Diferentes estratégias para modelagem de respostas politômicas ordinais em estudos longitudinais

Nívea Bispo da Silva

Dissertação apresentada ao
Programa de Pós-Graduação em Estatística da
Universidade Federal de Minas Gerais
para obtenção do título de Mestre em Estatística

Orientador: Prof. Dr. Enrico Antônio Colosimo (UFMG)
Co-orientadora: Profa. Dra. Leila Denise Alves Ferreira Amorim (UFBA)

Belo Horizonte
27 de fevereiro de 2013

Diferentes estratégias para modelagem de respostas politômicas ordinais em estudos longitudinais

Esta versão da dissertação contém as correções e alterações
sugeridas pela banca durante a defesa da versão original
do trabalho, realizada em 22 de fevereiro de 2013.

Comissão Julgadora:

- Prof. Dr. Enrico Antonio Colosimo (orientador) - UFMG
- Profa. Dra. Glaura da Conceição Franco - UFMG
- Prof. Dr. Juvêncio Santos Nobre - UFC

Agradecimentos

A jornada foi longa, cansativa, mas ao olhar para trás diria citando o sábio Fernando Pessoa que: *"Tudo vale a pena quando a alma não é pequena"*.

Dedico esse trabalho às pessoas que são o meu porto seguro: Meus pais José e Petronilia, que me ensinaram a ser perseverante e determinada e a jamais desistir das coisas que acredito. E minhas irmãs, Luciene e Marília, que sempre tinham uma palavra de incentivo e fé nos momentos em que o medo e a insegurança tentavam se aproximar de mim. Agradeço-lhes por todo amor, incentivo, orações e por sempre apoiarem as minhas decisões. Saber que posso contar com o carinho e compreensão de vocês me faz mais forte e segura de que alcançarei o que almejo, caminhando sempre com humildade e sem medo dos obstáculos que precisarei ultrapassar.

Agradeço a Deus por ser presença constante em minha vida, conduzindo-me sempre e não me deixando desistir.

Ao Enrico Colosimo agradeço imensamente pela orientação, momentos de aprendizado, e por acreditar que juntos poderíamos finalizar este trabalho. Agradeço também à profa Leila Amorim pela co-orientação e oportunidade de continuar compartilhando conhecimento. Espero que a parceria com vocês ainda renda muitos frutos!

Aos professores da Pós pelo conhecimento que adquiri durante as disciplinas, e também aos professores da UFBA pela minha formação acadêmica.

Aos amigos que conquistei (Claudia T., Marília Souza, Mariese Alves, Sheila Regina, Silvia Lemos, Marinalva Souza, Daniele Trindade, Suzane Carvalho, Carlos Trucios, Kelly Cadena, Marta Macufa, e muitos outros que agora me fugiram da mente, mas que não deixam de ser importantes) ao longo desses anos, e com quem compartilhei momentos únicos e inesquecíveis. Poder contar com o carinho e amizade de vocês foi e sempre será muito importante para mim.

Aos amigos que ganhei em BH, e com quem compartilhei bons momentos. Agradeço-lhes pela amizade, conselhos e pelo respeito mútuo que existe entre nós. Brigadão de coração a Talita Costa, Silvana Shneider, Izabella Alves, Mariana Araújo, e aos meus veteranos Paulo Cerqueira e Gabriel Caldas que foram anjos da guarda e me acolheram em seu grupo de braços abertos. Agradeço também ao André por toda paciência e ajuda. As reuniões

que tivemos durante o mestrado me ajudaram muito! Brigadão também às mineirinhas com quem moro (Grazi e Dani), que me receberam de braços abertos, e com quem dei boas risadas. Conviver com todos vocês foi muito bom!

Aos membros da banca, profesores Juvêncio Nobre (Universidade Federal do Ceará) e Glaura Franco, pelas ricas contribuições dadas ao texto, e que serviram para enriquecê-lo ainda mais.
Ao CNPq pelo apoio financeiro durante a execução desse trabalho, e ao Departamento de Estatística da UFMG pela estrutura e suporte oferecidos.

Enfim, obrigada a todos que direta ou indiretamente me ajudaram a tornar possível essa conquista!

"A mente que se abre a novas ideias jamais voltará ao seu tamanho original"
(Albert Einstein)

Resumo

A modelagem de respostas politômicas, em especial as ordinais, tem sido alvo de crescente interesse nos últimos anos, e vem ganhando espaço em pesquisas sobre qualidade de vida, indicadores de condição de saúde, avaliação da proficiência dos alunos em determinadas disciplinas, dentre outras. A sua utilização vai desde os estudos transversais, onde se assume independência entre as observações, até os estudos longitudinais, em que mais de uma resposta do mesmo indivíduo é observada ao longo do tempo. Existem na literatura várias metodologias propostas para modelar dados desta natureza em estudos transversais, sendo a mais usual na prática a que utiliza o modelo de logitos cumulativos, também conhecido como modelo de odds proporcionais devido ao pressuposto do modelo que assume proporcionalidade nas *odds*, ou seja, o modelo assume que há um crescimento aproximadamente linear das razões de chances para os coeficientes da regressão. Em muitas situações práticas o referido pressuposto pode não ser verificado, tornando desaconselhável a utilização do modelo. Há, contudo, outro modelo que generaliza o de *odds* proporcionais, conhecido por *odds* proporcionais parciais, que permite a não proporcionalidade para um subconjunto de covariáveis que violaram o pressuposto de proporcionalidade. Em estudos longitudinais, os modelos usuais - modelos marginais, modelos lineares generalizados mistos e modelos de transição - para análise de dados correlacionados também podem ser utilizados para modelar respostas politômicas. Nesse trabalho a modelagem de respostas politômicas ordinais em estudos longitudinais é discutida sob a ótica dos modelos marginais, modelos lineares generalizados mistos e de transição. A especificação e interpretação dos modelos é ilustrada e discutida através da análise de dois conjuntos de dados reais.

Palavras-chave: *respostas ordinais, modelo de odds proporcionais, modelo de odds proporcionais parciais, modelos marginais, modelos lineares generalizados mistos, modelos de transição.*

Abstract

The modeling of polytomous responses, especially ordinal, has been the subject of increasing interest in recent years, and has been gaining ground in research on quality of life, health status indicators, assessment of student proficiency, among others. Its use goes from cross-sectional studies, which assume independence among the observations, to longitudinal studies, where more than one response from the same individual is observed over time. There are several methods proposed in the literature for modeling polytomous responses in transversal studies, being the most commonly used the cumulative logits model, also known as proportional odds model due to the assumption of proportionality in odds, i.e., the model assumes that there is an approximately linear increase in odds ratios for the regression coefficients. In many practical situations this assumption is violated. There is, however, another model that generalizes the proportional odds, known as partial proportional odds, which allows no odds proportionality to a subset of covariates that violated that assumption. In longitudinal studies, the conventional models - marginal models, generalized linear mixed models and transition models - for analysis of correlated data can also be used to model polytomous responses. In this work modeling of ordinal polytomous responses in longitudinal studies is discussed from the perspective of marginal, generalized linear mixed models and transition models. The specification and interpretation of the models is illustrated and discussed by analyzing two real data sets.

Keywords: *ordinal responses, proportional odds model, partial proportional odds model, marginal models, GLMM, transition models.*

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Nos últimos anos muitas metodologias têm sido propostas para modelar respostas politômicas (nominais ou ordinais), e o interesse nesse tipo de modelagem tem sido crescente tanto em estudos transversais quanto em longitudinais.

Na área epidemiológica, por exemplo, há interesse frequente em estimar o risco de eventos adversos, e os pesquisadores geralmente escolhem classificar a resposta de interesse em 2 ou mais categorias, a fim de estimar o risco relativo ou a *odds ratio*, a depender do delineamento do estudo (Abreu et al, 2009; Ananth et al., 1997). Em ensaios clínicos respostas politômicas numa escala ordinal são muitas vezes utilizadas para quantificar os sintomas ou condição do paciente, bem como avaliar a eficácia de procedimentos pós-operatórios (Parsons et al, 2009). O uso de respostas politômicas, em particular as ordinais, tem ganhado bastante espaço em estudos sobre qualidade de vida, indicadores de condição de saúde, e até mesmo sobre a gravidade de certa doença.

Em estudos transversais a modelagem de respostas politômicas é feita ajustando-se um modelo logito multinomial, também conhecido como logito generalizado, caso as categorias da resposta sejam nominais. Para respostas politômicas ordinais existem na literatura diversos modelos propostos (Ananth et al., 1997, Agresti, 2002). Dentre eles, o mais utilizado na prática é o modelo logito cumulativo, também conhecido como modelo de *odds* proporcionais (McCullagh, 1980). Há também o modelo de *odds* proporcionais parciais (Peterson e Harrell, 1990) e o modelo de categorias adjacentes (Ananth et al., 1997, Agresti, 2002).

Quando se tem uma única resposta para cada indivíduo, os modelos anteriormente citados podem ser usados para avaliar a influência das variáveis preditoras sobre a resposta de interesse. Contudo, várias respostas do mesmo indíviduo podem ser obtidas ao longo de certo período de tempo, caracterizando assim um estudo longitudinal, e técnicas específicas de modelagem precisam ser usadas em tal caso, de forma a tentar captar a heterogeneidade entre indivíduos decorrente das medidas repetidas.

Assim como na modelagem de respostas binárias e de contagem, os modelos con-

vencionais (modelos marginais, modelos lineares generalizados mistos, modelos de transição) para modelar dados longitudinais podem ser utilizados na modelagem de respostas politômicas. Heagerty e Zeger (1996) apresentaram uma proposta para construção de mode-los para medidas ordinais agrupadas utilizando modelos marginais. Hedeker e Gibbons (1994; 2006) apresentaram um modelo com representação multinível, que acomoda múltiplos efeitos aleatórios para analisar respostas ordinais longitudinais, onde as variáveis preditoras podem ser incluídas em ambos os níveis do modelo de forma a tentar explicar a variação intra e entre indivíduos. Outra classe de modelos que também pode ser usada para modelar respostas politômicas são os modelos de transição, que avaliam a probabilidade de transição da resposta de uma categoria para outra, e consideram o efeito prévio da resposta sob a resposta atual (Diggle et al, 2002). Dentro dessa classe de modelos é possível caracterizar a correlação entre as observações repetidas incorporando no mesmo uma estrutura para a média marginal, sendo tal caracterização feita através dos chamados modelos de transição marginalizados, inicialmente propostos para respostas binárias longitudinais (Heagerty et al., 2000; Heagerty, 2002), mas sendo recentemente estendidos de forma a acomodar dados longitudinais ordinais (Lee e Daniels, 2007).

Nesse trabalho objetiva-se sistematizar a literatura sobre modelos marginais, lineares generalizados mistos e de transição para modelar respostas politômicas ordinais em estudos longitudinais. A especificação e interpretação dos modelos será ilustrada e discutida através da análise de dois conjuntos de dados reais. Na primeira aplicação usaremos os dados provenientes de um estudo epidemiológico desenvolvido com 664 crianças de até 5 anos, residentes em dois municípios do médio vale do Jequitinhonha em Minas Gerais. Nesse estudo o objetivo principal foi avaliar o impacto dos sistemas de captação de água da chuva na saúde da criança. A segunda aplicação refere-se a um estudo com 49 parturientes que foram acompanhadas até a hora do parto. O estudo foi conduzido pela Faculdade de Medicina da UFMG e pelo hospital municipal Odilon Berhens, e tinha como objetivo a compararação de duas técnicas de analgesia para a dor no trabalho de parto.

O presente trabalho está organizado da seguinte forma: apresentamos no Capítulo 2 a modelagem transversal de respostas politômicas, com definição do modelo para resposta nominal (seção 2.1), e dois modelos para respostas ordinais (seção 2.2). No Capítulo 3 é feita uma breve revisão sobre modelagem longitudinal, onde apresentamos três classes de modelos de regressão para dados longitudinais. Na subseção 3.1.1 trataremos dos modelos marginais. Na subseção 3.1.2 discutimos a teoria dos modelos lineares generalizados mistos. E, na subseção 3.1.3 apresentamos os modelos de transição. No Capítulo 4 os modelos para respostas ordinais longitudinais são apresentados sob a ótica das abordagens marginal (seção 4.1), mista (4.2) e de transição (subseção 4.3). No Capítulo 5 apresentamos a aplicação feita com dois conjuntos de dados reais. Por fim, o Capítulo 6 traz as considerações finais para este trabalho.

Capítulo 2

Modelagem Transversal

A modelagem de respostas não-normais é bastante comum na prática, e um exemplo clássico do seu uso começou com os modelos de regressão probito para modelagem de respostas binárias em meados do século XIX (McCulloch et al., 2001). Nelder e Wedderburn. (1972), buscando estender a classe de modelos lineares, propuseram uma classe mais ampla de modelos que contempla vários tipos de respostas não-normais, baseadas na família exponencial de distribuições, que possui propriedades interessantes para estimação, testes de hipóteses e outros aspectos de inferência. Essa classe de modelos conhecida como modelos lineares generalizados (GLM, em inglês) é definida pelas distribuições de probabilidade, membros da família exponencial, por um conjunto de variáveis preditoras, que descrevem a estrutura linear do modelo, e por uma função de ligação entre a média da variável resposta e sua estrutura linear (Cordeiro et al., 2006). Várias distribuições importantes de probabilidade (Normal, Gama, Poisson, Binomial, dentre outras) são membros da família exponencial, e alguns modelos anteriores à classe proposta por Nelder e Wedderburn são casos especiais dos GLM's: modelo linear normal, modelos log-lineares aplicados na análise de tabelas de contigência, modelo probito, dentre outros.

