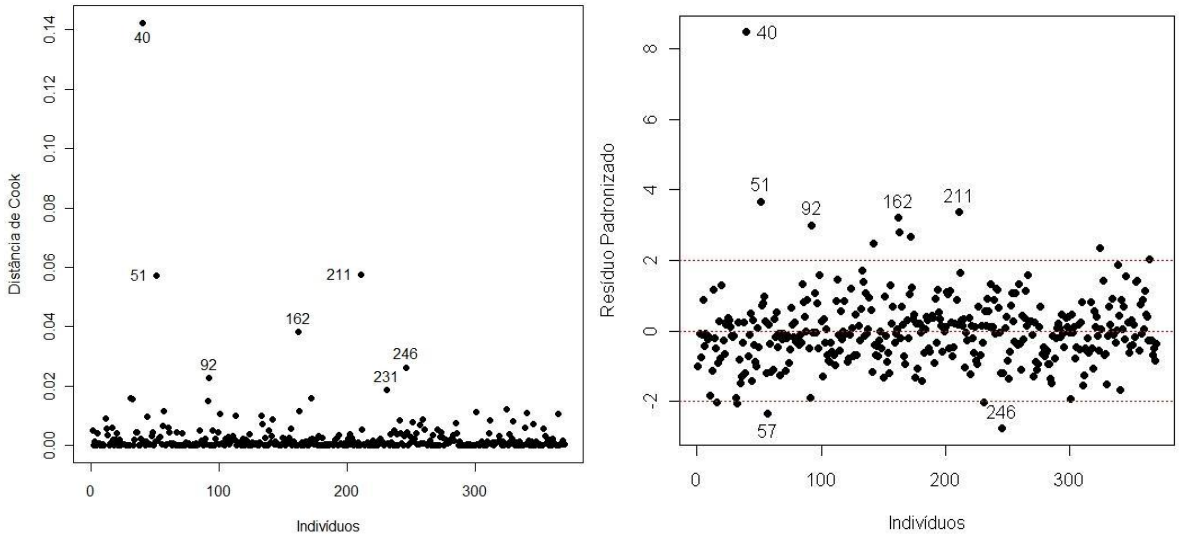

Figura 5.13: Distância de Cook e Resíduos padronizados do modelo marginal com resposta contínua ajustado com estrutura de correlação uniforme.

também se destaca no gráfico de envelope de simulação, mostrado na Figura (5.14), que apresenta quatro versões do mesmo gráfico, com escalas diferentes, para esclarecer se os pontos do canto inferior esquerdo estão dentro ou fora do intervalo.

O gráfico de envelope simulado tem por finalidade verificar a adequabilidade do modelo ajustado. Se muitos pontos estiverem fora do intervalo de credibilidade o ajuste não é recomendado. Olhando para a Figura (5.14), observa-se que os Gráficos de envelope simulado $(a.1)$, $(a.2)$ e $(a.3)$, mostram que a estrutura de correlação uniforme não é adequada para estes dados.

Refazendo a análise, agora sem o quarto indivíduo que possui uma observação discrepante, nota-se que as estimativas dos parâmetros (Tabela (5.2), (EX*)) não sofreram grandes modificações, indicando que esta observação não era muito influente. A Figura (5.15) se refere à análise de resíduos após a redirada do indivíduo com observação discrepante segundo a Figura (5.13), mostra a existência de outros valores discrepantes que não são tão influentes.

Quando analizamos a Figura (5.16), observa-se que o gráfico de envelope simulado tem uma leve melhora quanto a disposição dos pontos no gráfico, quando eliminamos a observação discrepante.

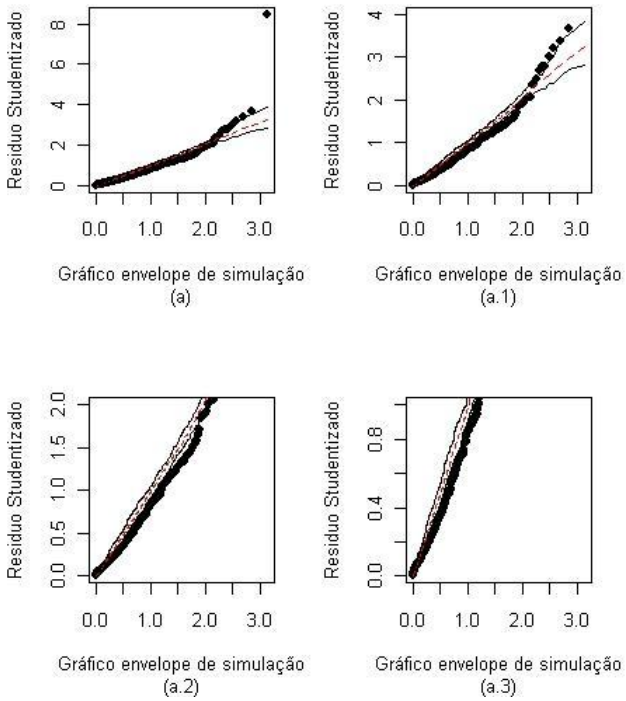

Figura 5.14: Envelopes simulados do modelo marginal com resposta contínua ajustado com estrutura de correlação uniforme.

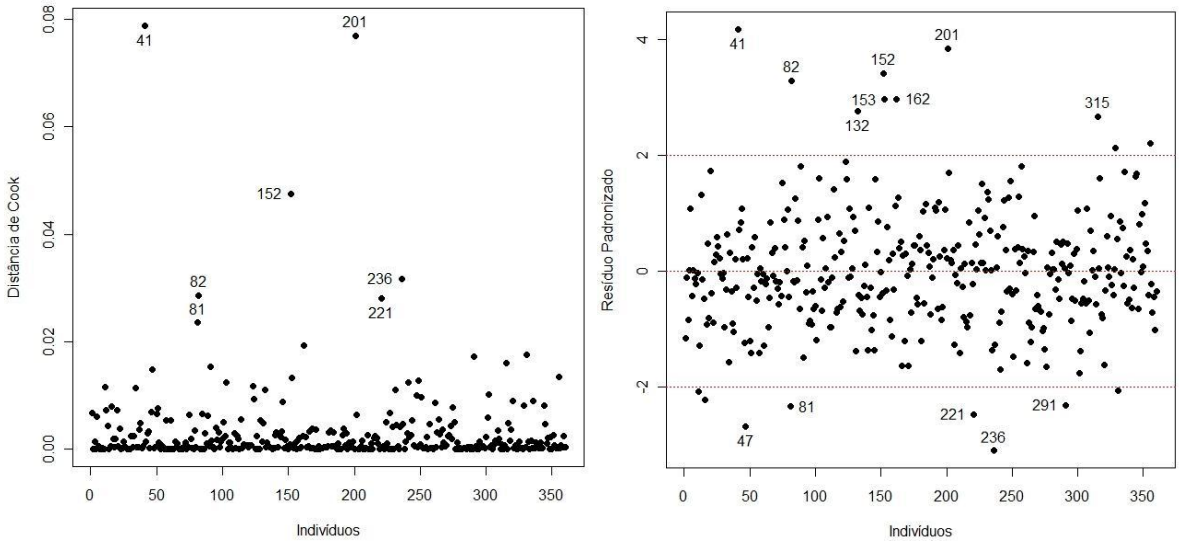

Figura 5.15: Distância de Cook e Resíduos padronizados do modelo marginal com resposta contínua ajustado com estrutura de correlação uniforme sem o quarto indivíduo.

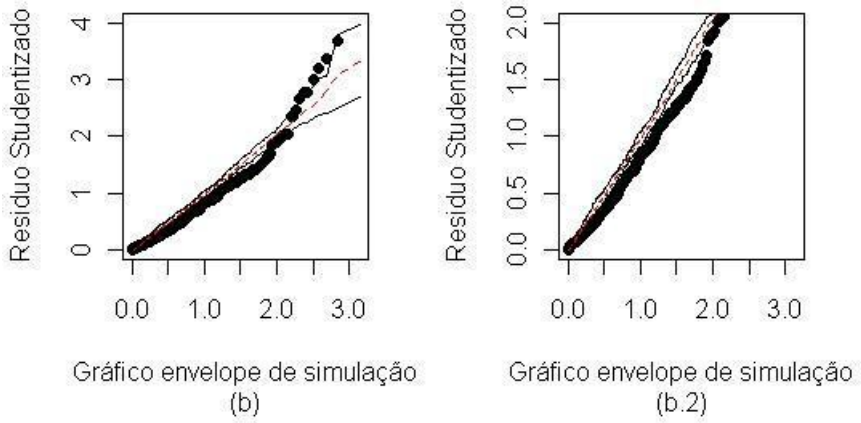

Figura 5.16: Envelope de simulação do modelo marginal com resposta contínua ajustado com estrutura de correlação uniforme sem o quarto indivíduo.

A seguir foi ajustado o modelo de regressão com estrutura de correlação AR-1 e pode-se observar que também aparece uma observação que se destaca das demais, como é mostrado nos Gráficos (5.19), da distância de Cook e do resíduo padronizado, podendo estar ou não influenciando na estimativa dos parâmetros. A figura (5.20), mostra que o modelo com a estrutura AR-1 é um pouco mais adequada do que o modelo com estrutura de correlação uniforme (EX).

Tabela 5.3: Estimativas dos parâmetros e P-valores do modelo marginal com resposta contínua ajustado com estrutura de correlação AR-1.

Coeficientes	**Estimativas(AR-1)**	**P-valor(AR-1)**	**Estimativas(AR-1*)**	**P-valor(AR-1*)**
(Intercepto)	2,13084	0,00222	2,11465	0,00396
x1	0,01020	0,02775	0,01099	0,02244
x2	-0,27171	0,06389	-0,25540	0,07645
x3	0,00233	0,03006	0,00192	0,07027
x4	-0,00206	0,21978	-0,00102	0,53963
x5	-0,00057	0,80560	-0,00053	0,82229
x6	-0,28321	0,57297	-0,30770	0,54267
x7	0,00184	0,12371	0,00176	0,14207
x8	0,00041	0,92001	-0,00046	0,90487
x9	-0,00051	0,09566	-0,00050	0,09362
x10	-0,00234	0,21882	-0,00232	0,21364
x11	-0,00434	0,04562	-0,00364	0,07222
t	-0,04499	0,00000	-0,04876	0,00000

(*) ajuste sem o quarto indivíduo

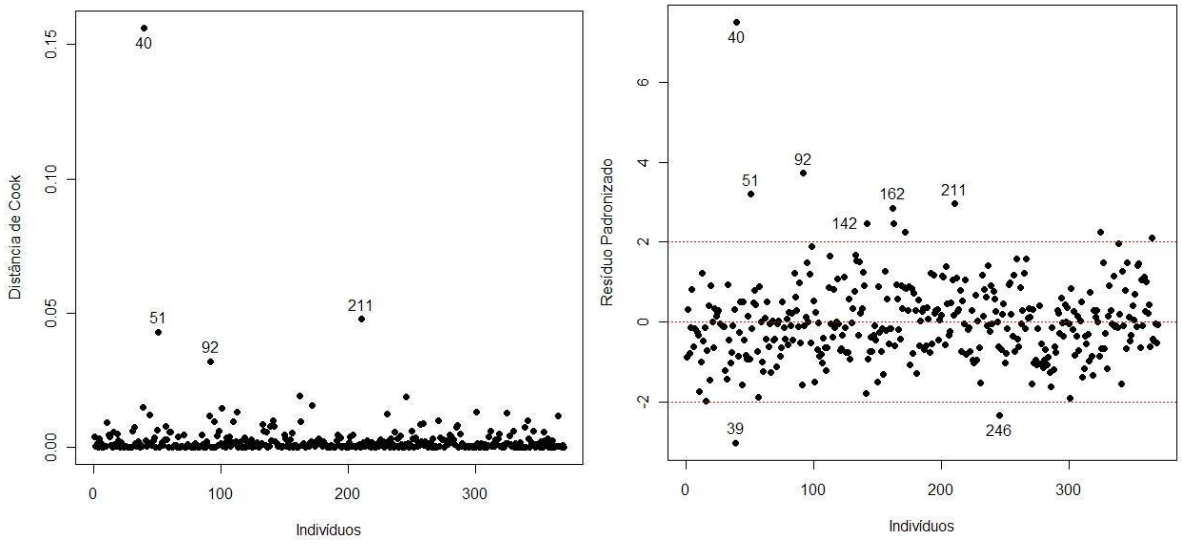

Figura 5.17: Distância de Cook e Resíduos padronizados do modelo marginal com resposta contínua ajustado com estrutura de correlação AR-1, com todos os indivíduos.