Em linhas gerais, a estrutura de um GLM é formada por 3 componentes:

i) Um componente aleatório, composto pela variável aleatória $\mathbf{Y} = (Y_1^T, \ldots, Y_n^T)^T$, com n observações independentes e identicamente distribuídas, com média μ, onde supõe-se que cada componente de $\mathbf{Y}$ segue uma distribuição da família exponencial, definida por:

$$f_{\mathbf{Y}}(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \mathbb{I}_A(y)$$

em que θ é o parâmetro que caracteriza a distribuição $f_Y(\cdot)$, a, b e c são funções conhecidas, ϕ é um parâmetro de dispersão, e $\mathbb{I}_A(y)$ é o suporte da resposta, que não pode depender de θ.

ii) Um componente sistemático, composto por variáveis preditoras $\mathbf{X} = (X_1^T, \ldots, X_p^T)^T$, que produzem o preditor linear $\eta = \mathbf{X}\beta$, sendo $\beta = (\beta_1^T, \ldots, \beta_p^T)^T$ um vetor $p \times 1$ de parâmetros;

iii) Uma função de ligação $g(\mu) = \eta$, que relaciona os dois componentes anteriores.

Assim, a função de verossimilhança para um GLM, pode ser expressa como:

$$L_{\mathbf{Y}}(\beta, y) = \prod_{i=1}^{n} \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}.$$

As estimativas de máxima verossimilhança para o vetor de parâmetros β são obtidas através do algoritmo iterativo de Newton-Raphson (Cordeiro et al., 2006).

Se a resposta de interesse for binária, por exemplo, o modelo logístico é frequentemente usado para modelar a relação entre a variável resposta e um conjunto de variáveis preditoras. Contudo, a resposta de interesse pode ser categórica e assumir mais que duas categorias, sendo então necessário estender o modelo logístico para acomodar respostas politômicas. Nas situações em que a resposta assume mais que duas categorias, as mesmas podem ou não ter uma ordenação natural. Se as categorias da resposta não forem naturalmente ordenadas tem-se um resposta politômica nominal, e seu uso é comum em estudos sobre a situação de emprego (empregado, desempregado, aposentado/estudante) em processos de imigração, estudos educacionais relacionados com a preferência de aprendizado do aluno (individual, grupo, sala de aula), dentre outros. Caso as categorias da resposta sejam ordenadas, então a resposta será politômica ordinal, e seu uso geralmente ocorre em estudos sobre qualidade de vida (raramente sente-se cansado, ocasionalmente sente-se cansado, cansaço frequente), indicadores de condição de saúde (ruim, regular, boa, ótima), e até mesmo sobre a gravidade de certa doença (baixo risco, risco intermediário, estágio avançado) (Pettit et al., 2002; Hedeker, 2003).

Quando as categorias da resposta forem nominais tem-se na literatura o modelo multinomial nominal, também conhecido como logito generalizado. Sendo as categorias da resposta ordenadas, existe mais de um modelo que pode ser usado para modelar a resposta ordinal. O modelo mais conhecido é o logístico ordinal (ou logito cumulativo), inicialmente proposto por Walker e Duncan (1967), e anos mais tarde chamado de modelo de *odds* proporcionais por McCullagh (1980). Outros três modelos também comumente usados na modelagem de respostas ordinais são: modelo de *odds* proporcionais parciais, modelo de categorias adjacentes, e modelo de razões contínuas (Agresti, 2002; Hosmer e Lemeshow, 2000; Ananth et al., 1997).

Nas seções a seguir os principais modelos para respostas politômicas são apresentados.

2.1 Modelo para resposta politômica nominal

Quando a resposta de interesse é politômica nominal, o modelo mais usual é o logístico politômico, também conhecido como logito generalizado. Esse modelo é uma extensão do modelo logístico para respostas binárias, e consiste de uma combinação de vários modelos logito, estimados simultaneamente. Assim, seja $\mathbf{Y}$ uma variável categórica com K categorias, $k = 1, 2, \ldots, K$, em que:

$$Y_{ik} = \begin{cases} 1, & \text{se o indivíduo } i \text{ está na categoria } k, \quad i = 1, \ldots, n \\ 0, & \text{c.c.} \end{cases}$$

Desta forma, existem $\binom{K}{2}$ pares de categorias, bem como $\binom{K}{2}$ preditores lineares.

Denotemos por π_k a probabilidade do indivíduo estar na categoria k. Ou seja, $\pi_k = \mathbb{P}(Y_{ik} = 1|\mathbf{X})$.

Usando a categoria K como referência, precisamos de apenas $K-1$ comparações para esta categoria. Assim, o logito para a k-ésima comparação é dado por:

$$\log\left[\frac{\pi_k}{\pi_K}\right] = \gamma_k + \mathbf{X}\beta_k, \quad k = 1, \ldots, K-1. \tag{2.1}$$

onde γ_k é o intercepto do modelo.

Em geral, para compararmos as categorias j e l, por exemplo, temos:

$$\log\left[\frac{\pi_j}{\pi_l}\right] = (\gamma_j - \gamma_l) + X_i(\beta_j - \beta_l). \tag{2.2}$$

E, dadas as $K-1$ expressões do logito em (2.1), é possível obter $K-1$ expressões diretamente para a probabilidade das categorias em termos de $K-1$ preditores lineares. Assim, a expressão resultante é do tipo:

$$\pi_k = \frac{e^{\gamma_k + \mathbf{X}\beta_k}}{1 + \sum_{k=1}^{K-1} e^{\gamma_k + \mathbf{X}\beta_k}}, \quad k = 1, 2, \ldots, K-1. \tag{2.3}$$

A categoria de referência pode ser expressa na forma:

$$\pi_K = 1 - (\pi_1 + \ldots + \pi_{K-1}) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\gamma_k + \mathbf{X}\beta_k}}.$$

A partir de tais definições é possível escrever a função de verossimilhança para

o i-ésimo indivíduo na amostra com base na probabilidade fornecida pela distribuição multinomial:

$$\mathbb{P}(Y_{i1} = y_{i1}, \ldots, Y_{iK} = y_{iK}) = (\pi_1)^{y_{i1}} \cdots (\pi_K)^{y_{iK}}.$$

Assim, a função de verossimilhança é descrita por:

$$L(\beta) = \prod_{i=1}^{n} \left[\prod_{k=1}^{K} \pi_k^{y_{ik}} \right]$$

$$= \prod_{i=1}^{n} \left[\prod_{k=1}^{K-1} \pi_k^{y_{ik}} \left(1 - \sum_{k=1}^{K-1} \pi_k \right)^{1 - \sum_{k=1}^{K-1} y_{ik}} \right]. \tag{2.4}$$

A função de log-verossimilhança, pode então, ser expressa na forma:

$$l = \sum_{i=1}^{n} \left[\sum_{k=1}^{K-1} y_{ik} \log \pi_k + \left(1 - \sum_{k=1}^{K-1} y_{ik} \right) \log \left(1 - \sum_{k=1}^{K-1} \pi_k \right) \right]$$

$$= \sum_{i=1}^{n} \left[\sum_{k=1}^{K-1} y_{ik} (\gamma_k + X_i^T \beta_k^T) - \log \left(1 + \sum_{k=1}^{K-1} e^{\gamma_k + X_i^T \beta_k^T} \right) \right].$$

Os parâmetros do modelo são estimados através do método de máxima verossimilhança, usando o algoritmo iterativo de Newton-Raphson. A interpretação dos parâmetros no modelo para respostas nominais é análoga à regressão logística para respostas binárias, sendo uma das categorias da resposta tomada como referência.

Consideremos, para melhor entendimento, o exemplo a seguir. Pesquisadores em educação estavam interessados em avaliar o desempenho dos alunos em matemática, e para tal entrevistaram os alunos acerca de qual programa de aprendizagem preferiam (1: individual, 2: grupo, 3: sala de aula), e se a preferência estaria associada com a escola (1,2 ou 3) e o período escolar (0: integral, 1: padrão). A Tabela 2.1 mostra a preferência dos alunos (resposta de interesse) por escola e período (Stokes et al., 2000).

Por se tratar de uma resposta categórica, cujas categorias não possuem uma ordenação natural, utilizamos o modelo de logitos generalizados. Para esse exemplo a forma funcional do modelo a ser considerada é

$$\log\left[\frac{\pi_{hjk}}{\pi_{hjK}}\right] = \gamma_k + X_{h_1}\beta_{1k} + X_{h_2}\beta_{2k} + X_j\beta_{3k},$$

em que $k = 1, 2, 3$; $h = 1, 2, 3$; $j = 1, 2$; X_{h1} denota o indicador para a escola 2; X_{h2} denota o indicador para a escola 3; e X_j denota o indicador para o período. Para esse modelo são definidos dois logitos generalizados, considerando a categoria 3 da resposta como sendo a referência. Ou seja:

$$\text{logito}_1 = \log\left[\frac{\pi_{hj1}}{\pi_{hj3}}\right] \quad \text{e} \quad \text{logito}_2 = \log\left[\frac{\pi_{hj2}}{\pi_{hj3}}\right]$$

em que π_{hjk} denota a probabilidade do aluno da escola h e período escolar j, preferir o programa de aprendizado k.

Tabela 2.1: Preferência dos alunos por escola e período

		Preferência de aprendizado			
Escola	Período	Individual	Grupo	Sala Aula	Total
1	Padrão	10	17	26	53
1	Integral	5	12	50	67
2	Padrão	21	17	26	64
2	Integral	16	12	36	64
3	Padrão	15	15	16	46
3	Integral	12	12	20	44

Para o modelo ajustado, os resultados são apresentados na Tabela 2.2.

Tabela 2.2: Estimativas dos parâmetros para o modelo logito generalizado

	logito(1|3)		logito(2|3)	
covariável	estimativa	ep	estimativa	ep
intercepto	-0,78	0,15	-0,66	0,14
Escola				
2	-0,79*	0,22	-0,28	0,19
3	0,28	0,19	-0,09	0,19
Período padrão	0,37*	0,14	0,37*	0,14

*p-valor$<$0.05; ep: erro-padrão

É possível observar que a escola 1 difere da 2 no logito (1|3), que compara o aprendizado individual ao em sala de aula. Já o período escolar apresenta efeitos significativos similares em ambos os logitos. Se interpretarmos os resultados em termos das razões de chances, observa-se que a chance dos alunos cujo período escolar é o padrão preferirem o aprendizado individual ao invés da sala de aula é 1,45 ($e^{0,37}$) vezes a chance dos alunos em escolas com período integral (maiores detalhes em Stokes et al., 2000).

2.2 Modelos para resposta politômica ordinal

Se a resposta de interesse for politômica ordinal, existem na literatura alguns modelos que podem ser utilizados para ajuste deste tipo de dado. A seguir são apresentados os modelos de *odds* proporcionais e *odds* proporcionais parciais.

2.2.1 Modelo de *odds* proporcionais

Este modelo foi inicialmente proposto por Walker e Duncan em 1967, sendo mais tarde chamado modelo de *odds* proporcionais por McCullagh (1980). Assim, se a variável resposta é ordinal e possui k categorias, então há $K-1$ dicotomizações da resposta, ou seja:

$$Y_{ik} = \begin{cases} 1, & \text{se} \quad y_{ik} \leq k, \quad i = 1, \ldots, n \\ 0, & \text{se} \quad y_{ik} > k \end{cases}$$

Nesse modelo uma forma de representar as probabilidades acumuladas é dada pela expressão:

$$\mathbb{P}(Y_{ik} = 1|\mathbf{X}) = \mathbb{P}(y_{ik} \leq k) = \pi_1 + \pi_2 + \ldots + \pi_k, \quad k = 1, 2, \ldots, K. \tag{2.5}$$

De forma geral, os logitos cumulativos são então representados por:

$$\log\left[\frac{\mathbb{P}(y_{ik} \leq k)}{1 - \mathbb{P}(y_{ik} \leq k)}\right] = \log\left[\frac{\pi_1 + \ldots + \pi_k}{\pi_{k+1} + \ldots + \pi_K}\right]$$

$$= \gamma_k + \mathbf{X}\beta. \tag{2.6}$$

De maneira equivalente, as probabilidades acumuladas descritas em (2.5) podem ser expressas por:

$$\mathbb{P}(y_{ik} \leq k) = \frac{e^{\gamma_k + \mathbf{X}\beta}}{1 + e^{\gamma_k + \mathbf{X}\beta}}.$$

Nota-se na expressão (2.6) que cada logito cumulativo tem seu próprio intercepto, e diferentemente do modelo para respostas nominais (descrito em (2.1)), os β's são os mesmos para cada logito, ou seja, $\beta_k = \beta$ para todo $k = 1, \ldots, K$. Isto ocorre devido ao pressuposto do modelo de que todas as razões de chance são idênticas entre os $K-1$ pontos de corte, ou seja, esse modelo assume que todas as observações possuem uma variância comum, implicando, assim, que há um crescimento aproximadamente linear das razões de

chances. McCullagh (1980) chamou esse pressuposto de *odds* proporcionais. Segundo ele, os parâmetros γ_k são, em geral, de pouco interesse na intepretação do modelo, e usualmente são referidos como "pontos de corte". Já o parâmetro β descreve o incremento no log da *odds* (ou chance) para qualquer categoria da resposta.

A verificação do pressuposto de que $\beta_k = \beta$ pode ser feita global ou individualmente através do teste da razão de verossimilhança, ou ainda através do teste baseado na estatística escore, cuja distribuição assintótica é qui-quadrado com $p \times (K-1)$ graus de liberdade, sendo p o número de variáveis preditoras usadas no modelo, e K o número de categorias da resposta ordinal. O teste global apenas avalia se o modelo como um todo violou o pressuposto de proporcionalidade das *odds*. A partir dele não é possível identificar qual ou quais variáveis preditoras consideradas no ajuste violaram o pressuposto. Para avaliar se cada covariável individualmente violou o pressuposto, realiza-se também um teste da razão de verossimilhança, ou ainda usa-se uma estatística de teste escore, cuja distribuição será qui-quadrado com $K-1$ graus de liberdade.

Caso o pressuposto do modelo não possa ser verificado para todas as variáveis preditoras ou um subconjunto delas, é possível relaxar o modelo de *odds* proporcionais ajustando-se o modelo conhecido como *odds* proporcionais parciais (Peterson e Harrel, 1990), a ser apresentado na subseção 2.2.2.

Assim, em um estudo cuja resposta possui, por exemplo, 3 categorias ordenadas, teremos 2 logitos cumulativos baseados nas probabilidades acumuladas, ou seja:

$$\begin{aligned} \text{logito}_1 = \log\left[\frac{\pi_1}{\pi_2+\pi_3}\right] \quad \text{e} \quad \text{logito}_2 = \log\left[\frac{\pi_1+\pi_2}{\pi_3}\right] \\ = \gamma_1 + \mathbf{X}\beta \qquad\qquad = \gamma_2 + \mathbf{X}\beta \end{aligned}$$

Os logitos 1 e 2, respectivamente, representam, em geral, o log da *odds* para a categoria mais favorável em relação às demais categorias, e o log da *odds* para as 2 categorias mais favoráveis em relação à última categoria.

Para esse modelo a função de verossimilhança é dada por:

$$L(\beta) = \prod_{i=1}^{n}\left[\prod_{k=1}^{K}\pi_k^{y_{ik}}\right] = \prod_{i=1}^{n}\left[\prod_{k=1}^{K}\left(\mathbb{P}(y_{ik}\leq k) - \mathbb{P}(y_{ik}\leq k-1)\right)^{y_{ik}}\right]$$

$$= \prod_{i=1}^{n}\left[\prod_{k=1}^{K}\left(\frac{e^{\gamma_k + X_i^T\beta^T}}{1+e^{\gamma_k+X_i^T\beta^T}} - \frac{e^{\gamma_{k-1}+X_i^T\beta^T}}{1+e^{\gamma_{k-1}+X_i^T\beta^T}}\right)^{y_{ik}}\right]$$

As estimativas do modelo são obtidas através do método de máxima verossimilhança, usando o algoritmo escore de Fisher, por exemplo.

Para melhor compreensão do modelo de *odds* proporcionais consideraremos como exemplo os dados de um estudo epidemiológico, do tipo coorte prospectiva, desenvolvido com 664 crianças de até 5 anos, selecionadas e acompanhadas pelo período de um ano (2009 a 2010). O objetivo principal do estudo foi avaliar o impacto da implantação dos sistemas de captação de água da chuva na saúde das crianças de famílias rurais, residentes em 2 municípios do Médio Vale do Jequitinhonha em Minas Gerais (maiores detalhes em Fonseca, 2012).

Para a aplicação consideramos apenas a etapa inicial do estudo. As variáveis preditoras usadas no ajuste do modelo são: sexo (0: Feminino; 1: Masculino), grupo (1: com cisterna; 0: sem cisterna), frequência de banho da criança (1: >1 vez ao dia; 0: uma vez ao dia) e idade da criança em meses. A resposta de interesse é a carga parasitária da criança (1: poliinfectada, 2: monoinfectada, 3: não infectada), avaliada via exame de fezes. Para a referida resposta teremos 2 logitos cumulativos representados como segue:

$$\log\left[\frac{\pi_1}{\pi_2+\pi_3}\right] = \log\left[\frac{\mathbb{P}(y_{ik}\leq 1)}{\mathbb{P}(y_{ik}>1)}\right] = \gamma_1 + \beta_1\text{grupo} + \beta_2\text{sexo} + \beta_3\text{freq. banho} + \beta_4\text{idade}$$

e

$$\log\left[\frac{\pi_1+\pi_2}{\pi_3}\right] = \log\left[\frac{\mathbb{P}(y_{ik}\leq 2)}{\mathbb{P}(y_{ik}>2)}\right] = \gamma_2 + \beta_1\text{grupo} + \beta_2\text{sexo} + \beta_3\text{freq. banho} + \beta_4\text{idade}.$$

O primeiro logito representa o log da chance de uma criança ser poliinfectada em comparação às crianças mono ou não infectadas. Já o segundo representa o log da chance de uma criança ser poli ou monoinfectada em comparação a uma criança sem infecção.

A análise descritiva mostrou que 53% das crianças analisadas são do sexo masculino, 84% delas tomam mais que 1 banho por dia, e possuem, em média, 26,4 meses de vida. Em relação à variável grupo, metade das crianças possuiam cisterna em casa no início do estudo, e apenas 3% das crianças eram poliinfectadas. Já a proporção de crianças monoinfectadas era de 10%.

O modelo ajustado considerando a definição (2.6) forneceu os resultados apresentados na Tabela 2.3.

Tabela 2.3: Estimativas dos parâmetros para o modelo logito cumulativo

variável	estimativa	ep	p-valor
γ_1	-3,493	0,328	-
γ_2	-1,923	0,292	-
grupo (com cisterna)	0,159	0,196	0,417
sexo (M)	-0,233	0,195	0,234
freq.banho (>1x ao dia)	-0,473	0,259	0,102
idade.criança	0,032	0,001	<0,001

A partir dos resultados apresentados na Tabela 2.3 é possível encontrar as probabilidades preditas para a carga parasitária das crianças em função das variáveis preditoras analisadas, conforme definido no Quadro 1.

Quadro 1: Probabilidades preditas para o modelo de *odds* proporcionais

poliinfecção	$\frac{e^{\gamma_1+\beta_1\text{grupo}+\beta_2\text{sexo}+\beta_3\text{freq. banho}+\beta_4\text{idade}}}{1+e^{\gamma_1+\beta_1\text{grupo}+\beta_2\text{sexo}+\beta_3\text{freq. banho}+\beta_4\text{idade}}}$
poli ou monoinfecção	$\frac{e^{\gamma_2+\beta_1\text{grupo}+\beta_2\text{sexo}+\beta_3\text{freq. banho}+\beta_4\text{idade}}}{1+e^{\gamma_2+\beta_1\text{grupo}+\beta_2\text{sexo}+\beta_3\text{freq. banho}+\beta_4\text{idade}}}$
sem infecção	$\frac{1}{1+e^{\gamma_2+\beta_1\text{grupo}+\beta_2\text{sexo}+\beta_3\text{freq. banho}+\beta_4\text{idade}}}$

Para o modelo ajustado testou-se o pressuposto de proporcionalidade das *odds*. Através do teste da razão de verossimilhança verificou-se que o referido pressuposto foi violado (p-valor=0,005). Para identificar qual ou quais variáveis preditoras consideradas no ajuste violaram o pressuposto, foram realizados testes da razão de verossimilhança, e verificou-se que as variáveis grupo e idade da criança em meses não possuem *odds* proporcionais (p-valor=0,001 e 0,044, respectivamente). Em decorrência da violação do pressuposto do modelo de *odds* proporcionais, o mais adequado é buscar outras alternativas para modelar a resposta ordinal. Uma delas seria utilizar o modelo de *odds* proporcionais parciais apresentado a seguir.

2.2.2 Modelo de *odds* proporcionais parciais

A motivação primária para o desenvolvimento desse modelo, por Peterson e Harrel (1990), foi relaxar o forte pressuposto do modelo de *odds* proporcionais. No modelo proposto por McCullagh (1980) se pelo menos uma covariável considerada no ajuste viola o pressusposto, então a qualidade do modelo ajustado fica comprometida. Assim, Peterson e Harrel propuseram relaxar o pressuposto de proporcionalidade das *odds* para um subconjunto de variáveis preditoras usadas no ajuste do modelo. O modelo por eles proposto pode ser definido como:

$$C_{ik} = \mathbb{P}(y_{ik} \leq k) = \frac{e^{\gamma_k + \mathbf{X}\beta + \widetilde{\mathbf{X}}\varrho_k}}{1 + e^{\gamma_k + \mathbf{X}\beta + \widetilde{\mathbf{X}}\varrho_k}} \tag{2.7}$$

em que $\mathbf{X}$ é uma matrix $n \times p$ de variáveis preditoras; $\widetilde{\mathbf{X}}$ é uma matriz $n \times q$ $(q \leq p)$ contendo um subconjunto de variáveis preditoras da matriz X para os quais o pressuposto de proporcionalidade das *odds* não pôde ser assumido, ou que ainda será testado; $\varrho_k = (\varrho_1^T, \ldots, \varrho_{K-1}^T)^T$ é um vetor $q \times 1$ de parâmetros.

De maneira equivalente, o modelo (2.7) pode ser descrito como:

$$\begin{aligned} \log\left[\frac{\mathbb{P}(y_{ik} \leq k)}{1 - \mathbb{P}(y_{ik} \leq k)}\right] &= \log\left[\frac{\pi_1 + \ldots + \pi_k}{\pi_{k+1} + \ldots + \pi_K}\right] \\ &= \gamma_k + \mathbf{X}\beta + \widetilde{\mathbf{X}}\varrho_k. \end{aligned} \tag{2.8}$$

Se $\varrho_k = 0$ para todo $k = 1, \ldots, K$, então o modelo (2.8) se reduz ao proposto por McCullagh (1980) (expressão (2.6)). Para testar se $\varrho_k = 0$, Peterson e Harrel propuseram um teste baseado na estatística escore, que, sob a hipótese nula (H_0), assume que as *odds* são proporcionais. Ainda segundo esses autores, somente se H_0 for rejeitada é que há a necessidade de se ajustar o modelo de *odds* proporcionais parciais. A distribuição da estatística proposta é assintoticamente qui-quadrado com $q(K-1)$ graus de liberdade.

Em caso de rejeição da hipótese testada é ainda possível avaliar através de um caso especial da estatística escore quais as variáveis preditoras violaram o pressuposto do modelo de *odds* proporcionais. A estatística nesse caso também tem distribuição assintótica qui-quadrado com $(K-1)$ graus de liberdade.

A função de verossimilhança para esse modelo é dada por:

$$L = \prod_{i=1}^{n} \prod_{k=1}^{K} [\mathbb{P}(y_{ik} = k|\mathbf{X})]^{y_{ik}}$$

$$= \prod_{i=1}^{n} \prod_{k=1}^{K} [\pi_{ik}]^{y_{ik}}. \tag{2.9}$$

em que

$$\pi_{ik} = \begin{cases} \pi_{i1} = C_{i1}, & \text{se} \quad y_{ik} = 1 \\ \pi_{ik} = C_{ik} - C_{iK-1}, & \text{se} \quad 1 < y_{ik} < K-1 \\ \pi_{iK} = C_{iK}, & \text{se} \quad y_{ik} = K-1. \end{cases}$$

Os parâmetros neste modelo são estimados pelo método de máxima verossimilhança, utilizando-se o algoritmo escore de Fisher. A interpretação do modelo é análoga ao modelo de *odds* proporcionais apresentado anteriormente, contudo, vale ressaltar que para as variáveis preditoras cujo pressuposto foi violado, as probabilidades são calculadas mudando não apenas o intercepto, mas também o seu respectivo coeficiente ϱ_k.

Assim, para o estudo sobre sistemas de captação de água da chuva, o modelo de *odds* proporcionais parciais ajustado considerando como proporcionais apenas o efeito das variáveis sexo e frequência de banho é expresso por:

$$\log\left[\frac{\mathbb{P}(Y_i \leq k)}{1-\mathbb{P}(Y_i \leq k)}\right] = \log\left[\frac{\pi_1 + \ldots + \pi_k}{\pi_{k+1} + \ldots + \pi_K}\right]$$

$$= \gamma_k + \beta_1 \text{sexo} + \beta_2 \text{freq. banho} + \varrho_{1k} \text{grupo} + \varrho_{2k} \text{idade}$$

Os resultados obtidos para o modelo anterior são apresentados na Tabela 2.4.