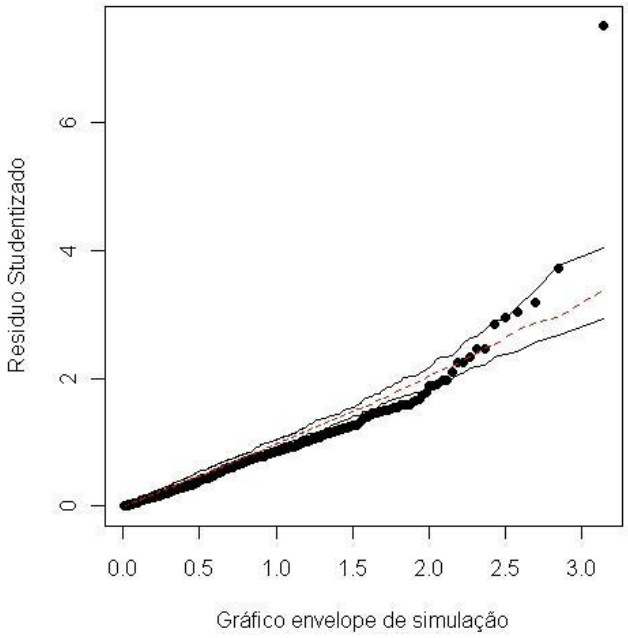

Figura 5.18: Envelope de simulação do modelo marginal com resposta contínua ajustado com estrutura de correlação AR-1, com todos os indivíduos.

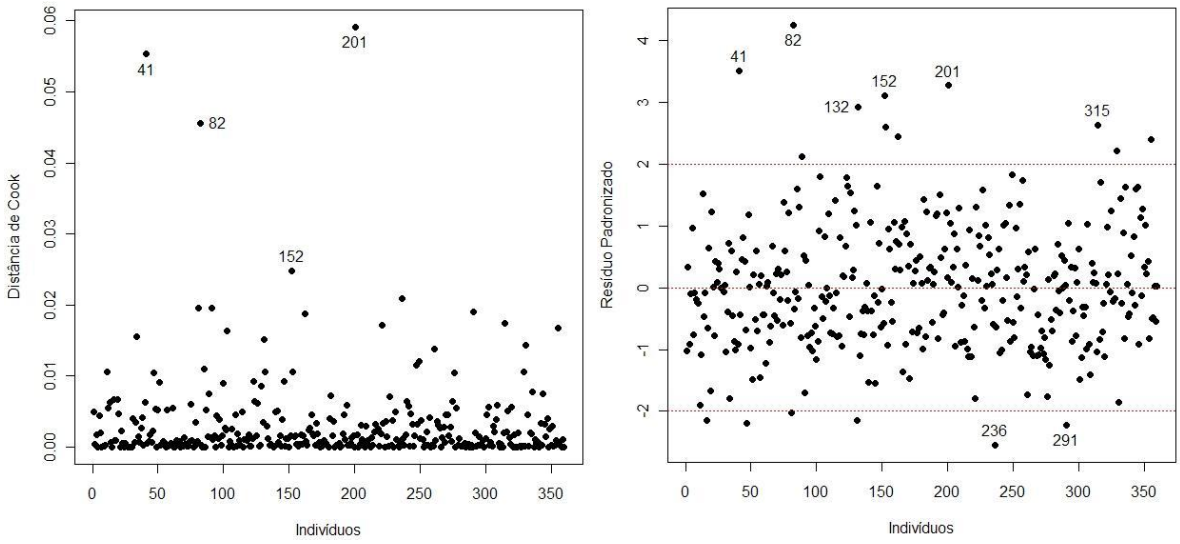

Figura 5.19: Distância de Cook e Resíduos padronizados do modelo marginal com resposta contínua ajustado com estrutura de correlação AR-1, sem o quarto indivíduo.

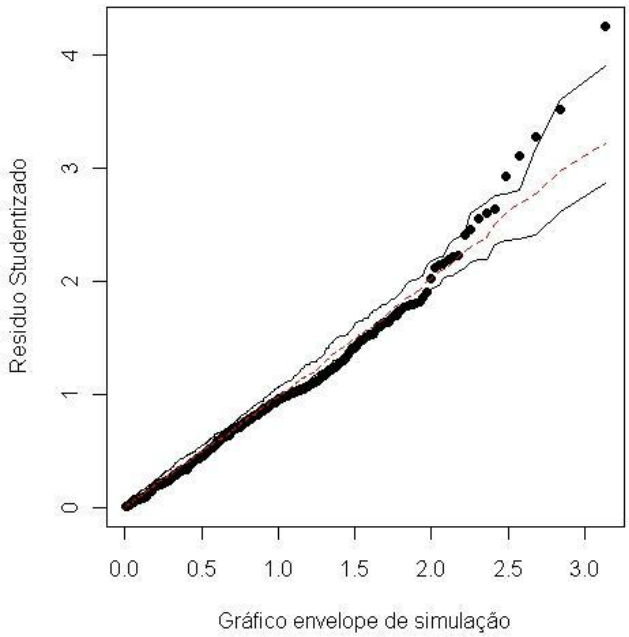

Figura 5.20: Envelope de simulação do modelo marginal com resposta contínua ajustado com estrutura de correlação AR-1, sem o quarto indivíduo.

5.2.2 Modelo com efeito aleatório

Agora, ajustamos aos dados um modelo de regressão com intercepto aleatório dado pela expressão (2.13), visto no capítulo 2. Este modelo se diferencia do modelo marginal, por ser capaz de captar características de cada indivíduo. Como estrutura de correlação, utilizou-se na modelagem a estrutura AR-1, e os resultados são apresentados na Tabela (5.4).

Na Figura (5.21), a observação 40 se destaca das demais, mas não parece influenciar nas estimativas dos parâmetros. Assim, pode-se dizer que este modelo é adequado aos dados, apesar de alguns pontos ficarem fora do intervalo de credibilidade mostrado na Figura (5.22).

Tabela 5.4: Estimativas dos parâmetros e P-valores do modelo misto com resposta contínua com intercepto aleatório ajustado com estrutura de correlação AR-1.

Coeficientes	**Estimativas**	**P-valor**
(Intercepto)	2,16184	0,00541
x1	0,01184	0,02207
x2	-0,27717	0,02411
x3	0,00102	0,37968
x4	-0,00175	0,33582
x5	0,00113	0,71052
x6	-0,32930	0,48599
x7	0,00188	0,15038
x8	0,00085	0,79875
x9	-0,00049	0,14216
x10	-0,00338	0,09505
x11	-0,00456	0,02580
t	-0,04928	0,00000

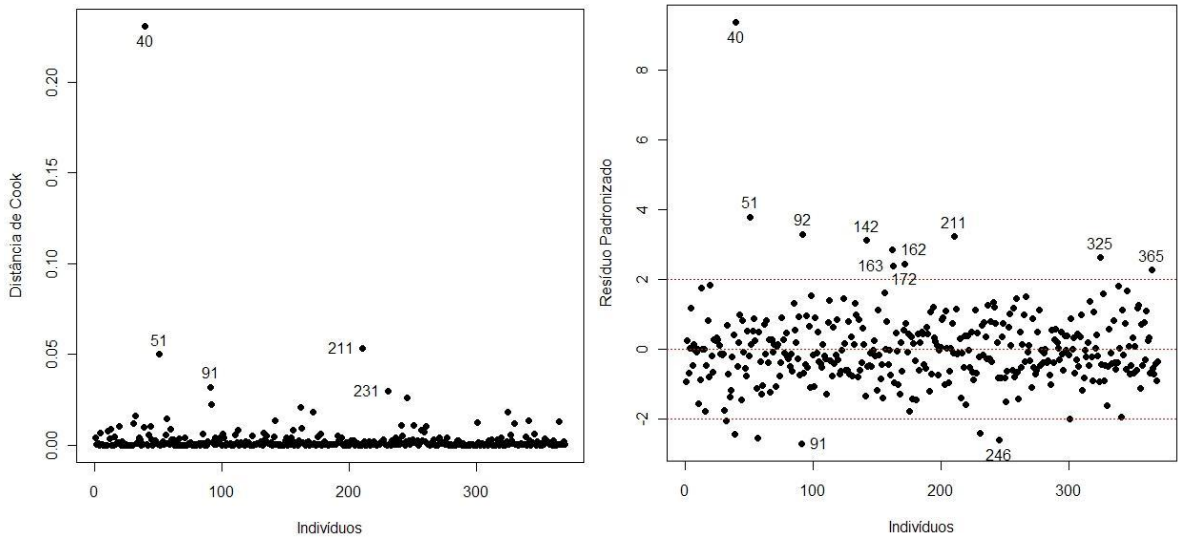

Figura 5.21: Distância de Cook e Resíduos padronizados do modelo misto com resposta contínua com intercepto aleatório ajustado com estrutura de correlação AR-1.

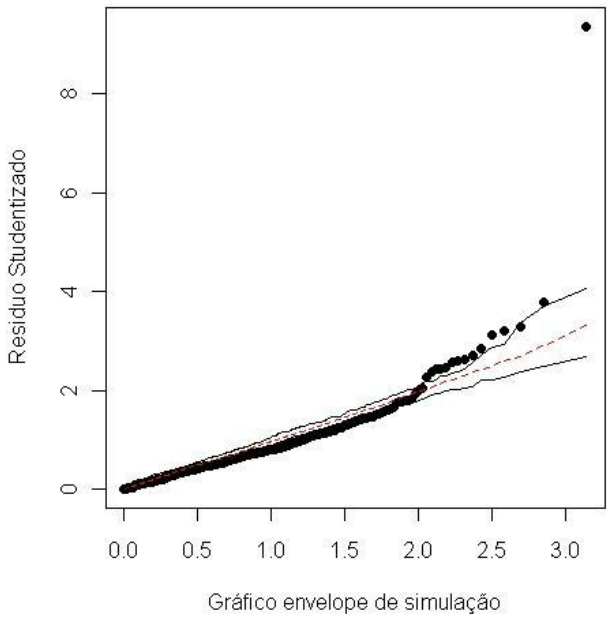

Figura 5.22: Envelope de simulação do modelo misto com resposta contínua com intercepto aleatório ajustado com estrutura de correlação AR-1.

5.3 Modelagem com variável resposta binária

Nesta etapa, a variável equilíbrio dinâmico foi categorizada tomando o valor 0 para valores $> 2,34$, e, 1 caso contrário. Esta nova variável recebeu o nome de *status*, Matsudo (2000). A transformação foi realizada com o intuito de utilizar à metodologia apresentada no capítulo 3, como também as técnicas de diagnóstico do capítulo 4. Quanto à modelagem, utilizaram-se o modelo marginal e o modelo misto com intercepto aleatório, como será visto a seguir.

5.3.1 Modelo marginal

O modelo marginal aplicado aqui, foi o apresentado na seção (2.3.1). Foi feita uma análise dos dados aplicando o modelo visto em (3.8), com estruturas de correlação uniforme e AR-1, e os gráficos distância de Cook, resíduos padronizados e envelope de simulação estão apresentados nas Figuras (5.23) a (5.26).