Tabela 2.4: Estimativas dos parâmetros para o modelo de *odds* proporcionais parciais

variável	estimativa	ep
γ_1	-3,571	0,525
γ_2	-1,926	0,294
Grupo_1(com cisterna)	-0,290	0,341
Grupo_2(com cisterna)	0.206	0.198
sexo (M)	-0,235	0,195
freq.banho (>1x ao dia)	-0,459	0,291
idade.criança_1	0,041*	0,014
idade.criança_2	0,031*	0,008

* p-valor<0,05

Assim, a chance de poli ou monoinfecção é maior ($e^{0,206} = 1,23$) entre as crianças no grupo com cisterna em comparação às crianças do grupo sem cisterna. Em relação à variável idade, a cada 1 mês que se aumenta na idade da criança, a chance dela ser poliinfectada aumenta ($e^{0,041} = 1,04$). O mesmo ocorre em relação à chance de poli ou monoinfecção ($e^{0,031} = 1,03$).

Capítulo 3

Modelagem Longitudinal

A análise de dados com respostas correlacionadas é comum em estudos biomédicos, epidemiológicos, econômicos, dentre outros, onde medidas do mesmo indivíduo podem ser obtidas em diferentes ocasiões. As observações repetidas para um mesmo indivíduo caracterizam um estudo longitudinal, cujo objetivo é detectar possíveis mudanças na resposta dos indíviduos sob o tempo, além de avaliar quais fatores influenciam na heterogeneidade entre indivíduos (Fitzmaurice et al., 2011). Essa heterogeneidade, em geral, pode estar associada a fatores genéticos, ambientais, sociais, dentre outros. Dessa forma, o desenho de estudo longitudinal permite conhecer características do indivíduo que podem explicar tal heterogeneidade, e como é a mudança ao longo do tempo.

Em um estudo longitudinal, os participantes, ou mais geralmente as unidades a serem estudadas, são referidos como indivíduos ou sujeitos. Nesse sentido, a resposta do i-ésimo indivíduo, $i = 1, 2, \ldots, n$, tomada repetidamente ao longo do tempo, é definida por Y_{ij}, e pode ser agrupada em um vetor $m_i \times 1$ do tipo $\mathbf{Y}_i = (Y_{i1}^T, \ldots, Y_{im_i}^T)^T$, com $j = 1, \ldots, m_i$.

Associado a cada Y_{ij}, há um vetor $p \times 1$ de variáveis preditoras X_{ij}, que podem ou não mudar ao longo do tempo, ou seja, $\mathbf{X}_{ij} = (X_{i11}^T, \ldots, X_{im_ip}^T)^T$. Os vetores X_{ij} podem ser representados por uma matriz $m_i \times p$:

$$\mathbf{X}_i = \begin{pmatrix} X_{i11} & X_{i12} & \ldots & X_{i1p} \\ X_{i21} & X_{i22} & \ldots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{im_i1} & X_{im_i2} & \ldots & X_{im_ip} \end{pmatrix}$$

A análise dos dados em um estudo longitudinal requer técnicas de modelagem que levem em conta o fato de que as medidas repetidas de um mesmo indivíduo podem ser correlacionadas. Assim, os modelos usuais (Modelos Lineares Generalizados - GLM (McCullagh e Nelder, 1989)), que supõem independência entre as observações, não são

apropriados para analisar dados dessa natureza.

Um dos trabalhos pioneiros em análise de dados longitudinais foi escrito por Laird e Ware (1982), que, com base em uma classe mais geral de modelos lineares mistos (MLM), inicialmente introduzida por Harville (1977), descreveram uma classe flexível de modelos lineares mistos para modelar respostas contínuas, que incluía como casos especiais a ANOVA univariada para medidas repetidas, e os modelos para curvas de crescimento em dados longitudinais (Fitzmaurice et al., 2009). Em meados da década de 80, paralelamente ao desenvolvimento dos MLM's, um notável avanço surgiu na metodologia para analisar dados longitudinais discretos quando Liang e Zeger (1986a; 1986b) propuseram as Equações de Estimação Generalizadas (EEG). Essas equações de estimação são uma extensão natural do método de Quase-Verossimilhança proposto por Wedderburn (1974) para modelar respostas multivariadas em GLM's. Nos últimos 25 anos vários modelos que estendem os GLM's foram propostos para modelar respostas longitudinais.

Nas subseções a seguir, três classes de modelos de regressão para dados longitudinais são apresentadas.

3.1 Classes de Modelos

3.1.1 Modelos Marginais

Uma das metodologias mais populares para modelar dados cuja resposta tem caráter longitudinal, os modelos marginais fornecem um método unificado para analisar vários tipos de respostas longitudinais, evitando suposições sobre a distribuição do vetor de respostas, e baseando-se exclusivamente em suposições sobre a resposta média. Essa classe de modelos caracteriza a esperança marginal de uma variável resposta discreta ou contínua, como função de um conjunto de variáveis preditoras, sendo apropriada quando o foco da análise é inferir sobre a população média (Diggle et al., 2002). Assim, o termo 'marginal' nesse contexto indica que o modelo para a resposta média em cada ocasião não incorpora dependência sobre nenhum efeito aleatório ou sobre respostas anteriores (Fitzmaurice et al., 2011).

De forma geral, um modelo marginal pode ser especificado por:

1. $\mathbb{E}(Y_{ij}|X_{ij}) = \mu_{ij}$, que se assume depender de X_{ij} através de uma função de ligação do tipo:

$$g(\mu_{ij}) = \eta_{ij} = X_{ij}^T\beta,$$

em que β é um vetor $p \times 1$ de parâmetros da regressão marginal.

2. A variância condicional de cada Y_{ij} é assumida depender da média μ_{ij}, ou seja:

$$\text{Var}(Y_{ij}|X_{ij}) = \phi\upsilon(\mu_{ij}),$$

em que ϕ é um parâmetro de dispersão, e $\upsilon(\mu_{ij})$ é uma função conhecida da média μ_{ij}.

3. A correlação/associação entre o vetor de respostas repetidas intra-indivíduo é assumida ser função de um vetor adicional de parâmetros, denotado por α. Por exemplo:

   - $\text{Corr}(Y_{ij}, Y_{ik}) = \alpha^{|k-j|}$, se Y_{ij} é contínua ou uma contagem, e é modelada assumindo uma estrutura auto-regressiva de ordem 1 (AR-1) para a correlação;
   - $\log OR(Y_{ij}, Y_{ik}) = \alpha_{jk}$, se Y_{ij} é categórica, e é modelada assumindo um padrão não-estruturado para o log da *odds ratio*.

Os dois primeiros componentes do modelo marginal correspondem às especificações de um modelo linear generalizado. Já o terceiro componente do modelo incorpora a associação entre o vetor de respostas repetidas intra-indivíduo, e representa a principal extensão de GLM's para dados longitudinais. Assim, os modelos marginais especificam um GLM para as respostas longitudinais em cada ocasião, mas também incluem um modelo para a associação entre o vetor de respostas repetidas intra-indivíduo (Fitzmaurice et al., 2009).

A estimação dos parâmetros no modelo marginal é feita utilizando-se uma extensão multivariada do método de quase-verossimilhança proposto por Wedderburn (1974). O estimador de quase-verossimilhança para β é encontrado resolvendo-se a seguinte equação estimação de quase-verossimilhança:

$$\mathcal{Q}(\beta) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T V_i^{-1}\{\mathbf{Y}_i - \mu_i(\beta)\} = 0, \tag{3.1}$$

em que $V_i = \text{Var}(\mathbf{Y}_i)$.

A extensão multivariada do método de quase-verossimilhança, usada para estimar os parâmetros do modelo marginal, foi proposta por Liang e Zeger (1986a; 1986b), e é conhecida como EEG's. A ideia das EEG's consiste em substituir $\mathbf{Y}_i$ e μ_i na função de quase-verossimilhança pelos vetores $m_i \times 1$ de $\mathbf{Y}_i$ e μ_i, respectivamente, além de usar uma matriz de pesos V_i dada por:

$$V_i = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2}, \tag{3.2}$$

em que ϕ é um parâmetro de dispersão; A_i é uma matriz diagonal do tipo $A_i = \text{diag}\upsilon(\mu_{ij})$; $R_i(\alpha)$ é uma matriz de correlação $m_i \times m_i$, conhecida como "matriz de trabalho". O parâmetro α na matriz R_i representa o vetor de parâmetros associados com o modelo especificado para a $\text{Corr}(Y_i)$.

Em EEG, V_i é usualmente referida como matriz covariância de trabalho (esse termo é usado para enfatizar que V_i é apenas uma aproximação da verdadeira covariância). Além disso, ao contrário das equações de quase-verossimilhança, ela depende não apenas de β, mas também do parâmetro α. Assim, as equações de estimação para o modelo marginal são expressas por:

$$\mathcal{U}_1(\beta, \alpha) = \sum_{i=1}^{n} D_i^T V_i^{-1} \{\mathbf{Y}_i - \mu_i\} = 0 \tag{3.3}$$

em que $D_i = \dfrac{\partial \mu_i}{\partial \beta}$ é uma matriz $m_i \times m_i$ de derivadas.

A equação de estimação em (3.3) é referida na literatura como EEG de 1^a ordem. Liang e Zeger (1986a; 1986b) provaram que o estimador para β, em tal caso, é consistente e assintoticamente normal, mesmo que a matriz $R_i(\alpha)$ não seja corretamente especificada. Além disso, o parâmetro α na referida equação de estimação é tratado como perturbação.

Através de um processo iterativo em 2 estágios, os autores calcularam as estimativas para β utilizando uma versão modificada do algoritmo escore de Fisher, e estimaram α e ϕ pelo método dos momentos. Deste modo,

1. Assumindo independência entre as observações, obtêm-se a estimativa inicial para β;

2. Dadas as estimativas iniciais, pelo método dos momentos, para α e ϕ, estima-se V_i, e uma estimativa atualizada para β é obtida como solução da equação de estimação em (3.3);

3. Dada a estimativa para β, obtida no passo 2, encontram-se estimativas atualizadas para α e ϕ através dos resíduos. Assim,

$$\widehat{\phi} = \frac{\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{m_i} \widehat{r}_{ij}^2}{\displaystyle\sum_{i=1}^{n} m_i - p} \tag{3.4}$$

em que $\widehat{r}_{ij}^2 = \dfrac{(Y_{ij} - \widehat{\mu}_{ij})}{\sqrt{\upsilon(\widehat{\mu}_{ij})}}$, e $\widehat{\mu}_{ij} = g^{-1}(\widehat{\eta})$, sendo $\widehat{\eta} = X_{ij}^T \widehat{\beta}$.

As estimativas atualizadas para $\widehat{\alpha}$ irão depender da estrutura escolhida para a matriz $R_i(\alpha)$, ou seja:

a. Se $R_i(\alpha)$ for do tipo independente, assume-se que observações repetidas para cada indivíduo são não-correlacionadas. Desta forma:

$$R_i(\alpha) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

b. Se $R_i(\alpha)$ for não-estruturada, tem-se um $\widehat{\alpha}$ para cada tempo, ou seja:

$$\hat{\alpha}_{jk} = \frac{\sum_{i=1}^{n} \widehat{r}_{ij}^2 \widehat{r}_{ik}^2}{\widehat{\phi}(n-p)}$$

Neste caso, $R_i(\alpha)$ tem a seguinte representação:

$$R_i(\alpha) = \begin{pmatrix} 1 & \alpha_{12} & \dots & \alpha_{1m_i} \\ \alpha_{21} & 1 & \dots & \alpha_{2m_i} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{m_i1} & \alpha_{m_i2} & \dots & 1 \end{pmatrix}.$$

c. Se $R_i(\alpha)$ for do tipo simétrica composta, ter-se-á um único $\widehat{\alpha}$ para todos os tempos, ou seja:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} \sum_{j \neq k} \widehat{r}_{ij}^2 \widehat{r}_{ik}^2}{\widehat{\phi}(N^* - p)},$$

em que $N^* = \sum_{i=1}^{n} \frac{m_i(m_i-1)}{2}$. $R_i(\alpha)$, neste caso, tem a seguinte representação:

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

d. Por fim, se $R_i(\alpha)$ for do tipo AR-1 (auto-regressiva de ordem 1), então $\hat{\alpha}$ é dado por:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} \sum_{j \leq m_i - 1} \widehat{r}_{ij}^2 \widehat{r}_{i,j+1}^2}{\widehat{\phi}(K^* - p)},$$

em que $K^* = \sum_{i=1}^{n} (m_i - 1)$, com $R_i(\alpha)$ sendo dada por:

$$R(\alpha) = \begin{pmatrix} 1 & \alpha & \dots & \alpha^{|1-m_i|} \\ \alpha & 1 & \dots & \alpha^{|2-m_i|} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^{|1-m_i|} & \alpha^{|2-m_i|} & \dots & 1 \end{pmatrix}.$$

Assim, o processo iterativo anterior é repetido até que se obtenha a convergência. Os autores sugerem a seguinte expressão para obter as atualizações de $\widehat{\beta}$:

$$\widehat{\beta}^{(m+1)} = \widehat{\beta} - \left[\sum_{i=1}^{n} D_i^{T(m)} \left\{\widetilde{V}_i^{(m)}\right\}^{-1} D_i^{(m)}\right]^{-1} \sum_{i=1}^{n} D_i^{T(m)} \left\{\widetilde{V}_i^{(m)}\right\}^{-1} \left\{Y_i - \mu_i(\widehat{\beta}^{(m)})\right\} \quad (3.5)$$

em que $\widetilde{V}_i^{(m)} = V_i\{\widehat{\beta}^{(m)}, \widehat{\alpha}^{(m)}\}$.

Sob certas condições de regularidade, Liang e Zeger (1986a - Teorema 2) mostraram que, quando $n \to \infty$, $\widehat{\beta}$ é um estimador consistente e assintoticamente não-viesado para β, ou seja,

$$\sqrt{n}(\widehat{\beta} - \beta) \sim N_M(\beta, V)$$

em que

$$V = \lim_{n\to\infty} n \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1} \left\{\sum_{i=1}^{n} D_i^T V_i^{-1} \mathrm{cov}(Y_i) V_i^{-1} D_i\right\} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1}, \quad (3.6)$$

e V é a matriz de covariância, e $\mathrm{Cov}(Y_i) = \mathbb{E}\left[(Y_i - \hat{\mu}_i)^T (Y_i - \hat{\mu}_i)\right]$.

Se a estimação de β é o interesse primário na modelagem, sabe-se que a escolha da matriz $R_i(\alpha)$ não afetará as propriedades assintóticas de $\widehat{\beta}$. Contudo, se o objetivo for também modelar o parâmetro α, então a especificação correta da matriz $R_i(\alpha)$ será importante. Há um critério proposto por Pan (2001) que ajuda a selecionar a melhor estrutura para a matriz de trabalho no modelo ajustado via EEG. O critério denotado por QIC (*Quasi Information Criterion*) substitui a função de verossimilhança constante na expressão do critério de Akaike (AIC) pela função de quase-verossimilhança. A 'melhor' estrutura para a matriz $R_i(\alpha)$ será aquela cujo QIC for o menor dentre todas as estruturas candidatas.

Toda a teoria anteriormente discutida sobre o procedimento utilizado para estimação dos parâmetros do modelo marginal encontra-se em dois artigos publicados por Liang e Zeger (1986a e 1986.b), onde eles apresentam as equações de estimação generalizadas, comumente conhecidas na literatura como EEG1. Como já foi dito, o foco principal das equações de estimação descritas por esses autores está na estimação do

parâmetro β.

Nos anos seguintes à publicação dos artigos de Liang e Zeger vários autores propuseram extensões das EEG's de 1^a ordem com o objetivo de modelar conjuntamente o vetor de parâmetros (β, α), ou até mesmo modelar a estrutura de correlação/associação entre as medidas repetidas, considerando medidas de associação apropriadas, quando, por exemplo, tem-se uma resposta binária (Prentice, 1988; Prentice e Zhao, 1991; Lipsitz et al., 1991; Carey et al., 1993). Carey et al. (1993) propuseram o procedimento conhecido como regressão logística alternada (ALR, em inglês), através da qual modelam simultaneamente o vetor de parâmetros (β, α), onde α mede a associação entre as respostas repetidas para cada indivíduo, sendo estimado em termos da razão de chance.

De forma geral, o procedimento proposto pelos autores combina as EEG's de 1^a ordem para estimação do parâmetro β, com uma nova equação na regressão logística para estimar o parâmetro α. Os autores denotaram por γ_{ijs} o log da razão de chances entre as respostas Y_{is} e Y_{ij}, e definiram $\mu_{ij} = \mathbb{P}(Y_{ij} = 1)$ e $\nu_{ijk} = \mathbb{P}(Y_{ij} = 1, Y_{is} = 1)$. Assim:

$$\text{logito} \quad \mathbb{P}(Y_{ij} = 1 | Y_{is} = y_{is}) = \gamma_{ijs} y_{is} + \log\left(\frac{\mu_{ij} - \nu_{ijs}}{1 - \mu_{ij} - \mu_{is} + \nu_{ijs}}\right). \tag{3.7}$$

Em geral, assume-se que $\gamma_{ijs} = \widetilde{Z}_{ijs}^T \alpha$, onde $\widetilde{Z}_{ijs}^T$ é um vetor $q \times 1$ de variáveis preditoras que especifica a forma da associação entre Y_{is} e Y_{ij}; e α é um vetor $q \times 1$ de parâmetros relacionados com as associações.

O procedimento proposto utiliza um processo iterativo para estimar o vetor $\delta = (\beta, \alpha)$, que se alterna entre dois passos até obter a convergência, ou seja:

Passo 1: Para um dado valor de α, estima-se β usando EEG1 (expressão(3.3));

Passo 2: Para o β estimado no passo 1, estima-se α usando uma regressão logística de Y_{ij} sobre cada $Y_{is}(s > j)$, considerando o lado direito da expressão (3.7) como *offset* no modelo.

Assim, a regressão logística alternada estima δ de forma simultânea através da solução das seguintes equações de estimação:

$$\mathcal{U}_\beta = \sum_{i=1}^{n} D_i^T [\text{cov}(Y_i)]^{-1} \{\mathbf{Y}_i - \mu_i(\beta)\} = 0 \qquad \text{(EEG1)}$$

$$\mathcal{U}_\alpha = \sum_{i=1}^{n} T_i^T M_i^{-1} H_i = 0 \tag{3.8}$$

onde T_i é uma matriz $C_{m_i}^2 \times q$ de derivadas, cujos elementos são $\dfrac{\partial \zeta_i}{\partial \alpha}$, sendo ζ_i um vetor de tamanho $C_{m_i}^2$, com elementos dados por:

$$\zeta_{ijs} = \mathbb{E}(Y_{ij}|Y_{is} = y_{is}) = \text{logito}^{-1}\left\{\gamma_{ijs}y_{is} + \log\left(\frac{\mu_{ij} - \nu_{ijs}}{1 - \mu_{ij} - \mu_{is} + \nu_{ijs}}\right)\right\}$$

em que M_i é uma matriz diagonal de dimensão $C^2_{m_i} \times C^2_{m_i}$, cujos elementos da diagonal são $\zeta_{ijs}(1 - \zeta_{ijs})$; H_i é um vetor de resíduos expresso por $H_{ijs} = Y_{ij} - \zeta_{ijs}$.

A proposta desses autores foi mais tarde estendida por Heagerty e Zeger (1996) para estimar os parâmetros do modelo marginal quando a resposta de interesse é politômica ordinal.

É válido ressaltar que nessa classe de modelos a interpretação dos parâmetros estimados é feita de forma análoga aos modelos apresentados no Capítulo 2.

3.1.2 Modelos Lineares Generalizados Mistos - GLMM

Como anteriormente mencionado, na análise de dados onde a mesma unidade experimental é medida várias vezes, os modelos clássicos de regressão, que apresentam apenas efeitos fixos além do erro experimental, não podem ser utilizados visto que a pressuposição básica de independência entre as observações não pode ser assumida. Outra classe de modelos bastante utilizada em estudos longitudinais são os modelos mistos, que incluem efeitos aleatórios no modelo de efeitos fixos, a nível do indivíduo, modelando a heterogeneidade entre indivíduos e induzindo, assim, uma covariância entre as respostas repetidas.

Os modelos lineares generalizados mistos (GLMM, em inglês) são uma combinação natural da classe de modelos lineares mistos (MLM) (Laird e Ware, 1982) e dos modelos lineares generalizados (Nelder e Wedderburn, 1972). Os MLM's são uma família de modelos para analisar dados longitudinais, que inclui como casos particulares os modelos para curvas de crescimento e a ANOVA univariada para medidas repetidas. Nessa família, que apresenta tanto efeitos fixos como aleatórios, além do erro experimental, o vetor de respostas segue uma distribuição normal. A metodologia descrita por Laird e Ware consiste em um modelo de dois níveis para medidas repetidas, sendo baseada no trabalho de Harville (1977), onde o autor discute o problema na estimação dos componentes da variância em modelos de análise de variância. No primeiro nível do modelo são introduzidos os parâmetros populacionais, efeitos individuais e a variação intra-indivíduo, enquanto que a variação entre-indivíduos entra no segundo nível do modelo.

Em princípio, todo modelo linear que contenha a média geral ou uma constante μ, tomada como fixa, e um termo referente ao erro, assumido como aleatório, é um modelo

misto. Contudo, a denominação 'modelo linear misto' é geralmente reservada a modelos lineares que contenham efeitos fixos além de μ, e qualquer outro termo aleatório além do erro. Assim, o modelo proposto é dado pela seguinte expressão:

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i, \tag{3.9}$$

em que Y_i é um vetor m_i-dimensional; β é um vetor $p \times 1$ de parâmetros desconhecidos; $\mathbf{X}_i$ e $\mathbf{Z}_i$ são matrizes conhecidas, de dimensão $m_i \times p$ e $m_i \times q$, respectivamente; $\mathbf{b}_i$ é um vetor $q \times 1$ de efeitos aleatórios a nível do indivíduo, independente de $\mathbf{X}_i$; e $\epsilon \sim N(0, \Sigma_i)$, onde $\mathbf{\Sigma}_i$ é a matriz de covariâncias, cuja dimensão é $m_i \times m_i$. Os efeitos $\mathbf{b}_i$'s seguem uma distribuição normal, com média 0 e matriz de covariância $\mathbf{G}$, de dimensão $m_i \times m_i$, e são independentes de ϵ_i.

Marginalmente, os $\mathbf{Y}_i$ são independentes e seguem uma distribuição normal com média $\mathbf{X}_i\beta$ e matriz de covariâncias $\mathbf{\Sigma}_i$, dada por $\mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^T + \sigma^2\mathbf{I}$. Assim, a função de verossimilhança para o modelo descrito em (3.9) é dada por:

$$L(\beta, \sigma^2, \mathbf{G}) = \prod_{i=1}^{n} \mid \mathbf{\Sigma}_i \mid^{-1/2} \exp\left\{\frac{1}{2}(\mathbf{Y}_i - \mathbf{X}_i\beta)^T\mathbf{\Sigma}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\beta)\right\}. \tag{3.10}$$

Para estimar os parâmetros do modelo linear misto, Laird e Ware sugerem o uso do algoritmo EM para obtenção das estimativas de máxima verossimilhança - MV. Contudo, a estimação por MV fornece estimativas viesadas para os componentes da variância. Isso se deve ao fato de que a função de verossimilhança completa, descrita em (3.10), envolve o vetor β, de parâmetros fixos, que precisa ser estimado juntamente com os componentes de variância, e a perda de graus de liberdade na estimação dos efeitos fixos não é levada em consideração na estimação de MV dos componentes de variância. Uma solução é utilizar o método de máxima verossimilhança restrita (Patterson e Thompson, 1971), que em geral, diminui o viés dos estimadores de máxima verossimilhança para os componentes de variância. Para esse método a seguinte função de verossimilhança é considerada:

$$L(\sigma^2, \mathbf{G}) = \prod_{i=1}^{n} \mid \mathbf{\Sigma}_i \mid^{-1/2} \mid \mathbf{X}_i^T\mathbf{\Sigma}_i\mathbf{X}_i \mid^{-1/2} \exp\left\{\frac{1}{2}(\mathbf{Y}_i - \mathbf{X}_i\beta)^T\mathbf{\Sigma}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\beta)\right\}.$$

Existem, contudo, muitas situações onde o vetor de respostas não pode ser modelado a partir da distribuição normal, como é pressuposto nos MLM's. Em tais casos, as opções de distribuição deste vetor de respostas podem ser estendidas de forma que ele pertença à família exponencial de distribuições. Feito isto, os MLM's são estendidos para uma classe mais ampla de modelos conhecida como GLMM. A premissa básica subjacente

ao GLMM para dados longitudinais é a suposição de heterogeneidade entre os indivíduos na população estudada, em um subconjunto dos coeficientes de regressão a partir de um GLM (Fitzmaurice et al., 2011).

Assim, seja Y_{ij} a resposta de interesse para o i-ésimo indivíduo na j-ésima ocasião. Assumimos que a distribuição condicional de cada Y_{ij}, dado um vetor $q \times 1$ de efeitos aleatórios $\mathbf{b}_i$, é membro da família exponencial. Além disso, dados os $\mathbf{b}_i$, os Y_{ij} são condicionalmente independentes. Os modelos lineares generalizados mistos podem ser especificados por:

1. $\mathbb{E}(Y_{ij}|b_i) = \mu_{ij}$, que se assume depender dos efeitos fixos e aleatórios através de uma função de ligação do tipo:

$$g(\mu_{ij}) = \eta_{ij} = X_{ij}^T\beta + Z_{ij}^T b_i.$$

2. $\text{Var}(Y_{ij}|b_i) = \phi\upsilon(\mu_{ij})$, em que ϕ é um parâmetro de dispersão e $\upsilon(\cdot)$ uma função conhecida da média condicional;

3. Assume-se que os efeitos b_i seguem distribuição normal multivariada, com média 0 e matriz de variância e covariância $\mathbf{G}$.