Observa-se que a estutura de correlação AR-1 melhor se adaptou aos dados e uma confirmação deste ajuste pode ser visto no envelope de simulação mostrado na Figura (5.26).

Tabela 5.5: Estimativas dos parâmetros e P-valores do modelo marginal com resposta binária ajustado com estruturas de correlação uniforme(EX) e AR-1.

Coeficientes	**Estimativas(EX)**	**P-valor(EX)**	**Estimativas(AR-1)**	**P-valor(AR-1)**
(Intercepto)	5,41090	0,37464	3,26254	0,5840
x1	0,06353	0,17451	0,01099	0,02244
x2	-0,82755	0,48984	-0,54757	0,58810
x3	0,01059	0,48901	0,02259	0,13650
x4	0,01796	0,44993	-0,00136	0,95140
x5	-0,00564	0,72698	0,00284	0,83240
x6	-4,20234	0,31982	-3,19791	0,43460
x7	0,00247	0,90142	0,00879	0,65640
x8	-0,03345	0,38757	-0,05150	0,17180
x9	-0,00317	0,34350	-0,00360	0,20700
x10	-0,04154	0,07816	-0,01678	0,5011
x11	-0,02078	0,30433	-0,02358	0,3130
t	-0,37130	0,00000	-0,36984	0,00000

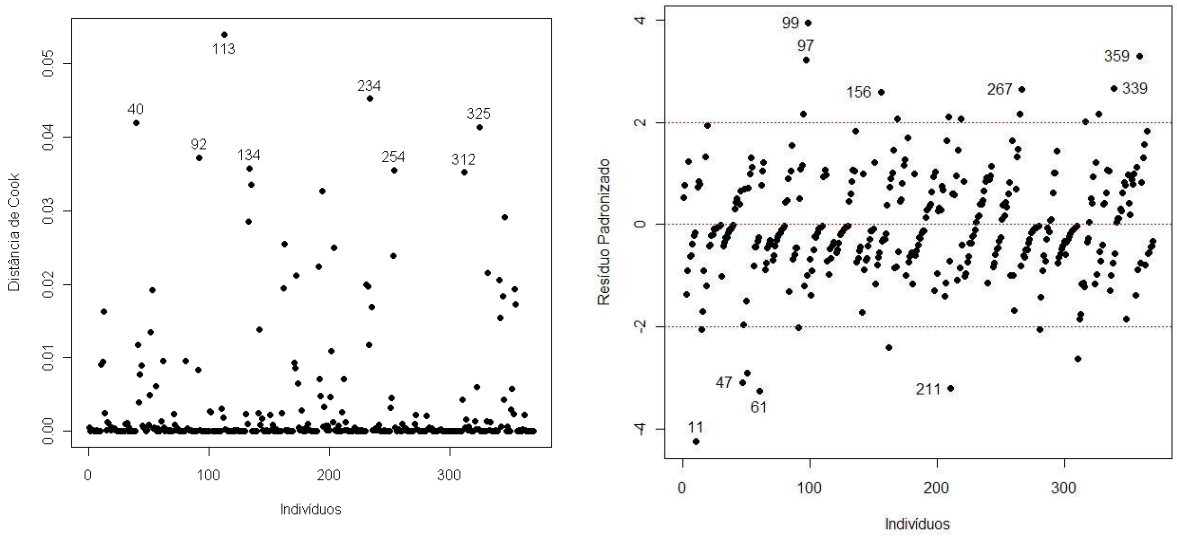

Figura 5.23: Distância de Cook e Resíduos padronizados do modelo marginal com resposta binária ajustado com estrutura de correlação uniforme.

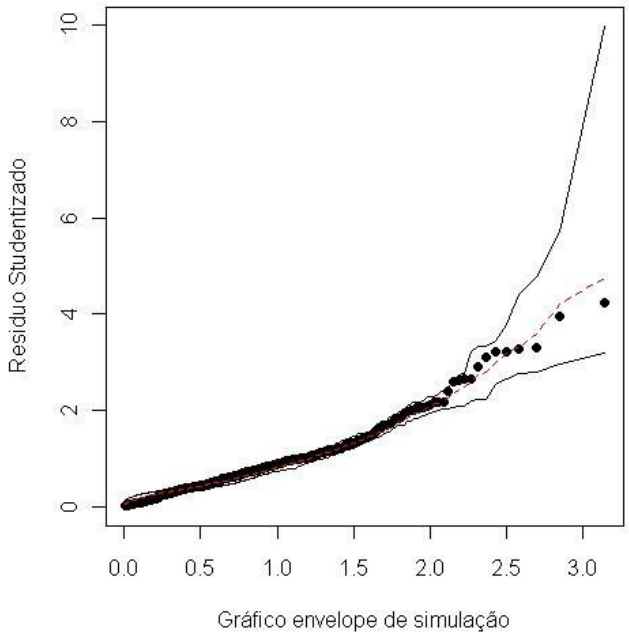

Figura 5.24: Envelope de simulação do modelo marginal com resposta binária ajustado com estrutura de correlação uniforme.

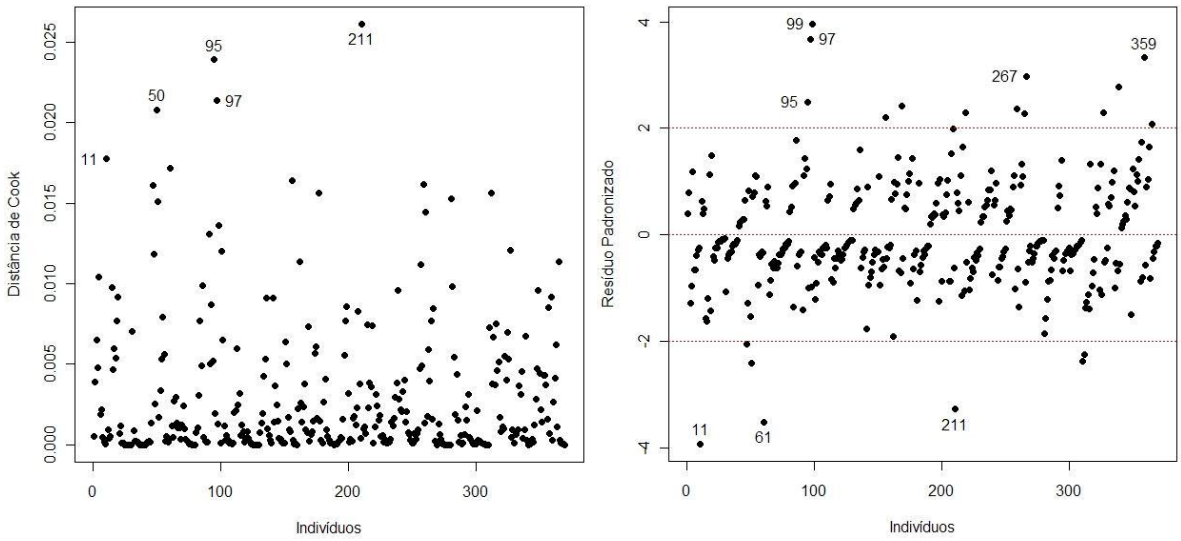

Figura 5.25: Distância de Cook e Resíduos padronizados do modelo marginal com resposta binária ajustado com estrutura de correlação AR-1.

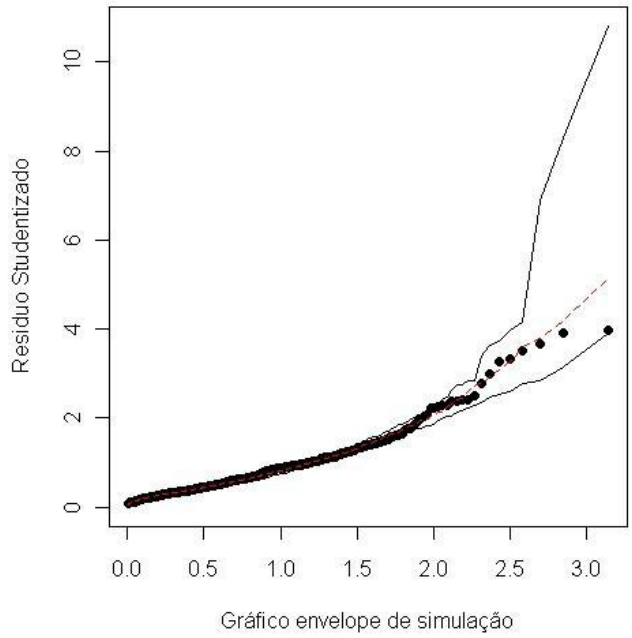

Figura 5.26: Envelope de simulação do modelo marginal com resposta binária ajustado com estrutura de correlação AR-1.

5.3.2 Modelo com efeito aleatório

Por último ajustamos aos dados o modelo misto com intercepto aleatório, conforme expressão vista em (3.11), com estrutura de correlação AR-1, apresentada na seção (2.3.1). Podemos observar que esta estrutura se ajusta bem como mostram as Figuras 5.27 e 5.28, respectivamente.

Tabela 5.6: Estimativas dos parâmetros e P-valores do modelo misto com resposta binária com intercepto aleatório ajustado com estrutura de correlação AR-1.

Coeficientes	**Estimativas**	**P-valor**
(Intercepto)	5,58426	0,51694
x1	0,11207	0,04578
x2	-1,59829	0,25283
x3	0,00530	0,72861
x4	0,02798	0,25593
x5	-0,00198	0,95362
x6	-6,33352	0,21968
x7	0,00485	0,79166
x8	-0,01133	0,80287
x9	-0,00425	0,27496
x10	-0,05446	0,03412
x11	-0,02101	0,41378
t	-0,49936	0,00000

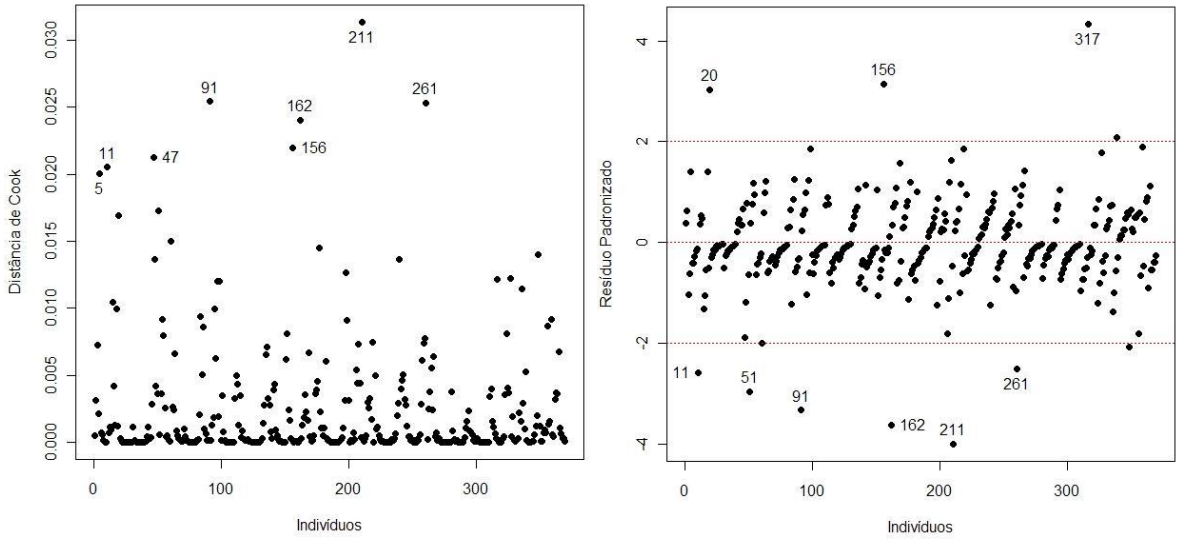

Figura 5.27: Distância de Cook e Resíduos padronizados do modelo misto com resposta binária com intercepto aleatório ajustado com estrutura de correlação AR-1.