Embora a introdução de efeitos aleatórios no modelo misto possa dar conta da correlação entre as respostas longitudinais, ela tem importantes implicações na interpretação dos coeficientes da regressão. A interpretação dos β's nessa classe de modelos é ao nível do indivíduo, ou seja, β mede a mudança na resposta do indivíduo i a cada unidade que aumentamos (ou diminuímos) em X_{ij}. Assim, a forma como interpretam-se os coeficientes do modelo misto é o que o difere dos modelos marginais, cujo alvo da inferência é a média da população (Fitzmaurice et al, 2011).

É possível, através do teste da razão de verossimilhança, verificar se o efeito $\mathbf{b}_i$ é estatisticamente significativo. Ou seja, sob H_0 queremos testar se $\sigma_b^2 = 0$, onde σ_b^2 é a variância de b_i. Se o teste apontar para um efeito estatisticamente significativo implica, então, que a inclusão de $\mathbf{b}_i$ no modelo é importante para explicar a variação entre indivíduos decorrente das medidas repetidas. O teste usual para tal verificação possui distribuição assintótica qui-quadrado com 1 grau de liberdade. Esse teste é, em geral, conservador por testar na fronteira do espaço paramétrico, sendo aconselhável considerar uma mistura de qui-quadrados $(\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2)$ como distribuição.

Além de avaliar se o efeito σ_b^2 é estatisticamente significativo, é também de frequente interesse expressá-la em termos de uma correlação intra-classe (Snijders e Bosker, 1999), nos modelos onde apenas o intercepto é aleatório. O coeficiente de correlação intra-classe indica a proporção não explicada da variância a nível do indivíduo, ou seja, ele

reflete a magnitude da variância entre indivíduos (Hedeker e Gibbons, 2006). De forma geral, o coeficiente de correlação intra-classe (ICC, em inglês) é expresso por:

$$\text{ICC} = \frac{\widehat{\sigma}_b^2}{\widehat{\sigma}_b^2 + \widehat{\sigma}_\epsilon^2}$$

em que $\widehat{\sigma}_b^2$ representa a variância estimada para o efeito aleatório, e $\widehat{\sigma}_\epsilon^2$ a variância estimada dos erros do modelo.

A função de verossimilhança nessa classe de modelos é expressa por:

$$\begin{aligned} L(\beta, \phi, G) &= \prod_{i=1}^{n} \int_{b_i} f(Y_i, b_i) db_i \\ &= \prod_{i=1}^{n} \int_{b_i} f(Y_i|b_i) f(b_i) db_i \\ &= \prod_{i=1}^{n} \int_{b_i} \left[\prod_{j=1}^{m_i} f(Y_{ij}|b_i) \right] f(b_i) db_i. \end{aligned} \tag{3.11}$$

Os parâmetros do modelo linear generalizado misto são obtidos maximizando-se a função em (3.11). Contudo, em geral, a integral anterior não possui forma analítica, e para obtenção das estimativas por máxima verossimilhança faz-se então necessária a utilização de métodos de aproximação numérica, como Laplace ou Quadratura gaussiana (Pinheiro e Bates, 1995). Os métodos aproximados, por sua vez, podem ser computacionalmente intensivos se o número de efeitos aleatórios no modelo for relativamente grande. Uma alternativa em tais casos é fazer uso do método de quase-verossimilhança penalizada (PQL, em inglês) que estende o método de quase-verossimilhança aplicando uma penalidade na função de quase-verossimilhança, referente à distribuição do efeito aleatório (Breslow e Clayton, 1993), ou utilizar a verossimilhança H, proposta por Lee e Nelder (1996).

3.1.3 Modelos de Transição

Em um estudo longitudinal se o foco/objetivo for fazer predição da resposta, os modelos marginais e mistos são inadequados, e em tal caso aconselha-se o uso dos chamados modelos de transição.

Os modelos de transição são uma classe de modelos que estende os GLM's para dados longitudinais, onde a distribuição condicional da resposta Y_{ij} é descrita como uma

função explícita das respostas passadas e de um vetor de variáveis preditoras. No modelo de transição a distribuição condicional de Y_{ij} dado as respostas passadas é especificada através de um GLM. Sob esse modelo assume-se correlação entre $Y_{i1}, \ldots, Y_{im_i}$, pois as respostas passadas $Y_{i1}, \ldots, Y_{ij-1}$ influenciam de forma explícita na resposta atual, entrando no modelo como preditoras adicionais (Diggle et. al, 2002).

O modelo de transição mais utilizado são os Markovianos, onde a distribuição de $Y_{ij}|\mathcal{H}_{ij}$, sendo $\mathcal{H}_{ij} = \{Y_{id}, d = 1, \ldots, j-1\}$ a história do indivíduo i no tempo j, depende apenas de Q observações a priori, $Y_{ij-1}, \ldots, Y_{ij-Q}$,sendo Q a ordem da cadeia.

De forma geral, um modelo de transição pode ser especificado por:

1. $\mathbb{E}(Y_{ij}|X_{ij}, \mathcal{H}_{ij}) = \mu_{ij}$, onde a média condicional de Y_{ij} é expressa como função do vetor de variáveis preditoras X_{ij} e de $\mathcal{H}_{ij}$, através da seguinte função de ligação:

$$g(\mu_{ij}) = X_{ij}^T\beta + \sum_{q=1}^{Q} f_q(\mathcal{H}_{ij}, \alpha) \tag{3.12}$$

   em que $f_q(\cdot)$ são funções conhecidas.

2. A variância condicional de Y_{ij} é assumida depender de μ_{ij}, ou seja:

$$\text{Var}(Y_{ij}|X_{ij}, \mathcal{H}_{ij}) = \phi\upsilon(\mu_{ij})$$

3. A correlação entre $Y_{i1}, \ldots, Y_{in_i}$ é avaliada através do parâmetro α que aparece na função $f_q(\cdot)$.

Assim, se a resposta de interesse for binária, por exemplo, a formulação do modelo de transição de ordem Q seria:

$$\text{logito}\mathbb{P}(Y_{ij} = 1|X_{ij}, \mathcal{H}_{ij}) = X_{ij}^T\beta + \sum_{q=1}^{Q} \alpha_q y_{ij-q},$$

em que β relaciona a resposta média às variáveis preditoras após ajustar pelas Q respostas passadas; e α_q descreve a dependência da resposta atual nas respostas passadas.

É importante ressaltar que a interpretação do coeficiente β é específica para a ordem do modelo, e condicional às Q respostas passadas. Ressalta-se, ainda, que essa classe de modelos não é interessante se o objetivo for fazer inferências marginais, ou seja, inferências que descrevam a associação entre Y_{ij} e X_{ij} apenas.

Nos modelos de transição os parâmetros podem ser estimados por máxima verossimilhança. Desta maneira, a distribuição conjunta de $Y_{i1}, \ldots, Y_{im_i}$ pode ser escrita na forma:

$$f(Y_{i1},\ldots,Y_{im_i}) = f(Y_{im_i}|Y_{im_{i-1}},\ldots,Y_{i1})f(Y_{im_{i-1}}|Y_{im_{i-2}},\ldots,Y_{i1})\ldots f(Y_{i2}|Y_{i1})f(Y_{i1})$$

Ou ainda:

$$f(Y_{i1},\ldots,Y_{im_i};\beta,\alpha) = f(Y_{i1};\beta,\alpha)\prod_{j=2}^{m_i} f(Y_{ij}|\mathcal{H}_{ij}) \tag{3.13}$$

Já a distribuição condicional de Y_{ij} em um modelo de ordem Q é dada por:

$$f(Y_{ij}|\mathcal{H}_{ij}) = f(Y_{ij}|Y_{ij-1},\ldots,Y_{ij-Q})$$

Assim, a contribuição da verossimilhança para o $i-$ésimo indivíduo será:

$$L(Y_{i1},\ldots,Y_{iM_i}) = f(Y_{i1},\ldots,Y_{iQ})\prod_{j=Q+1}^{m_i} f(Y_{ij}|Y_{ij-1},\ldots,Y_{ij-Q}) \tag{3.14}$$

Ressalta-se, novamente, que essa classe de modelos especifica apenas a distribuição condicional de $Y_{ij}|\mathcal{H}_{ij}$. Dessa maneira, a verossimilhança das Q primeiras observações $f(Y_{i1},\ldots,Y_{iQ})$ não é diretamente especificada. Assim sendo, os parâmetros β e α são estimados maximizando-se a função de verossimilhança condicional, que é obtida omitindo-se $f(Y_{i1},\ldots,Y_{iQ})$ na função descrita em (3.14).

A função de verossimilhança condicional é então expressa por:

$$\begin{aligned} L(Y_{ij};\beta,\alpha) &= \prod_{i=1}^{n} f(Y_{iQ+1},\ldots,Y_{im_i}|Y_{i1},\ldots,Y_{iQ}) \\ &= \prod_{i=1}^{n}\prod_{j=Q+1}^{m_i} f(Y_{ij}|\mathcal{H}_{ij}) \end{aligned}$$

Na classe dos modelos de transição é ainda possível caracterizar a correlação entre as observações repetidas incorporando uma estrutura para a média marginal. Tal caracterização é feita através dos chamados modelos de transição marginalizados (Azzalini, 1994; Heagerty, 2002), que não serão discutidos neste trabalho.

Capítulo 4

Modelos para respostas ordinais longitudinais

Na literatura sobre modelos para dados longitudinais há uma vasta referência de livros e artigos que discutem as diferentes técnicas para modelar respostas discretas e contínuas. Contudo, nos últimos anos tem sido crescente o interesse em modelar respostas politômicas, em especial as ordinais, em estudos longitudinais, e algumas metodologias alternativas têm sido utilizadas para tal (Miller et al., 1993; Hedeker e Gibbons, 1994; Heagerty e Zeger, 1996; Hedeker e Mermelstein, 2000; Lee e Daniels, 2007).

Os modelos usuais para analisar dados longitudinais, descritos na seção 3, podem ser utilizados para modelar respostas politômicas. Seguindo, por exemplo, a linha dos modelos marginais, alguns autores propuseram diferentes procedimentos para modelar res-postas politômicas em estudos longitudinais (Clayton, 1992; Gange et al., 1993; Miller et al., 1993; Lipsitz et al., 1994). Na maioria dos trabalhos os autores adotaram as equações de estimação propostas por Liang e Zeger (1986a; 1986b), apresentadas na subseção 3.1.1, e fizeram uso do modelo de chances proporcionais para modelar a resposta. Clayton (1992), por exemplo, usou as EEG's de 1^a ordem para modelar respostas ordinais, mas as reescreveu a partir de $K-1$ variáveis binárias, e utilizou o modelo de chances proporcionais para ajustar os $K-1$ modelos. Gange et al. (1993) trataram o parâmetro relacionado à associação entre as medidas repetidas como perturbação, modelando apenas a resposta média. Uma limitação dos dois trabalhos citados é que os autores utilizaram a matriz de trabalho que assume independência entre as observações no processo de estimação dos parâmetros do modelo. Já Miller et al. (1993) estenderam as EEG's de 1^a ordem para acomodar respostas politômicas. Os referidos autores definiram um segundo conjunto de equações de estimação para o parâmetro α, e mostraram que, sob certas condições, as EEG's de 1^a ordem se reduzem às equações de mínimos quadrados. Lipsitz et al. (1994) estenderam as equações de Liang e Zeger (1986a; 1986b) para modelar a correlação

entre respostas repetidas nominais ou ordinais. Os autores especificaram a matriz de correlação em termos das correlações entre os pares de respostas repetidas para o i-ésimo indivíduo, e das matrizes de cova-riâncias marginais. Posterior a esses trabalhos, Heagerty e Zeger (1996) propuseram um novo conjunto de equações de estimação para analisar respostas ordinais correlacionadas. Os autores reescreveram as EEG's de 1^a e 2^a ordem, e a regressão logística alternada, de forma a acomodar as respostas ordinais. Eles introduziram dois modelos de regressão: um modelo para descrever as probabilidades acumuladas da resposta ordinal, e outro para a associação entre os pares de resposta. Para o primeiro modelo, assumiram um modelo de chances proporcionais para modelar a resposta média, e no segundo modelo utilizaram a razão de chances global como medida de associação. É também possível modelar a resposta ordinal, via EEG's, utilizando-se o modelo de chances proporcionais parciais, contudo, sua implementação, até então, somente está disponível no *software* SAS.

Além das propostas de modelagem utilizando a classe de modelos marginais, surgiram inúmeros trabalhos propondo modelar respostas ordinais a partir da classe de modelos lineares generalizados mistos (Hedeker e Gibbons, 1994; Hedeker e Mermelstein, 1998,2000; Hartzel et al., 2001; Hedeker, 2003). Hedeker e Gibbons (1994) propuseram um modelo de efeitos aleatórios para analisar respostas ordinais em estudos longitudinais fazendo uso das funções logística e probito. Os autores definiram um modelo de efeitos aleatórios usando a terminologia multinível, e para obter as estimativas dos parâmetros utilizaram o método de máxima verossimilhança. Hedeker e Mermelstein (1998; 2000) descreveram uma extensão do modelo de chances proporcionais para dados longitudinais que permite a não-proporcionalidade das *odds* em um subconjunto de vari-áveis preditoras. Nesse artigo, os autores descrevem o modelo de efeitos aleatórios usando a proposta de Peterson e Harrell (1990), relaxando, assim, o pressuposto de chances proporcionais. Hartzel et al. (2001) apresentaram um procedimento geral para modelar respostas nominais ou ordinais. Os autores descreveram um modelo logístico de categorias adjacentes para as respostas ordinais, usando a definição de um GLMM. Para respostas nominais, estenderam o modelo logito generalizado, permitindo uma estrutura geral para a correlação dos efeitos aleatórios. Hedeker (2003) novamente descreve um GLMM para modelar respostas politômicas nominais, onde o modelo proposto acomoda múltiplos efeitos aleatórios e permite, adicionalmente, uma forma mais geral para as variáveis preditoras do modelo. O método de máxima verossimilhança é utilizado para estimar os parâmetros do modelo.

Outra classe de modelos que também pode ser usada para modelar respostas politômicas são os modelos de transição, que avaliam a probabilidade de transição da resposta de uma categoria para outra, descrevendo a distribuição condicional de cada resposta Y_{ij} como uma função explícita das respostas passadas e de variáveis preditoras (Diggle et al, 2002; Ganjali e Rezaee, 2007). Nessa classe de modelos é também possível

ajustar um modelo de *odds* proporcionais parciais, caso o pressuposto de proporcionalidade das *odds* não possa ser verificado no modelo de chances proporcionais. Ainda dentro dessa classe de modelos é possível caracterizar a correlação entre as observações repetidas incorporando uma estrutura para a média marginal, sendo tal caracterização feita através dos chamados modelos de transição marginalizados, propostos inicialmente para modelar respostas binárias longitudinais (Heagerty e Zeger, 2000; Heagerty, 2002), e mais recentemente estendidos de forma a acomodar dados longitudinais ordinais (Lee e Daniels, 2007).

A escolha da estratégia de modelagem a ser utilizada no estudo longitudinal dependerá da pergunta de interesse do pesquisador. Nas seções a seguir os modelos para respostas politômicas ordinais considerados nesse trabalho são apresentados sob a ótica das três classes de modelos de regressão para dados longitudinais apresentadas na seção 3.1 para o caso geral.

4.1 Modelos marginais para respostas ordinais

O modelo marginal para a resposta ordinal Y_{ij} pode ser definido de forma geral através do modelo de *odds* proporcionais parciais (Peterson e Harrel, 1990). Uma extensão natural do referido modelo para dados longitudinais é dada por:

$$\log\left[\frac{\mathbb{P}(Y_{ij}\leq k)}{1-\mathbb{P}(Y_{ij}\leq k)}\right]=\gamma_k+X_{ij}^T\beta+\widetilde{X}_{ij}^T\varrho_k,\quad k=1,2,\ldots,K-1. \tag{4.1}$$

É válido ressaltar que o modelo (4.1) é uma extensão do modelo de *odds* proporcionais onde o efeito ϱ_k varia de acordo com os k pontos de corte da resposta, sendo $\widetilde{X}_{ij}$ um subconjunto de X_{ij} para os quais o efeito está variando em k. Se $\varrho_k=0$, o modelo acima se reduz ao modelo de *odds* proporcionais:

$$\log\left[\frac{\mathbb{P}(Y_{ij}\leq k)}{1-\mathbb{P}(Y_{ij}\leq k)}\right]=\gamma_k+X_{ij}^T\beta,\quad k=1,2,\ldots,K-1 \tag{4.2}$$

No modelo definido em (4.2), as mudanças nos $K-1$ logitos cumulativos, ao longo do tempo, estão relacionadas com as variáveis preditoras. Embora o modelo inclua $K-1$ interceptos γ_k, ele assume que o efeito das variáveis preditoras é o mesmo entre os $K-1$ logitos, o que é equivalente a assumir que o efeito das variáveis preditoras sobre as *odds* cumulativas são proporcionais (Fitzmaurice et al, 2011). Quando o pressuposto de

proporcionalidade das *odds* não for verificado, pode-se considerar o modelo (4.1).

A construção de um MLG para as probabilidades acumuladas requer tratar as respostas ordinais como um conjunto de $K-1$ variáveis binárias da forma:

$$U_{ijk} = \begin{cases} 1, & \text{se} \quad Y_{ij} \leq k \\ 0, & \text{se} \quad Y_{ij} > k. \end{cases}$$

Uma especificação geral do modelo marginal para respostas ordinais é dada por:

1. logito $(F_{ijk}) = \gamma_k + X_{ij}^T\beta$, em que $F_{ijk} = \mathbb{P}(U_{ijk} = 1|X_{ij}) = \mathbb{P}(Y_{ij} \leq k|X_{ij})$;
2. $\text{Var}(U_{ijk}|X_{ij}) = F_{ijk}(1 - F_{ijk})$;
3. Ao especificarmos a associação intra-indivíduo (α), a correlação entre os componentes de $(U_{ij1}, \ldots, U_{ij,K-1})$ na j-ésima ocasião é uma função conhecida da média por meio de F_{ij}.

Os componentes de $(U_{ij1}, \ldots, U_{ij,K-1})$ na especificação 3 são correlacionados, e tal correlação decorre do fato de que as probabilidades das K respostas multinomiais devem necessariamente somar 1 (Fitzmaurice et al., 2011). Assim, por exemplo, a correlação entre U_{ij1} e U_{ij2}, pode ser expressa por:

$$\text{Corr}(U_{ij1}, U_{ij2}) = \frac{F_{ij1} - F_{ij1}F_{ij2}}{\sqrt{F_{ij1}F_{ij2}(1 - F_{ij1})(1 - F_{ij2})}}.$$

Alternativamente aos modelos para correlação, é possível especificar a associação entre os pares de respostas ordinais fazendo-se uso da razão de chances global como medida de associação. Lipsitz et al. (1991) usaram a razão de chances como medida de associação quando a resposta de interesse era binária, em contrapartida ao modelo para correlação proposto por Prentice (1988), enquanto Heagerty e Zeger (1996) consideraram a razão de chances global como medida de associação ao definirem o modelo marginal para os pares de respostas repetidas ordinais. Assim, a razão de chances global para o i-ésimo indivíduo, com respostas $Y_{is} = k_1$ e $Y_{ij} = k_2$, por exemplo, é definida como:

$$\psi_{isj}(k_1, k_2) = \frac{\overline{F}_{isj}(k_1, k_2)\left[1 - F_{isk_1} - F_{ijk_2} + \overline{F}_{isj}(k_1, k_2)\right]}{\left[F_{isk_1} - \overline{F}_{isj}(k_1, k_2)\right]\left[F_{ijk_2} - \overline{F}_{isj}(k_1, k_2)\right]} \tag{4.3}$$

em que $\overline{F}_{isj}(k_1, k_2) = \mathbb{P}(Y_{is} \leq k_1, Y_{ij} \leq k_2)$.

Segundo Heagerty e Zeger (1996), o parâmetro $\psi_{isj}(k_1, k_2)$ na expressão (4.3) pode ser visto como uma razão de chances marginal, que busca capturar a associação entre os

pares de respostas Y_{is} e Y_{ij}, nas categorias k_1 e k_2, respectivamente. É válido ressaltar que é possível assumir uma única razão de chances global, independente dos K pontos de corte da resposta ordinal se a matriz de trabalho R_i escolhida for do tipo simétrica composta.

A estimação dos parâmetros no modelo marginal para resposta ordinal é feita utilizando-se a extensão proposta por Heagerty e Zeger (1996). Assim, segundo estes autores, se o interesse primário é na estimação do parâmetro β, as equações de estimação de 1^a ordem, assumindo a natureza ordinal da resposta, são representadas matricialmente por:

$$\begin{bmatrix} \mathcal{U}_1(\beta,\alpha) \\ \mathcal{U}_2(\beta,\alpha) \end{bmatrix} = \sum_{i=1}^{n} \begin{bmatrix} \frac{\partial \mu_i}{\partial \beta} & 0 \\ 0 & \frac{\partial \sigma_i}{\partial \alpha} \end{bmatrix}^T \begin{bmatrix} V_{i11} & 0 \\ 0 & V_{i22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{Y}_i - \mu_i(\beta) \\ S_i - \sigma_i(\alpha,\beta) \end{bmatrix} = 0 \tag{4.4}$$

em que V_{i11} segue da expressão (3.2); V_{i22} denota a matriz de covariância correspondente ao seguinte produto de Kronecker: $S_{i(s,j)} = (Y_{is} - \mu_{is}) \otimes (Y_{ij} - \mu_{ij})$. Além disso, $\sigma_i = \mathbb{E}(S_i)$.

A representação matricial em (4.4) é obtida a partir de uma transformação linear que depende da média marginal. Por facilidade computacional, isto é, facilidade na inversão das matrizes V_{i11} e V_{i22}, sugere-se escrever separadamente as equações de estimação em (4.4). Assim, tem-se que:

$$\mathcal{U}_1^{\star}(\beta,\alpha) = \sum_{i=1}^{n} \left[\frac{\partial \mu_i}{\partial \beta}\right]^T V_{i11}^{-1}(\mathbf{Y}_i - \mu_i(\beta)) \tag{4.5}$$

e

$$\mathcal{U}_2^{\star}(\beta,\alpha) = \sum_{i=1}^{n} \left[\frac{\partial \sigma_i}{\partial \alpha}\right]^T \widehat{V}_{i22}^{-1}(S_i - \sigma_i(\beta,\alpha)) \tag{4.6}$$

A estimação de (β,α) em respostas ordinais é análoga à proposta de Lipsitz et al. (1991). Contudo, a diferença para respostas ordinais é que a matriz de covariância para o vetor de respostas $\mathbf{Y}_i$ possui uma estrutura bloco-diagonal, com a covariância para cada Y_{ij} determinada pela média μ_{ij}. Assim, o seguinte algoritmo sumariza o processo de estimação para as EEG's de 1^a ordem no caso de respostas ordinais:

1. Obter as estimativas iniciais para $(\widehat{\beta}^{(0)}, \widehat{\alpha}^{(0)})$. Em geral, assume-se que $\alpha^{(0)} = 0$, e estima-se $\widehat{\beta}^{(0)}$ a partir de um modelo de chances proporcionais para observações independentes;

2. Usando o processo iterativo de Gauss-Seidel, obtêm-se as atualizações para β e α,

e o processo continua até que se obtenha a convergência, tal que:

$$\widehat{\beta}^{(m+1)} = \widehat{\beta}^{(m)} + \left(\sum_{i=1}^{n}\left[\frac{\partial\mu_i}{\partial\beta}\right]^T V_{i11}^{-1}\left[\frac{\partial\mu_i}{\partial\beta}\right]\right)^{-1}\left(\sum_{i=1}^{n}\mathcal{U}_1^{\star}(\beta^{(m)},\alpha^{(m)})\right)$$

$$\widehat{\alpha}^{(m+1)} = \widehat{\alpha}^{(m)} + \left(\sum_{i=1}^{n}\left[\frac{\partial\sigma_i}{\partial\alpha}\right]^T V_{i22}^{-1}\left[\frac{\partial\sigma_i}{\partial\alpha}\right]\right)^{-1}\left(\sum_{i=1}^{n}\mathcal{U}_2^{\star}(\beta^{(m)},\alpha^{(m)})\right)$$

É válido lembrar que para as EEG's de 1^a ordem o parâmetro α é tratado como uma perturbação, e nos artigos publicados por Liang e Zeger (1986a; 1986b), o mesmo foi inicialmente estimado pelo método dos momentos. Na abordagem marginal com respostas ordinais, o coefiente de regressão descreve como a *odds* da resposta pode aumentar (ou diminuir) a cada unidade que se aumenta na covariável, se a mesma for contínua. E, no caso de variáveis preditoras categóricas, a *odds* da resposta pode aumentar (ou diminuir) a depender do grupo de comparação.

Como já foi dito anteriormente, Carey et al. (1993) propuseram a ALR, e estimaram o parâmetro α usando a razão de chances como medida de associação. Heagerty e Zeger (1996) estenderam a metodologia de Carey et al. (1993) para acomodar respostas ordinais, e utilizaram a razão de chances global como medida de associação. Assim, o segundo conjunto de equações de estimação para o parâmetro α é dado por:

$$\mathcal{U}_2^{\star\star}(\beta,\alpha) = \sum_{i=1}^{n}\left[\frac{\partial\zeta_i}{\partial\alpha}\right]^T M_i^{-1}(Y_i^{\star} - \zeta_i) = 0 \tag{4.7}$$

onde M_i e ζ_i são como descritos em (3.8), mas $\zeta_{ijs(k_1,k_2)} = \mathbb{E}(Y_{is,k_1}|Y_{ij,k_2})$. Além disso, $Y_i^{\star} - \zeta_i$ representam os resíduos condicionais formados por todos os diferentes pares de respostas ordinais, sendo $Y_i^{\star}$ definido como:

$$Y_i^{\star} = ((Y_{i1}\otimes\mathbf{1}_K)^T, (Y_{i2}\otimes\mathbf{1}_K)^T, \ldots, (Y_{i(n_i-1)}\otimes\mathbf{1}_K)^T)^T$$

O produto de Kronecker é usado com um vetor de 1's para representar o fato de que o elemento do vetor Y_{ij} é repetido K vezes.