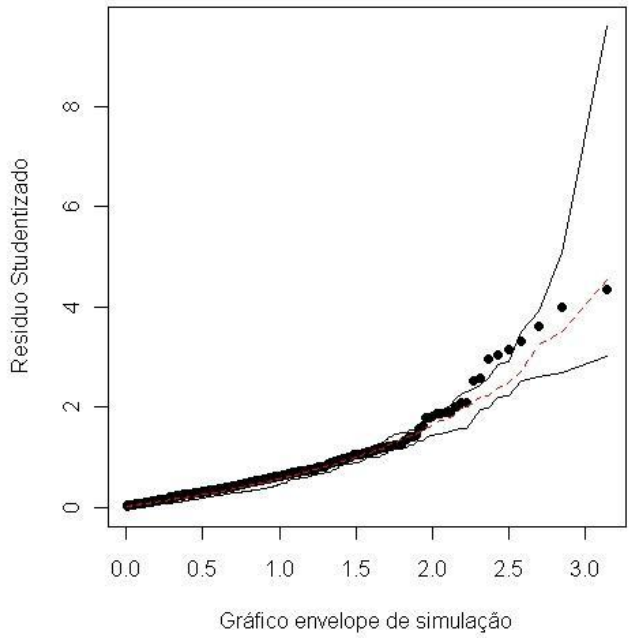

Figura 5.28: Envelope de simulação do modelo misto com resposta binária com intercepto aleatório ajustado com estrutura de correlação AR-1.

5.4 Conclusões

Analisando os modelos aplicados aos dados vistos nas seções anteriores, observou-se que o modelo misto com apenas o intercepto aleatório, tanto com resposta contínua quanto com resposta binária, mostrou-se mais sensível em detectar características típicas individuais, em relação ao modelo marginal. Constatou-se que para a variável resposta dicotomizada, o gráfico envelope de simulação, sob a estrutura de correlação AR-1, manteve em seus intervalos todos os pontos, ao contrário do mesmo gráfico para o caso em que a variável resposta era contínua, em alguns pontos estavam bem distantes dos demais.

Com relação à significância das variáveis, observou-se que idade (x1), sexo (x2) e equilíbrio perna direita (x11), juntamente com o tempo (t), mostraram-se significativas para o modelo, a um nível de probabilidade de aceitação de 0,05, no caso contínuo. Já no caso binário apenas as variáveis idade (x1), equilíbrio perna esquerda (x10) e o tempo foram significativas sob o modelo com estrutura de correlação AR-1.

Assim, neste estudo longitudinal o modelo misto apenas com o intercepto aleatório e estrutura de correlação AR-1 apresentou-se mais adequado para o caso contínuo como para o caso discreto.

Capítulo 6

Conclusões e sugestões futuras

No desenvolvimento deste trabalho, foram encontrados na literatura no que diz respeito à modelagem de dados longitudinais, quando a variável resposta é contínua. Também encontraram-se trabalhos descrevendo as técnicas de seleção e de diagnósticos para este tipo de modelagem. Porém o mesmo não ocorreu quando a variável resposta é binária, pois alguns trabalhos abordam o ajuste de modelo, mas quanto à parte de seleção e diagnóstico, poucos trabalhos apresentavam a metodologia de forma completa e organizada.

As metodologias descritas foram aplicadas a um conjunto de dados reais com o objetivo de utilizar as técnicas de diagnóstico para definir um melhor modelo. No caso do exemplo apresentado, observou-se que o modelo misto com intercepto aleatório e estrutura de correlação AR-1 se adaptou adequadamente aos dados quando a variável resposta é contínua e quando é binária.

Dentre as sugestões para continuidade deste trabalho podem ser citadas, a exploração do modelo misto com covariáveis aleatórias quando a variável resposta é binária e a imputação para observações. Outro tópico de interesse é a parte de diagnóstico e seleção de modelos, que consiste em pesquisar sobre gráficos de resíduos parciais e critério de seleção do tipo AIC para modelos mistos longitudinais assumindo que a variável resposta é binária.

Apêndice A

Conjunto de dados

Tabela A.1: Dados do projeto de reavitalização de adultos/DFisio - UFSCar

o	t	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	y	status
1	1	65	0	140	90	83,70	1,69	60	30	340	5	19	2,02	1
1	2	65	0	130	90	86,50	1,69	56	29	355	10	27	2,06	1
1	3	65	0	140	90	86,50	1,69	64	31	357	21	25	1,89	1
1	4	65	0	130	80	87,30	1,70	64	30	373	24	20	1,97	1
1	5	65	0	140	100	84,90	1,70	60	28	345	24	12	2,11	1
1	6	65	0	130	90	87,60	1,69	56	30	346	23	28	1,76	1
1	7	66	0	150	90	87,00	1,69	76	31	364	24	23	1,86	1
1	8	66	0	130	90	88,50	1,69	72	30	365	28	22	1,75	1
1	9	66	0	120	80	86,20	1,69	84	31	336	30	30	1,69	1
1	10	66	0	130	80	86,60	1,69	72	33	370	29	30	1,63	1
2	1	68	0	120	80	72,10	1,58	68	23	240	20	0	1,94	1
2	2	68	0	100	70	72,90	1,57	80	18	250	8	2	2,07	1
2	3	69	0	115	80	72,40	1,58	88	28	250	6	13	2,42	1
2	4	69	0	120	80	71,80	1,57	68	24	214	4	3	2,16	1
2	5	70	0	110	80	74,40	1,58	64	24	250	7	10	1,99	1
2	6	70	0	130	80	71,80	1,58	76	25	218	14	7	1,71	1
2	7	70	0	110	70	71,20	1,57	84	22	245	19	7	1,85	1
2	8	70	0	120	80	71,00	1,57	78	26	255	5	5	2,05	1
2	9	70	0	140	90	70,00	1,56	80	24	275	5	8	1,79	1
2	10	71	0	120	70	70,00	1,58	84	26	262	9	9	2,14	1

3	1	57	1	130	90	81,60	1,81	72	62	204	27	30	1,79	1
3	2	57	1	140	90	80,20	1,80	84	54	210	30	30	1,69	1
3	3	58	1	120	80	79,40	1,81	80	56	252	29	30	1,77	1
3	4	59	1	130	90	79,70	1,81	64	57	225	30	30	1,72	1
3	5	58	1	120	70	80,00	1,80	64	64	245	30	30	1,70	1
3	6	58	1	110	80	80,80	1,80	60	59	226	30	30	1,61	1
3	7	59	1	140	90	79,70	1,81	72	66	217	30	30	1,58	1
3	8	59	1	120	80	81,40	1,81	80	65	195	30	30	1,51	1
3	9	59	1	120	80	80,80	1,80	80	62	180	30	30	1,46	1
3	10	59	1	120	80	82,10	1,81	80	67	175	30	30	1,42	1
4	1	47	0	120	80	63,80	1,54	100	36	300	30	30	1,71	1
4	2	47	0	90	60	59,60	1,53	100	40	350	30	30	1,61	1
4	3	47	0	100	70	56,80	1,54	88	42	359	30	30	1,94	1
4	4	48	0	105	60	56,80	1,53	84	44	335	30	30	1,74	1
4	5	48	0	110	70	57,10	1,52	88	42	342	30	30	1,56	1
4	6	48	0	90	70	58,10	1,52	72	37	352	30	30	1,49	1
4	7	48	0	100	70	59,10	1,53	80	43	358	30	30	1,64	1
4	8	49	0	105	75	59,10	1,53	80	42	343	30	30	1,71	1
4	9	49	0	100	60	58,70	1,54	80	38	350	30	30	1,41	1
4	10	49	0	100	60	60,70	1,54	80	43	362	30	30	3,13	0
5	1	75	0	130	90	72,90	1,49	80	22	220	21	16	2,22	1
5	2	75	0	130	90	73,00	1,51	84	22	235	25	30	2,20	1
5	3	78	0	130	80	72,90	1,50	72	26	240	30	11	2,36	0
5	4	75	0	120	80	74,80	1,50	68	27	206	3	13	2,01	1
5	5	78	0	130	90	72,80	1,49	72	22	233	4	20	2,26	1
5	6	78	0	130	80	72,50	1,49	60	22	233	18	21	2,03	1
5	7	79	0	140	90	71,10	1,50	60	25	232	8	18	1,94	1
5	8	79	0	120	80	73,00	1,50	60	23	222	10	23	1,84	1
5	9	79	0	140	90	73,50	1,50	64	20	212	8	14	2,05	1
5	10	79	0	160	80	73,50	1,49	60	23	240	15	22	1,87	1
6	1	65	0	120	70	46,90	1,62	76	28	275	16	18	3,02	0
6	2	65	0	120	80	46,40	1,59	80	28	325	25	14	2,40	0
6	3	66	0	110	70	44,60	1,61	84	32	348	18	23	2,35	0
6	4	66	0	120	80	45,40	1,60	68	32	374	27	28	2,24	1
6	5	66	0	130	90	46,70	1,60	80	30	373	20	24	2,12	1
6	6	66	0	140	70	46,90	1,62	68	31	351	23	21	1,84	1
6	7	66	0	115	75	47,50	1,60	68	32	354	30	28	1,49	1
6	8	67	0	120	80	47,60	1,60	84	32	380	30	19	1,97	1
6	9	67	0	120	80	48,10	1,60	68	28	350	30	30	1,75	1
6	10	67	0	150	80	50,00	1,60	72	33	390	30	30	1,59	1
7	1	64	0	130	80	60,10	1,53	68	16	195	30	30	1,96	1
7	2	65	0	170	90	58,90	1,53	60	26	280	29	30	2,16	1
7	3	65	0	120	80	59,90	1,53	80	26	232	30	30	2,04	1
7	4	65	0	110	80	61,30	1,53	64	25	210	30	30	2,09	1
7	5	66	0	110	70	60,00	1,52	72	23	243	30	30	1,89	1
7	6	66	0	130	80	58,50	1,53	72	27	262	30	30	1,90	1
7	7	66	0	100	70	58,30	1,52	64	22	250	30	30	1,61	1
7	8	66	0	115	65	59,40	1,52	72	24	199	30	30	1,85	1
7	9	66	0	125	80	58,30	1,53	68	25	225	30	30	1,73	1
7	10	67	0	140	90	61,60	1,52	68	22	215	30	30	1,73	1

8	1	59	0	120	80	65,20	1,64	68	50	395	30	30	1,78	1
8	2	60	0	120	80	65,00	1,63	72	46	422	30	30	1,91	1
8	3	60	0	110	80	65,00	1,64	68	45	420	30	30	1,85	1
8	4	60	0	110	70	63,30	1,63	76	43	411	30	30	1,70	1
8	5	60	0	120	70	63,00	1,64	72	42	415	30	30	1,72	1
8	6	61	0	120	80	63,90	1,64	68	42	433	30	30	1,86	1
8	7	61	0	120	80	64,00	1,63	64	42	395	30	30	1,74	1
8	8	61	0	110	80	63,80	1,64	60	40	402	30	30	1,60	1
8	9	61	0	115	75	65,30	1,64	72	38	408	30	30	1,66	1
8	10	62	0	110	80	65,30	1,65	68	46	410	30	30	1,50	1
9	1	56	0	120	80	65,20	1,55	80	26	380	20	0	2,27	1
9	2	56	0	130	90	63,00	1,54	72	28	390	13	14	2,12	1
9	3	57	0	120	80	65,10	1,55	68	24	385	30	13	2,08	1
9	4	57	0	130	80	65,30	1,56	68	22	384	29	15	1,99	1
9	5	57	0	140	90	66,00	1,55	60	26	366	28	11	2,26	1
9	6	57	0	120	80	66,40	1,55	60	26	386	17	30	2,05	1
9	7	58	0	120	80	67,00	1,55	60	26	375	24	30	1,92	1
9	8	58	0	120	80	67,40	1,55	64	24	404	25	23	1,82	1
9	9	58	0	130	70	66,80	1,55	72	24	387	26	30	1,95	1
9	10	58	0	140	80	66,90	1,55	60	26	390	22	27	1,65	1
10	1	69	1	110	60	84,00	1,68	80	42	160	27	25	1,90	1
10	2	69	1	120	80	86,80	1,69	76	49	180	28	22	2,71	0
10	3	70	1	105	70	83,70	1,68	64	46	193	28	28	2,06	1
10	4	70	1	115	75	83,80	1,69	64	42	167	29	30	2,01	1
10	5	70	1	110	70	83,00	1,69	44	46	183	30	30	2,12	1
10	6	70	1	120	80	85,10	1,69	68	48	161	30	30	1,92	1
10	7	70	1	120	80	83,20	1,68	64	54	174	30	30	2,01	1
10	8	71	1	120	80	87,00	1,69	68	46	164	30	30	1,75	1
10	9	71	1	120	75	82,90	1,69	64	44	183	26	30	2,08	1
10	10	71	1	100	60	82,80	1,69	64	44	195	30	30	1,78	1
11	1	60	1	120	80	76,00	1,57	84	45	261	12	17	1,65	1
11	2	60	1	140	90	79,50	1,58	68	50	265	26	8	1,86	1
11	3	61	1	130	90	79,50	1,57	68	50	290	26	25	1,60	1
11	4	61	1	120	90	80,50	1,57	64	52	287	30	30	1,57	1
11	5	61	1	130	100	80,90	1,58	84	49	290	30	30	1,41	1
11	6	61	1	130	80	81,40	1,58	76	47	242	28	25	1,41	1
11	7	62	1	140	90	79,80	1,57	72	52	290	30	30	1,30	1
11	8	62	1	130	90	79,60	1,57	72	43	277	30	30	1,34	1
11	9	62	1	120	90	79,80	1,57	80	44	262	29	26	1,28	1
11	10	62	1	160	90	81,30	1,58	72	46	270	30	30	1,22	1
12	1	68	1	120	80	81,50	1,67	68	46	128	26	11	2,01	1
12	2	68	1	140	90	82,30	1,68	72	50	160	24	16	2,27	1
12	3	68	1	120	80	83,40	1,68	68	44	137	5	18	2,38	1
12	4	68	1	120	70	83,60	1,67	72	43	135	23	23	2,00	1
12	5	69	1	130	80	82,90	1,68	76	40	145	14	16	1,93	1
12	6	69	1	140	90	84,00	1,68	64	42	192	29	30	1,92	1
12	7	69	1	140	90	82,20	1,67	68	41	140	21	14	1,86	1
12	8	69	1	140	90	83,50	1,68	72	44	149	30	9	1,85	1
12	9	70	1	130	85	82,80	1,67	72	42	175	14	30	1,91	1
12	10	70	1	140	90	81,50	1,68	80	38	170	12	30	1,71	1

13	1	50	0	100	70	62,30	1,60	60	42	418	30	26	1,73	1
13	2	50	0	100	70	62,30	1,59	72	36	415	30	28	1,71	1
13	3	51	0	110	70	62,80	1,61	76	40	444	27	30	1,80	1
13	4	51	0	90	60	62,40	1,60	84	41	432	30	30	2,00	1
13	5	51	0	100	60	61,50	1,60	68	39	447	30	30	1,72	1
13	6	51	0	110	70	62,10	1,61	76	44	432	30	30	1,54	1
13	7	51	0	105	70	63,00	1,60	64	44	430	30	30	1,47	1
13	8	52	0	100	70	64,00	1,59	68	44	426	30	30	1,65	1
13	9	52	0	100	60	63,80	1,59	76	42	430	30	30	1,42	1
13	10	52	0	110	70	65,30	1,60	80	48	420	30	30	1,53	1
14	1	68	1	140	80	75,10	1,73	64	38	106	21	20	2,25	1
14	2	68	1	150	80	77,30	1,73	60	40	120	17	27	2,34	1
14	3	69	1	150	70	76,40	1,73	64	38	139	20	22	2,53	0
14	4	69	1	140	65	77,70	1,73	64	40	160	13	21	2,44	0
14	5	69	1	170	70	76,10	1,73	64	42	130	13	20	2,16	1
14	6	69	1	140	80	77,80	1,72	64	38	170	22	28	2,19	1
14	7	69	1	130	70	78,80	1,72	60	42	140	19	19	2,00	1
14	8	70	1	140	70	79,60	1,72	60	40	121	19	30	1,96	1
14	9	70	1	130	80	78,80	1,73	60	32	154	30	29	1,98	1
14	10	70	1	140	80	80,50	1,73	64	42	135	30	20	1,95	1
15	1	65	0	120	80	60,90	1,63	72	35	200	30	30	1,92	1
15	2	65	0	120	80	60,10	1,62	76	32	230	23	30	2,55	0
15	3	66	0	120	80	59,20	1,63	68	39	268	30	30	1,93	1
15	4	66	0	120	70	59,60	1,63	68	36	220	30	30	1,87	1
15	5	66	0	120	80	57,00	1,63	64	32	231	30	30	1,83	1
15	6	66	0	120	80	59,50	1,61	72	35	283	30	30	1,73	1
15	7	66	0	120	90	59,30	1,63	68	38	283	30	30	1,73	1
15	8	67	0	120	80	57,50	1,63	80	42	319	30	30	1,76	1
15	9	67	0	120	80	58,70	1,62	72	35	280	30	30	1,63	1
15	10	67	0	130	80	59,80	1,63	68	33	275	30	30	1,43	1
16	1	48	0	110	60	48,50	1,52	72	28	330	30	30	2,19	1
16	2	48	0	120	80	45,60	1,51	64	30	335	30	30	1,89	1
16	3	49	0	110	60	45,60	1,51	72	28	340	30	30	1,76	1
16	4	49	0	90	60	44,80	1,51	64	25	310	30	30	1,73	1
16	5	49	0	110	70	48,50	1,52	64	29	337	30	30	1,57	1
16	6	49	0	120	60	48,10	1,51	68	28	310	30	30	2,03	1
16	7	50	0	100	60	49,10	1,51	60	26	360	30	30	1,73	1
16	8	50	0	100	65	48,60	1,52	64	24	363	30	30	1,76	1
16	9	50	0	100	70	47,90	1,51	80	23	313	30	30	1,57	1
16	10	50	0	100	60	47,70	1,52	80	24	363	30	30	1,56	1
17	1	57	0	120	90	53,20	1,45	100	27	317	30	30	2,33	1
17	2	57	0	120	80	55,50	1,44	96	28	325	24	27	2,90	0
17	3	58	0	110	80	59,50	1,46	84	28	310	30	30	2,73	0
17	4	58	0	120	70	57,60	1,46	72	25	307	30	30	2,14	1
17	5	58	0	130	80	57,60	1,46	64	30	305	30	27	2,25	1
17	6	58	0	120	70	58,00	1,46	76	30	296	30	30	2,13	1
17	7	59	0	120	70	57,50	1,45	72	31	295	30	30	1,92	1
17	8	59	0	120	80	58,70	1,45	72	30	279	30	26	1,84	1
17	9	59	0	120	80	57,40	1,45	76	30	275	27	30	2,02	1
17	10	59	0	100	80	55,90	1,45	68	28	302	30	30	1,81	1

18	1	65	0	110	60	59,70	1,59	72	24	365	28	19	2,55	0
18	2	66	0	140	70	56,80	1,58	76	29	390	25	30	2,78	0
18	3	66	0	120	70	57,30	1,59	76	26	401	30	30	2,42	0
18	4	66	0	110	50	57,40	1,58	76	27	399	24	29	2,26	1
18	5	67	0	120	60	59,00	1,58	68	22	400	30	30	2,20	1
18	6	67	0	120	70	58,80	1,57	76	26	401	30	30	1,82	1
18	7	67	0	140	60	59,30	1,58	76	22	385	30	30	2,10	1
18	8	67	0	125	60	59,20	1,58	80	22	407	30	30	1,80	1
18	9	68	0	120	60	59,10	1,58	72	24	395	30	30	1,98	1
18	10	68	0	120	70	59,80	1,58	84	28	393	30	30	1,89	1
19	1	61	0	110	70	67,20	1,62	68	34	326	27	29	1,83	1
19	2	61	0	110	70	68,90	1,61	84	30	340	11	28	2,14	1
19	3	61	0	100	70	67,10	1,60	76	34	388	28	30	1,85	1
19	4	61	0	110	70	68,40	1,60	72	32	399	30	25	1,94	1
19	5	61	0	110	80	69,40	1,61	80	34	393	30	30	1,93	1
19	6	61	0	100	70	69,10	1,60	72	35	375	30	30	1,87	1
19	7	61	0	130	80	69,00	1,60	64	34	375	30	30	1,67	1
19	8	62	0	100	70	68,00	1,61	68	30	363	30	30	1,61	1
19	9	62	0	105	75	68,70	1,61	72	31	373	30	30	1,75	1
19	10	62	0	110	90	69,00	1,60	88	32	380	30	28	1,70	1
20	1	65	0	130	80	65,20	1,53	96	23	131	23	23	2,28	1
20	2	66	0	130	80	66,00	1,53	76	30	145	30	23	2,51	0
20	3	66	0	120	80	65,70	1,52	96	20	171	30	24	2,24	1
20	4	66	0	130	90	66,00	1,52	92	22	148	30	30	2,44	0
20	5	67	0	140	90	65,30	1,52	88	19	159	30	30	2,28	1
20	6	67	0	130	80	65,70	1,51	100	26	175	26	30	2,24	1
20	7	67	0	120	80	66,00	1,52	100	25	149	30	30	2,14	1
20	8	67	0	120	80	65,10	1,52	96	20	160	30	28	1,95	1
20	9	67	0	145	85	64,90	1,52	100	19	140	30	30	2,03	1
20	10	68	0	140	80	66,40	1,52	88	24	170	30	30	1,99	1
21	1	62	0	130	80	73,50	1,58	88	30	265	30	17	2,59	0
21	2	62	0	135	80	74,40	1,59	96	25	295	29	14	2,55	0
21	3	63	0	130	80	71,30	1,59	84	26	303	16	14	2,24	1
21	4	63	0	120	80	70,30	1,60	76	24	332	30	19	2,37	0
21	5	63	0	130	80	70,00	1,60	76	32	310	14	23	2,22	1
21	6	63	0	130	80	70,50	1,59	76	32	318	14	24	1,99	1
21	7	63	0	140	90	70,00	1,59	72	25	330	30	24	1,91	1
21	8	64	0	120	70	69,30	1,59	88	30	319	24	9	2,14	1
21	9	64	0	120	70	68,70	1,60	80	26	310	22	22	2,14	1
21	10	64	0	140	80	71,10	1,59	76	28	300	26	30	1,92	1
22	1	55	0	140	70	72,60	1,53	72	28	229	19	5	3,05	0
22	2	56	0	130	80	72,50	1,55	84	30	230	23	16	2,63	0
22	3	56	0	120	80	73,80	1,55	68	28	230	11	19	2,32	1
22	4	56	0	135	80	74,60	1,54	68	28	240	11	6	2,32	1
22	5	56	0	140	80	73,40	1,56	60	33	267	7	28	2,19	1
22	6	57	0	130	80	73,50	1,55	64	27	268	19	20	1,89	1
22	7	57	0	130	80	71,70	1,54	68	31	320	16	15	2,04	1
22	8	57	0	140	80	72,40	1,55	60	31	275	18	9	1,99	1
22	9	57	0	120	75	72,50	1,54	68	30	270	14	17	2,04	1
22	10	58	0	130	80	74,00	1,55	72	32	290	24	25	1,64	1

23	1	58	0	140	90	58,80	1,47	68	26	380	30	30	2,04	1
23	2	59	0	100	70	59,60	1,48	76	28	390	30	30	1,94	1
23	3	59	0	110	80	56,70	1,47	72	26	425	30	30	1,78	1
23	4	59	0	110	80	56,40	1,47	64	26	431	30	30	1,85	1
23	5	60	0	130	75	54,50	1,47	64	27	425	30	30	1,67	1
23	6	60	0	110	70	54,30	1,47	60	26	426	30	30	1,58	1
23	7	60	0	130	80	55,10	1,46	80	30	418	30	30	1,88	1
23	8	60	0	110	80	56,90	1,48	76	24	425	30	30	1,54	1
23	9	60	0	120	80	55,70	1,47	92	24	397	30	30	1,70	1
23	10	61	0	110	70	56,60	1,47	60	22	415	30	30	1,51	1
24	1	60	0	150	90	67,30	1,45	80	28	373	5	9	2,07	1
24	2	60	0	140	90	70,80	1,44	76	28	370	17	12	2,42	0
24	3	61	0	140	90	69,90	1,46	88	30	358	16	4	2,37	0
24	4	61	0	150	90	72,20	1,46	56	29	350	29	9	2,35	0
24	5	61	0	130	80	72,00	1,45	80	30	319	17	5	2,36	0
24	6	61	0	140	80	74,10	1,46	64	26	347	10	14	2,17	1
24	7	62	0	140	90	74,10	1,45	72	29	383	17	8	2,34	1
24	8	62	0	160	90	74,00	1,44	76	30	360	21	14	2,08	1
24	9	62	0	160	90	73,70	1,45	84	29	330	11	28	2,16	1
24	10	62	0	150	80	71,70	1,44	72	32	360	9	19	1,96	1
25	1	59	0	110	70	94,80	1,55	72	28	265	30	25	2,34	1
25	2	59	0	130	90	91,80	1,54	76	30	360	30	30	2,21	1
25	3	59	0	130	90	88,60	1,55	72	32	335	30	30	2,07	1
25	4	59	0	125	80	91,70	1,54	64	33	350	30	30	1,89	1
25	5	59	0	120	80	89,60	1,53	80	35	335	26	29	1,68	1
25	6	60	0	140	80	91,50	1,55	80	34	355	30	30	1,35	1
25	7	60	0	130	80	93,60	1,55	68	32	338	30	30	1,56	1
25	8	60	0	125	90	94,70	1,54	76	29	328	30	30	1,74	1
25	9	60	0	120	80	93,30	1,53	60	28	321	30	30	1,74	1
25	10	61	0	120	80	92,40	1,53	84	34	350	29	30	1,51	1
26	1	69	0	140	100	75,60	1,66	68	27	312	15	3	2,17	1
26	2	69	0	120	80	79,00	1,66	72	30	275	29	8	2,24	1
26	3	69	0	130	90	80,10	1,64	72	31	271	6	6	2,51	0
26	4	70	0	140	90	79,50	1,66	68	26	305	25	12	2,43	0
26	5	70	0	155	100	79,30	1,66	64	30	330	15	5	2,28	1
26	6	70	0	140	90	75,00	1,66	68	33	311	6	25	2,08	1
26	7	70	0	140	100	67,00	1,64	64	32	315	26	3	2,28	1
26	8	71	0	130	80	68,90	1,65	68	30	329	30	8	1,97	1
26	9	71	0	120	80	66,90	1,65	68	30	332	10	30	2,18	1
26	10	71	0	130	90	66,60	1,65	64	30	330	7	28	1,83	1
27	1	57	0	90	60	55,10	1,63	84	20	320	30	30	1,98	1
27	2	58	0	110	70	56,20	1,61	68	26	310	30	30	2,21	1
27	3	58	0	90	60	56,90	1,63	72	26	326	30	30	2,03	1
27	4	58	0	110	60	57,20	1,64	76	26	320	30	30	2,12	1
27	5	58	0	90	60	58,00	1,62	72	30	318	30	30	2,19	1
27	6	59	0	90	60	57,80	1,64	76	26	300	30	30	1,98	1
27	7	59	0	100	70	59,10	1,62	64	29	305	30	30	2,17	1
27	8	59	0	100	60	58,80	1,64	80	28	320	30	30	1,92	1
27	9	59	0	110	70	58,20	1,63	88	23	330	30	30	1,87	1
27	10	60	0	100	60	60,50	1,63	80	27	313	30	30	1,77	1

28	1	56	1	100	70	62,10	1,66	84	50	193	30	30	1,48	1
28	2	57	1	110	60	60,30	1,66	72	50	190	30	30	1,73	1
28	3	57	1	110	60	58,20	1,66	60	47	212	30	30	1,44	1
28	4	57	1	90	60	58,80	1,66	76	49	263	30	30	1,33	1
28	5	58	1	100	60	58,00	1,66	64	48	240	30	30	1,32	1
28	6	58	1	100	60	58,50	1,65	60	52	230	30	30	1,47	1
28	7	58	1	100	70	59,20	1,65	64	56	257	30	30	1,50	1
28	8	58	1	100	70	60,00	1,65	68	52	182	30	30	1,25	1
28	9	58	1	90	60	59,60	1,65	72	48	177	30	30	1,25	1
28	10	59	1	110	70	61,30	1,66	60	52	215	30	30	1,13	1
29	1	58	0	120	80	71,80	1,67	80	31	158	12	27	1,92	1
29	2	58	0	120	80	71,40	1,65	76	24	135	24	30	1,83	1
29	3	58	0	130	80	71,30	1,66	76	28	151	26	30	1,71	1
29	4	59	0	130	80	73,50	1,64	76	33	180	25	30	1,67	1
29	5	59	0	140	80	73,20	1,64	68	36	188	11	30	1,59	1
29	6	59	0	130	90	72,70	1,64	76	27	238	28	28	1,42	1
29	7	59	0	120	80	70,90	1,63	64	34	250	24	24	1,63	1
29	8	60	0	130	90	71,50	1,66	64	30	205	19	30	1,47	1
29	9	60	0	120	80	71,40	1,63	68	28	254	30	27	1,50	1
29	10	60	0	140	90	70,80	1,65	64	30	275	30	30	1,47	1
30	1	54	0	130	90	69,90	1,56	72	28	236	30	30	2,17	1
30	2	55	0	120	80	70,60	1,54	64	38	259	30	30	2,10	1
30	3	54	0	140	90	71,20	1,54	80	37	258	30	30	2,03	1
30	4	55	0	120	80	70,40	1,55	68	36	252	30	30	2,08	1
30	5	55	0	120	80	68,70	1,54	76	36	285	30	30	1,93	1
30	6	56	0	110	70	71,10	1,55	68	32	245	30	26	1,85	1
30	7	56	0	120	80	68,20	1,54	60	36	251	30	30	1,91	1
30	8	56	0	120	80	69,10	1,55	76	34	253	30	30	1,84	1
30	9	56	0	125	80	69,60	1,54	68	34	232	30	30	1,85	1
30	10	57	0	110	70	66,30	1,53	80	34	272	30	22	1,77	1
31	1	56	1	130	90	81,30	1,73	68	50	169	30	29	1,44	1
31	2	57	1	110	80	80,70	1,72	68	52	172	30	30	1,83	1
31	3	57	1	120	80	82,40	1,72	60	52	165	30	30	1,69	1
31	4	58	1	110	80	81,80	1,72	68	51	132	30	30	1,70	1
31	5	58	1	130	80	82,30	1,72	76	55	170	30	30	1,51	1
31	6	58	1	120	80	83,20	1,72	72	52	170	30	30	1,50	1
31	7	58	1	140	80	83,20	1,72	76	56	145	30	30	1,61	1
31	8	59	1	130	80	83,80	1,72	80	52	153	30	30	1,44	1
31	9	59	1	120	80	83,50	1,72	72	52	155	30	30	1,50	1
31	10	59	1	110	80	79,50	1,72	68	51	142	30	30	1,54	1
32	1	63	0	120	80	67,10	1,52	80	28	356	18	8	1,88	1
32	2	63	0	140	80	70,80	1,50	84	32	340	28	4	1,92	1
32	3	63	0	130	90	72,80	1,52	72	32	355	14	12	1,95	1
32	4	64	0	115	75	73,90	1,51	64	30	320	17	3	1,95	1
32	5	64	0	130	90	72,50	1,50	60	28	350	15	11	1,90	1
32	6	64	0	130	80	72,30	1,50	72	28	333	4	8	1,88	1
32	7	64	0	120	80	73,80	1,51	76	27	330	18	9	2,04	1
32	8	65	0	145	85	75,80	1,51	76	28	345	6	27	1,73	1
32	9	65	0	125	80	75,90	1,50	68	28	260	7	12	1,66	1
32	10	65	0	120	90	74,70	1,51	68	26	305	7	24	1,80	1

33	1	51	0	140	90	71,30	1,56	64	36	150	30	21	2,20	1
33	2	51	0	160	100	71,70	1,55	48	32	145	26	21	2,19	1
33	3	52	0	120	70	71,00	1,56	68	34	170	30	10	2,13	1
33	4	52	0	120	70	70,10	1,55	64	36	198	22	24	1,95	1
33	5	52	0	140	80	70,20	1,54	68	35	170	25	30	2,42	0
33	6	52	0	130	80	71,60	1,55	68	32	140	29	21	1,98	1
33	7	53	0	120	80	71,00	1,54	56	32	187	30	30	2,09	1
33	8	53	0	110	70	70,00	1,54	64	34	151	30	30	1,70	1
33	9	53	0	140	90	69,50	1,55	64	32	90	27	30	1,69	1
33	10	53	0	120	70	69,40	1,55	56	34	135	30	30	1,45	1
34	1	55	0	130	90	70,80	1,54	76	38	240	30	20	2,26	1
34	2	55	0	140	90	70,40	1,54	76	36	214	25	30	2,36	0
34	3	56	0	140	90	72,50	1,55	72	36	237	27	30	2,20	1
34	4	56	0	140	80	71,70	1,55	76	40	233	30	30	2,13	1
34	5	56	0	140	90	72,80	1,55	64	33	237	30	30	2,21	1
34	6	56	0	160	100	72,90	1,55	72	34	230	30	30	2,00	1
34	7	57	0	140	90	72,50	1,55	64	34	235	29	28	1,91	1
34	8	57	0	140	80	70,20	1,55	80	38	235	30	30	1,96	1
34	9	57	0	120	80	68,80	1,54	68	30	220	30	30	2,21	1
34	10	57	0	140	80	69,50	1,54	80	36	260	30	30	1,93	1
35	1	69	0	150	90	61,50	1,52	76	19	161	23	9	2,19	1
35	2	69	0	130	80	63,00	1,51	68	16	185	25	19	2,52	0
35	3	69	0	130	90	62,30	1,53	64	19	204	19	10	2,34	1
35	4	69	0	120	80	62,40	1,52	76	21	212	30	6	2,42	0
35	5	70	0	140	80	63,00	1,51	68	19	215	26	7	2,23	1
35	6	70	0	130	80	62,30	1,51	64	21	211	30	30	2,37	0
35	7	70	0	120	80	62,00	1,51	64	25	200	30	23	2,11	1
35	8	70	0	120	70	62,70	1,51	64	20	186	27	6	2,02	1
35	9	71	0	140	80	61,00	1,50	72	18	210	13	26	1,98	1
35	10	71	0	130	80	62,30	1,50	72	21	200	20	20	2,06	1
36	1	62	0	120	80	65,00	1,64	84	40	320	30	26	2,22	1
36	2	62	0	120	70	66,40	1,65	84	38	303	11	12	2,43	0
36	3	62	0	110	70	70,00	1,63	88	39	347	27	22	2,20	1
36	4	62	0	110	70	68,50	1,65	72	37	283	24	10	2,49	0
36	5	63	0	120	70	69,40	1,63	64	35	335	20	21	2,40	0
36	6	63	0	120	80	69,70	1,64	64	36	320	14	22	1,99	1
36	7	63	0	130	90	68,60	1,66	80	34	305	21	16	2,21	1
36	8	63	0	120	70	69,20	1,63	80	37	346	24	22	1,99	1
36	9	64	0	110	70	68,70	1,63	88	32	347	28	27	2,09	1
36	10	64	0	110	70	69,30	1,63	88	33	345	22	20	1,94	1
37	1	51	0	130	90	70,20	1,52	76	40	355	30	30	2,28	1
37	2	51	0	120	70	69,80	1,52	88	39	360	30	30	2,15	1
37	3	52	0	110	80	71,10	1,54	88	38	375	30	30	2,06	1
37	4	52	0	110	80	69,80	1,52	84	40	363	30	30	1,89	1
37	5	52	0	125	90	70,20	1,51	88	37	386	30	30	2,26	1
37	6	52	0	120	80	70,90	1,52	60	42	365	30	30	1,76	1
37	7	53	0	130	80	72,80	1,52	72	40	352	30	30	1,69	1
37	8	53	0	120	80	72,20	1,52	60	40	355	30	30	1,64	1
37	9	53	0	120	90	73,00	1,52	88	39	360	30	27	1,57	1
37	10	53	0	120	90	70,80	1,52	60	40	360	30	30	1,54	1

Apêndice B

Comandos no R

```
##### ANÁLISE EXPLORATÓRIA #####
# INSERINDO OS DADOS
rm(list=ls())
ls()
options(digits=4)
dados<-read.table("ed.txt",header=T)
dados$x2<-factor(dados$x2, labels = c("F", "M"))
attach(dados)
dados[1:5,]

# RESUMO ESTATÍSTICO
summary(dados)

# GRÁFICO DE DISPERSÃO
pairs(dados[,c(-1,-2,-4,-15)])

# GRÁFICOS - HISTOGRAMA E BOXPLOT
# O procedimento para gerar os gráficos - histograma e boxplot é o mesmo para
# as variáveis x3, x4, x5, x6, x7, x8, x9, x10, x11, y.

boxplot(x1 ~ t, data = dados, col="light blue", outline = TRUE,
xlab="Tempo", ylab="x1")
xyplot(x1 ~ t, groups = o, data = dados, type = "l",
xlab="Tempo", ylab="x1")
```

```
##### AJUSTANDO O MODELO MARGINAL #####
# AJUSTE DO MODELO QUANDO A VARIÁVEL RESPOSTA É CONTÍNUA
# ESTRUTURA DE CORRELAÇÃO UNIFORME
# ENTRADA DE DADOS
rm(list=ls())
ls()
options(digits=4)
require(nlme)
require(car)
require(MASS)
require(gee)
require(lattice)
dados<-read.table("ed.txt", header=T)
dados<-dados[,-15]
dados[1:5,]

# ELIMINANDO UM INDIVÍDUO
#dados<-dados[-(31:40),]

# AJUSTANDO UM MODELO COM ESTRURUTA DE CORRELAÇÃO UNIFORME
fit.gee<-gee(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+t, id=o, data=dados,
family = "gaussian", corstr="exchangeable")
summary(fit.gee)
fit<-summary(fit.gee)
fit$coef[,c(1,5)]
n<-nrow(dados)/10
t<-as.vector(rep(10,n))

# DISTÂNCIA DE COOK E RESÍDUO PADRONIZADO
X1<-dados[,c(-1,-2,-14)]
intercept<-rep(1,nrow(X1))
X<-cbind(intercept,X1)
X<-t(t(X))
y<-fit.gee$y
beta<-fit.gee$coef
R<-fit.gee$work
mi<-fitted(fit.gee)
N<-nrow(X)
p<-ncol(X)

# RESÍDUO DE PEARSON
r<-(y-mi)

# CÁLCULO DO phi
invphi<-(sum(r^2)/(N-p))
phi<-1/invphi
```

```
# MATRIZ C
A<-diag(1,N)
C<-A

# MATRIZ OMEGA (variancia e covariancia de y)
Omega<-matrix(0,N,N)
invOmega<-matrix(0,N,N)
i<-1
l<-1
while (l<N+1)
{
Omega[l:(l+t[i]-1),l:(l+t[i]-1)]<-sqrt(A[l:(l+t[i]-1),l:(l+t[i]-1)])%*%R%*%
sqrt(A[l:(l+t[i]-1),l:(l+t[i]-1)])
invOmega[l:(l+t[i]-1),l:(l+t[i]-1)]<-solve(Omega[l:(l+t[i]-1),l:(l+t[i]-1)])
l<-l+t[i]
i<-i+1
}

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

# MATRIZ H e W
W<-C%*%invOmega%*%C
H<-solve(t(X)%*%W%*%X)
raizW<-matrix(0,N,N)
i<-1
l<-1
while(l<N+1)
{
auto<-eigen(W[l:(l+t[i]-1),l:(l+t[i]-1)])
raizW[l:(l+t[i]-1),l:(l+t[i]-1)]<-auto$vectors%*%sqrt(diag(auto$values))%*
%t(auto$vectors)
l<-l+t[i]
i<-i+1
}
H<-raizW%*%X%*%H%*%t(X)%*%raizW
h<-diag(H)

# RESÍDUO PADRONIZADO
rsd<-as.vector(rep(0,N))
part.rsd<-raizW%*%solve(C)%*%(y-mi)
for (l in 1:N)
{
e<-as.vector(rep(0,N))
e[l]<-1
rsd[l]<-t(e)%*%part.rsd/sqrt(1-h[l])
}
```

```
# DISTÂNCIA DE COOK
cd<-as.vector(rep(0,N))
for(l in 1:N) cd[l]<-(rsd[l]^2*h[l])/((1-h[l])*p)

# CONSTRUÇÃO DOS GRÁFICOS DC e RSD
plot(cd,xlab="Indivíduos",ylab="Distância de Cook", pch=16)
identify(cd)
plot(rsd,xlab="Indivíduos",ylab="Resíduo Padronizado",pch=16)
abline(h=0,lty=3, col=2)
abline(h=2,lty=3, col=2)
abline(h=-2,lty=3, col=2)
identify(rsd)

# GERAR RESPOSTAS CORRELACIONADAS
dad.fit<-fitted(fit.gee)
media<-as.vector(tapply(dad.fit, list(dados[,2]), mean))
sd<-as.vector(sqrt(tapply(y, list(dados[,2]), var)))
repl<-25
random.y<-array(dim=c(N,repl))
sim.y1<-matrix(0,N,1)
for(i in 1:repl){
    for(j in 1:t[i]){
        for(k in 1:n) sim.y<-abs(rnorm(n, media[j],sd[j]))
        sim.y1[(1+n*(j-1)):(n*j),1]<-sim.y
        j<-j+1
    }
    random.y[,i]<-c(sim.y1)
    i<-i+1
}
random.y[1:10,]

# CONSTRUÇÃO DO GRÁFICO - ENVELOPE DE SIMULAÇÃO
orig.res<-rsd
ABSorig.res<-abs(orig.res)
SORTorig.res<-sort(ABSorig.res)
dados2<-cbind(dados,random.y)
attach(dados2)
random.res<-array(dim=c(N,repl))
for(k in 1:repl)
{
temp.fit.gee<-gee(random.y[,k]~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+t,
id=o, data=dados2, family = "gaussian", corstr="exchangeable")
y<-temp.fit.gee$y
beta<-coef(temp.fit.gee)
R<-temp.fit.gee$work
mi<-fitted(temp.fit.gee)
```

```
# CÁLCULO DO RESÍDUO DE PEARSON
r<-y-mi

# CÁLCULO DE phi
invphi<-(sum(r^2)/(N-p))
phi<-1/invphi

# MATRIZ C
A<-diag(1,N)
C<-A

# MATRIZ OMEGA (variancia e covariancia de y)
Omega<-matrix(0,N,N)
invOmega<-matrix(0,N,N)
i<-1
l<-1
while (l<N+1)
{
Omega[l:(l+t[i]-1),l:(l+t[i]-1)]<-sqrt(A[l:(l+t[i]-1),l:(l+t[i]-1)])%*%R
%*%sqrt(A[l:(l+t[i]-1),l:(l+t[i]-1)])
invOmega[l:(l+t[i]-1),l:(l+t[i]-1)]<-solve(Omega[l:(l+t[i]-1),l:(l+t[i]-1)])
l<-l+t[i]
i<-i+1
}

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

# MATRIZ H e W
W<-C%*%invOmega%*%C
H<-solve(t(X)%*%W%*%X)
raizW<-matrix(0,N,N)
i<-1
l<-1
while(l<N+1)
{
auto<-eigen(W[l:(l+t[i]-1),l:(l+t[i]-1)])
raizW[l:(l+t[i]-1),l:(l+t[i]-1)]<-auto$vectors%*%sqrt(diag(auto$values))%*%
t(auto$vectors)
l<-l+t[i]
i<-i+1
}
H<-raizW%*%X%*%H%*%t(X)%*%raizW
h<-diag(H)
```

```
random.rsd<-as.vector(rep(0,N))
part.rsd<-raizW%*%solve(C)%*%(y-mi)
for (l in 1:N)
{
e<-as.vector(rep(0,N))
e[l]<-1
random.rsd[l]<-t(e)%*%part.rsd/sqrt(1-h[l])
}
random.res[,k]<-random.rsd
}
random.res[1:10,]

ABSrandom.res<-abs(random.res)
SORTrandom.res<-array(dim=c(N,repl))
for(k in 1:repl) SORTrandom.res[,k]<-sort(ABSrandom.res[,k])
descritiva<-array(dim=c(N,3))
for(k in 1:N)
{
descritiva[k,1]<-min(SORTrandom.res[k,])
descritiva[k,2]<-median(SORTrandom.res[k,])
descritiva[k,3]<-max(SORTrandom.res[k,])
}
Z<-array(dim=c(N,1))
for(i in 1:N) Z[i]<-qnorm((i+N-1/8)/(2*N+1/2))

final<-cbind(Z, descritiva, SORTorig.res)
faixa<-range(final[,5], final[,2], final[,4])
par(mfrow=c(1,1))
par(pty="s")
plot(final[,1], final[,5], xlab="Gráfico envelope de simulação",
ylab="Residuo Studentizado", ylim=faixa, pch=16)
par(new=T)
lines(final[,1], final[,2])
lines(final[,1], final[,3], lty=2, col=2)
lines(final[,1], final[,4])
```

```
##### NOTAS COMPLEMENTARES #####
# NOTA 1: A estrutura básica deste programa é a mesma quando modelamos os dados
# com a estrutura de correlação AR-1, mudando apenas a matriz de correlação.

# NOTA 2: Quando modelamos os dados com o modelo misto, o programa base é
# o mesmo usado na modelagem marginal, fazendo algumas modificações:
# NO AJUSTE (estrutura de correlação AR-1)
fit.lme<-lme(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+t,
random = ~1 | o, data = dados)
summary(fit.lme)
fit.lme2<-update(fit.lme, correlation=corAR1())
fit.lme3<-intervals(fit.lme2)[3]
fit<-summary(fit.lme)
names(fit)
round(fit$tTable[,c(1,5)],5)
fit.lme3$cor[2]

R.AR1<-diag(t[1])
q<-2
a<-fit.lme3$cor[2]
for(i in 1:(t[1]-1)) {
  for(j in q:t[1])R.AR1[i,j]<-a^(abs(j-i))
  q<-q+1
  }
R.AR1<-R.AR1+t(R.AR1)-diag(nrow(R.AR1))

# RESÍDUO DE PEARSON
r<-fit.lme$res[,2]

#GRÁFICO - ENVELOPE DE SIMULAÇÃO
orig.res<-rsd
ABSorig.res<-abs(orig.res)
SORTorig.res<-sort(ABSorig.res)
dados2<-cbind(dados,random.y)
attach(dados2)
random.res<-array(dim=c(N,repl))
for(k in 1:repl)
{
temp<-random.y[,k]
temp.fit.lme<-lme(temp~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+t, random = ~1 | o,
data = dados2)
temp.fit.lme2<-update(temp.fit.lme, correlation=corAR1())
temp.fit.lme3<-intervals(temp.fit.lme2)[3]
y<-random.y[,k]
b<-summary(fit.lme2)
beta<-b$tTable[,1]
mi<-temp.fit.lme$fitt[,2]
```

```
R.AR1<-matrix(0,t[1],t[1])
q<-2
a<-temp.fit.lme3$cor[2]
for(i in 1:(t[1]-1)) {
  for(j in q:t[1])R.AR1[i,j]<-a^(abs(j-i))
  q<-q+1
      }
R<-R.AR1+t(R.AR1)-diag(nrow(R.AR1))

# RESÍDUO DE PEARSON
r<-temp.fit.lme$res[,2]

# NOTA 3: Na modelagem dos dados com o modelo marginal logístico, a estrutura
# básica do programa (modelo marginal contínuo) sofre algumas modificações:
# NO AJUSTE (estrutura de correlação AR-1)
fit.gee<-gee(status ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+t, id=o,
data=dados, family = "binomial", corstr="AR-M", Mv=1, scale.fix=T,
scale.value=1)

# MATRIZ C
A<-diag(mi*(1-mi),N)
C<-A

# GERAR RESPOSTAS BINÁRIAS CORRELACIONADAS
repl<-25
random.y<-array(dim=c(N,repl))
dif<-matrix(0,N,1)
for (i in 1:repl){
        for (j in 1:N){
        dif<-runif(N)-mi
        dif[dif>=0]<-0
        dif[dif<0]<-1
        j<-j+1
        }
        random.y[,i]<-dif
        i<-i+1
}
random.y[1:10,]

# GRÁFICO - ENVELOPE SIMULADO
orig.res<-rsd
ABSorig.res<-abs(orig.res)
SORTorig.res<-sort(ABSorig.res)
dados2<-cbind(dados,random.y)
attach(dados2)
random.res<-array(dim=c(N,repl))
```

```
for(k in 1:repl)
{
temp.fit.gee<-gee(random.y[,k] ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+t, id=o,
data=dados2, family = "binomial", corstr="AR-M", Mv=1, scale.fix=T,
scale.value=1)
summary(temp.fit.gee)
y<-temp.fit.gee$y
beta<-coef(temp.fit.gee)
R<-temp.fit.gee$work
mi<-fitted(temp.fit.gee)

# MATRIZ C
A<-diag(mi*(1-mi),N)
C<-A
}

# NOTA 4: No modelo logístico misto a estrutura básica do programa (modelo
# marginal) se repete sofrendo algumas modificações para uso da modelagem
# dos dados:
# NO AJUSTE (estrutura de correlação ar-1)
formula<-status ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+t
fit.glmmPQL2<- glmmPQL(formula, random = ~1 | o, family = binomial,
data = dados, corr = corAR1(form=~t))
fit.glmmPQL3<-intervals(fit.glmmPQL2)
fit<-summary(fit.glmmPQL2)
round(fit$tTable[,c(1,5)],5)
fit.glmmPQL3$cor[2]

R.AR1<-diag(t[1])
q<-2
a<-fit.glmmPQL3$cor[2]
for(i in 1:(t[1]-1)) {
  for(j in q:t[1])R.AR1[i,j]<-a^(abs(j-i))
  q<-q+1
  }
R.AR1<-R.AR1+t(R.AR1)-diag(nrow(R.AR1))

# MATRIZ C
A<-diag(mi*(1-mi),N)
C<-A

# GRÁFICO - ENVELOPE DE SIMULAÇÃO
orig.res<-rsd
ABSorig.res<-abs(orig.res)
SORTorig.res<-sort(ABSorig.res)
dados2<-cbind(dados,random.y)
attach(dados2)
```

```
random.res<-array(dim=c(N,repl))
k<-1
while(k < repl+1)
{

temp<-random.y[,k]
temp.fit.glmmPQL2<- glmmPQL(temp~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+t,
    random = ~1 | o, family = binomial, data = dados, corr = corAR1(form=~t))
temp.fit.glmmPQL3<-intervals(temp.fit.glmmPQL2)

R.AR1<-matrix(0,t[1],t[1])
q<-2
a<-temp.fit.glmmPQL3$cor[2]
for(i in 1:(t[1]-1)) {
  for(j in q:t[1])R.AR1[i,j]<-a^(abs(j-i))
  q<-q+1
      }
R<-R.AR1+t(R.AR1)-diag(nrow(R.AR1))

y<-random.y[,k]
b<-summary(fit.glmmPQL2)
beta<-b$tTable[,1]
mi<-exp(temp.fit.glmmPQL2$fitt[,2])/(1+exp(temp.fit.glmmPQL2$fitt[,2]))

################################### F I M #######################################
```

Bibliografia

[1] ARTES, Rinaldo e BOTTER, Denise A. Funções de estimação em modelos de regressão. São Paulo: ABE, 2005.

[2] BAIA, Lusane Leão. As equações de estimação generalizadas e aplicações. Dissertação, UNICAMP. Campinas - SP, 1997.

[3] BANERJEE, M. & FREES, E.W. Infuence Diagnostics for Linear Longitudinal Models. *Journal of the American Statistical Association* **92**, 999-1005, 1997.

[4] CHATTERJEE, S. & HADI, A.S. Infuential Observations, High Leverage Points, and Outliers in Linear Regression (with discussion). *Statistical Science* **1**, 379-393, 1986.

[5] CHATTERJEE, S. & HADI, A.S. Sensitivity Analysis in Linear Regression. New York: John Wiley & Sons, 1988.

[6] CHRISTENSEN, R. & PEARSON, L.M. Case-deletion diagnostics for mixed models. *Technometrics* **34**, 38-45, 1992.

[7] CLEVELAND, W.S. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, p. 829-836, 1979.

[8] CORDEIRO, Gauss Moutinh e NETO, Eufrásio de A. L. Modelos paramétricos. Recife: UFPE, 2004.

[9] COSTA, Silvano Cesar da. Modelos lineares generalizados mistos para dados longitudinais. Tese, ESALQ-USP. São Paulo, 2003.

[10] DAVID, Jacqueline Sant' E. (1999). Regressão logística, regressão de poisson e modelos lineares generalizados. Iniciação científica, IME-USP. São Paulo, 1999.

[11] DEMÉTRIO, Clarice G. B. Modelos lineares generalizados em experimentação agronômica, ESALQ-USP. São Paulo, 2002.

[12] DIGGLE, Peter J.; LIANG, Kung-Yee and ZEGER, Scott L. Analysis of longitudinal data, Inglaterra: Oxford University Press Inc., 1996.

[13] FARHAT, Cecília A.V. Análise de diagnóstico em regressão logística. Dissertação, IME-USP. São Paulo, 2003.

[14] FIRTH, D. Generalized linear models. EUA: Chapman & Hall, 1991.

[15] HEDEKER, Donald and GIBBONS, Robert D. Longitudinal Data Analysis. John Wiley & Sons, Inc., 2006.

[16] HENDERSON, C.R. Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* **31**, 423-447, 1975.

[17] JONES, R. H. Longitudinal Data with Serial Correlation: A State-Space Approach. London: Chapman & Hall, 1993.

[18] KOCH, G. C.; LANDIS, J. R.; FREEMAN, J. L.; FREEMAN, D. H. and LEHMAN, R. B. A general methodology for the analysis of repeated measuremenst of categorial data. *Biometrics* **33**, p. 133-158, 1977.

[19] LAIRD, N. M. and WARE, J. H. Random-effects models for longitudinal data. *Biometrics* **38**, p. 963-974, 1982.

[20] LANDWEHR, J.M.; PREGIBON, D. and SHOEMAKER, A. C. Graphical methods for asseing logistic regression models. *Journal of the American Statistical Association* **79**, p. 385 61-71, 1984.

[21] LARA, Idemauro A. R. Modelos de transição para dados binários. Tese, ESALQ-USP. São Paulo, 2007.

[22] LIANG, Kung-Yee and ZEGER, Scott L. Longitudinal analysis using generalized linear models. *Biometrika* **73**, p. 13-22, 1986.

[23] LIU, C. & RUBIN, D.B.. The ECME algorithm: A simple Extension of EM and ECM with Faster Monotone Convergence. *Biometrika* **81**, 633-648, 1994.

[24] MATSUDO, Sandra Marcela M. Avaliação do idoso. Londrina: Micrograf, 2000.

[25] McLACHLAN, G.J. & KRISHNAN, T. The EM algorithm and extensions. New York: John Wiley & Sons. 1997.

[26] McCULLAGH, P.; NELDER, J.A. Generalized linear models. 2nd ed. London: Chapman and Hall, 511p, 1989.

[27] MENG, Xiao-Li & VAN DYK, D. Fast EM-type implementations for mixed efects models. *Journal of the Royal Statistical Society B* **60**, 559-578, 1998.

[28] MILLS, Joanna E. The analysis of longitudinal binary data. These. Canada: Dalhousie University, 2000.

[29] MOLENBERGHS, Geert and VERBEKE, Geert. Models for discrete longitudinal data. New York: Springer, 2005.

[30] NATIS, L. Modelos lineares hierárquicos, Dissertação. IME-USP. São Paulo, 2002.

[31] NELDER, J. A. and MEAD, R. A simplex method for function minimization. *The Computer Journal*, **7**, 941-946, 1965.

[32] NELDER, J. A. and WEDDERBURN,R. W. M. Generalized linear models. *Journal of the Royal Statistical Society A* **135**, p.370-384, 1972.

[33] NOBRE, Juvêncio Santos. Métodos de diagnóstico para modelos lineares mistos. Dissertação. IME-USP. São Paulo, 2004.

[34] PAN, W. Akaike's information criterion in generalized estimating equations. *Biometrics* **57**, p. 120-125, 2001.

[35] PAULA, Gilberto A. Modelos de regressão com apoio computacional. São Paulo: IME-USP, 2004.

[36] PATTERSON, H.D. & THOMPSON, R. Recovery of interblock information when block sizes are unequal. *Biometrika* **58**, p. 545-554, 1971.

[37] PREISSER, J. S. and QAQISH, B. F. Deletion diagnostics for generalised estimating equations. *Biometrika* **83**, p. 551-562, 1996.

[38] ROCHA, Francisco Marcelo Monteiro da. Seleção de Estruturas de Covariância para Dados com Medidas Repetidas. Dissertação. IME-USP. São Paulo, 2004.

[39] SAAVEDRA, Pedro A. Tores. Percentile curves in binary longitudinal data. These. University of Puerto Rico, 2006.

[40] SINGER, Julio M.; NOBRE, Juvêncio S.; ROCHA, Francisco Marcelo M. Análise de dados longitudinais. São Paulo: IME-USP, versão preliminar, 2007.

[41] SNIJDERS T., & BOSKER R. Multilevel analysis: An introduction to basic and advanced multilevel modeling. London: Sage, 1999.

[42] SOUZA, Édila Cristina. Análise de influência local no modelo de regressão logística. Dissertação. Piracicaba, São Paulo: ESALQ-USP, 2006.

[43] TAN, M., QU, Y. and KUTNER, M. H. Model diagnostics for marginal regression analysis of correlated binary data. *Commun. Statist. - Simula.* **26**, p. 539-558, 1997.

[44] VENEZUELA, Maria K. Modelos lineares generalizados para análise de dados com medidas repetidas. Dissertação. IME-USP. São Paulo, 2003.

[45] VERBEKE, Geert and MOLENBERGHS, Geert. Linear mixed models for longitudinal data. New Yourk: Springer, 2000.

[46] WARE, J. H. Linear models for the analysis of longitudinal studies. The American Statistician **39**, p. 95-101, 1985.

[47] WEDDERBURN,R. W. M. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, part 3, p. 439, 1974.