A equação de estimação em (4.7) nada mais é do que uma regressão de $Y_i^{\star}$ sob $Y_i^{\star\star}$, onde as esperanças condicionais são dadas por $\zeta_i = \mathbb{E}(Y_i^{\star}|Y_i^{\star\star})$, sendo

$$Y_i^{\star\star} = ((\mathbf{1}_K\otimes Y_{i2})^T, (\mathbf{1}_K\otimes Y_{i3})^T, \ldots, (\mathbf{1}_K\otimes Y_{in_i})^T)^T$$

A partir da ALR é possível estimar a razão de chances global para o par (Y_{is}, Y_{ij}) nos pontos de corte (k_1, k_2), por exemplo, usando as variáveis indicadoras U_{ijk}. O algoritmo

de estimação para ALR é similar ao anteriormente descrito para as EEG's de 1^a ordem, alternando-se entre os passos para atualizar α e β. O β estimado via ALR possui a mesma robustez, consistência e normalidade assintótica do estimador obtido usando as EEG's de 1^a ordem. A diferença entre as duas metodologias refere-se à eficiência do parâmetro α (Heagerty e Zeger, 1996).

4.2 Modelos mistos para respostas ordinais

O modelo de efeitos mistos para as probabilidades acumuladas pode ser descrito, como em Hedeker e Mermelstein (2000), em termos do seguinte logito:

$$\log\left[\frac{\mathbb{P}(Y_{ij} \leq k|b_i)}{1-\mathbb{P}(Y_{ij} \leq k|b_i)}\right] = \gamma_k + X_{ij}^T\beta + \widetilde{X}_{ij}^T\varrho_k + Z_i^T b_i, \quad k = 1, 2, \ldots, K-1, \tag{4.8}$$

em que X_i representa a matriz de variáveis preditoras para os efeitos fixos, Z_i uma matriz desenho para os efeitos aleatórios, sendo um subconjunto de X_{ij}, e $\widetilde{X}_{ij}$ é um subconjunto de variáveis preditoras da matriz X_{ij} para os quais o efeito do coeficiente ϱ varia em k.

Se o efeito ϱ_k em (4.8) for igual a zero, a expressão acima se reduz à de um modelo de efeitos mistos descrito a partir do modelo de *odds* proporcionais, ou seja:

$$\log\left[\frac{\mathbb{P}(Y_{ij} \leq k|b_i)}{1-\mathbb{P}(Y_{ij} \leq k|b_i)}\right] = \gamma_k + X_{ij}^T\beta + Z_i^T b_i. \tag{4.9}$$

Uma especificação geral para o modelo de efeitos mistos em respostas ordinais a partir da expressão (4.9) é dada por:

1. $\eta_{ij} = \gamma_k + X_{ij}^T\beta + Z_i^T b_i$;
2. Condicional ao vetor de efeitos aleatórios b_i, os Y_{ij} são independentes e têm distribuição multinomial;
3. Os efeitos b_i são assumidos seguir distribuição normal bivariada, com média 0 e matriz 2×2 de covariância $\mathbf{G}$ (para um modelo com intercepto e slope (coeficiente angular) aleatórios).

Assim, os logitos acumulados para uma resposta Y_{ij} que possui, por exemplo, três categorias, são expressos por:

$$\log\left[\frac{\mathbb{P}(Y_{ij}\leq 1|b_i)}{1-\mathbb{P}(Y_{ij}\leq 1|b_i)}\right]=\log\left[\frac{\mathbb{P}(Y_{ij}=1)}{\mathbb{P}(Y_{ij}=2\quad \text{ou}\quad 3)}\right]=\gamma_1+X_{ij}^T\beta+Z_i^Tb_i$$

$$\log\left[\frac{\mathbb{P}(Y_{ij}\leq 2|b_i)}{1-\mathbb{P}(Y_{ij}\leq 2|b_i)}\right]=\log\left[\frac{\mathbb{P}(Y_{ij}=1\quad \text{ou}\quad 2)}{\mathbb{P}(Y_{ij}=3)}\right]=\gamma_2+X_{ij}^T\beta+Z_i^Tb_i$$

Como os coeficientes da regressão não dependem do índice k, assume-se então que o efeito de certa covariável é o mesmo entre os $K-1$ logitos. Desta maneira, a chance de uma resposta em uma categoria maior que K, para algum K fixo, é dada por $\exp(\beta)$ para cada 1 unidade de mudança em certa covariável X_i. Assim, para uma resposta ordinal contendo 3 categorias, o modelo descreve simultaneamente o efeito de X_i sob todas as $K-1$ comparações entre as probabilidades (Hedeker e Mermelstein, 2000).

Outra maneira de representar o modelo descrito em (4.9) é usando uma representação multinível (Hedeker e Gibbons, 1994; 2006). Para tal, o referido modelo é particionado em 2 níveis. Assim, o modelo para o 1^o nível é dado pela seguinte expressão:

$$\log\left[\frac{\mathbb{P}(Y_{ij}\leq k|b_i)}{1-\mathbb{P}(Y_{ij}\leq k|b_i)}\right]=\gamma_k+X_{ij}^T\beta+Z_i^Tb_i.$$

O 2^o nível do modelo avalia a variação entre indivíduos, sendo expresso por:

$$b_i=\mu+\ddot{X}_i^T\tilde{\beta}+\delta_i.$$

Nesse nível os b_i's são influenciados pelo efeito global μ e por $\ddot{X}_i$, onde é um subconjunto de X_{ij}. Além disso, o componente aleatório δ_i é normalmente distribuído com média 0 e matriz de variância e covariância G_b.

Como anteriormente mencionado, a intepretação dos coeficientes de regressão nessa classe de modelos é a nível do indivíduo, e a inclusão de efeitos aleatórios no modelo tem implicações na interpretação dos parâmetros. Nesse sentido, se houver uma covariável categórica no modelo, o coeficiente de regressão associado a ela descreverá como a *odds* de uma dada categoria da resposta aumenta (ou diminui), dados dois indivíduos (cada um de um grupo) com o mesmo efeito aleatório. Nos casos em que a covariável é contínua, a *odds* da resposta para algum indivíduo aumenta (ou diminui) a cada uma unidade que se acrescenta na referida covariável.

Se combinarmos os dois níveis do modelo, teremos a seguinte expressão:

$$\log\left[\frac{\mathbb{P}(Y_{ij}\leq k|\delta_i)}{1-\mathbb{P}(Y_{ij}\leq k|\delta_i)}\right]=\gamma_k+X_{ij}^T\beta+Z_i^T\left(\mu+\ddot{X}_i^T\tilde{\beta}+\delta_i\right)$$

A representação multinível mostra que, assim como as variáveis preditoras X_{ij}

são incluídas no primeiro nível modelo para explicar a variação da resposta a nível do indivíduo, as variáveis preditoras $\ddot{X}_i$ são incluídas no segundo nível para explicar a variação de b_i entre indivíduos (Hedeker e Gibbons, 1994).

Dadas as três especificações do modelo misto para a resposta ordinal, tem-se que a probabilidade conjunta para Y_i e b_i pode ser expressa como:

$$f(Y_i, b_i) = f(Y_i|b_i)f(b_i)$$

Assim, a função de verossimilhança é expressa por:

$$\begin{aligned} L(\beta, \phi, G) &= \prod_{i=1}^{n} \int_{b_i} f(Y_i|b_i)f(b_i)db_i \\ &= \prod_{i=1}^{n} \int_{b_i} \left[\prod_{k=1}^{K} (\mathbb{P}(Y_i \leq k|b_i) - \mathbb{P}(Y_i \leq k-1|b_i))^{y_{ik}} \right] f(b_i)db_i \end{aligned} \tag{4.10}$$

Os parâmetros do modelo misto são estimados maximizando-se a função em (4.10), sendo necessários para tal a utilização de métodos de aproximação numérica, como Laplace ou Quadratura gaussiana, pois tal expressão não possui forma fechada. É também possível fazer uso do método de quase-verossimilhança penalizada (PQL). É válido ressaltar que em algumas situações o método PQL é preferível aos métodos de aproximação numérica devido ao tempo computacional ser muito maior nos métodos numéricos.

4.3 Modelos de transição para respostas ordinais

A especificação do modelo de transição para respostas ordinais longitudinais é feita de forma análoga às abordagens marginal e mista. As probabilidades acumuladas da resposta podem ser definidas de forma geral a partir do seguinte modelo de *odds* proporcionais parciais:

$$\begin{aligned} \log \left[\frac{\mathbb{P}(Y_{ij} \leq k|X_i, \widetilde{X}_i, \mathcal{H}_{ij})}{1 + \mathbb{P}(Y_{ij} \leq k|X_i, \widetilde{X}_i, \mathcal{H}_{ij})} \right] &= \gamma_k + X_i^T\beta + \widetilde{X}_i^T\varrho_{Qk} + (\alpha_1 y_{ij-1} + \alpha_2 y_{ij-2} + \ldots + \alpha_q y_{ij-q}) \\ &= \gamma_k + X_i^T\beta + \widetilde{X}_i^T\varrho_{Qk} + \sum_{q=1}^{Q} \alpha_q y_{ij-q} \end{aligned} \tag{4.11}$$

Como descrito na subseção 3.1.3, os modelos de transição são uma classe de modelos que expressam de forma explícita a influência de respostas passadas sob a resposta atual, modelando a esperança condicional da variável resposta como função de respostas passadas e variáveis preditoras.

Caso o efeito ϱ na expressão (4.11) seja igual a zero, estamos assumindo que as *odds* em quaisquer categorias da resposta são proporcionais. Assim, para modelarmos a dependência de Y_{ij} sob $Y_{i1}, \ldots, Y_{ij-1}$, predizendo-a a partir das respostas passadas, especificamos o seguinte modelo de *odds* proporcionais:

$$\log\left[\frac{\mathbb{P}(Y_{ij} \leq k|X_i, \mathcal{H}_{ij})}{1+\mathbb{P}(Y_{ij} \leq k|X_i, \mathcal{H}_{ij})}\right] = \gamma_k + X_i^T\beta + (\alpha_1 y_{ij-1} + \alpha_2 y_{ij-2} + \ldots + \alpha_q y_{ij-q})$$

$$= \gamma_k + X_i^T\beta + \sum_{q=1}^{Q} \alpha_q y_{ij-q} \tag{4.12}$$

Assim, para Y_{ij} ordinal, a verossimilhança condicional é dada por:

$$L(Y_i;\beta,\alpha) = \prod_{i=1}^{n}\prod_{j=2}^{n_i} f(Y_i|\mathcal{H}_{ij}, X_i) \tag{4.13}$$

$$= \prod_{i=1}^{n}\prod_{j=2}^{n_i}\left\{\prod_{k=1}^{K}\left[\mathbb{P}(Y_{ij} \leq k|\mathcal{H}_{ij}, X_i) - \mathbb{P}(Y_{ij} \leq k-1|\mathcal{H}_{ij}, X_i)\right]^{y_{ijk}}\right\}$$

$$= \prod_{i=1}^{n}\prod_{j=2}^{n_i}\left\{\prod_{k=1}^{K}\left[\frac{e^{\gamma_k + X_i^T\beta + \sum_{q=1}^{Q}\alpha_q y_{ij-q}}}{1+e^{\gamma_k + X_i^T\beta + \sum_{q=1}^{Q}\alpha_q y_{ij-q}}} - \frac{e^{\gamma_{k-1} + X_i^T\beta + \sum_{q=1}^{Q}\alpha_q y_{ij-q}}}{1+e^{\gamma_{k-1} + X_i^T\beta + \sum_{q=1}^{Q}\alpha_q y_{ij-q}}}\right]^{y_{ijk}}\right\}$$

Nesse modelo os parâmetros β e α_q são estimados usando a função de verossimilhança condicional, pois como descrito em 3.1.3, a verossimilhança das Q primeiras observações não é diretamente especificada.

Nesse capítulo três classes de modelos para dados longitudinais foram utilizadas para modelar respostas politômicas ordinais. A escolha da metodologia mais adequada para análise dependerá do objetivo do estudo. Na modelagem marginal a interpretação é populacional, sendo seu uso indicado quando o objetivo do estudo é inferir sobre a população média. A modelagem marginal é uma das metodologias mais populares, em decorrência, talvez, da facilidade na interpretação dos parâmetros estimados. Quando o foco do estudo é o indivíduo, o modelo linear generalizado mistos é o mais indicado. Nessa

classe de modelos a heterogeneidade intra-indivíduos é modelada através da inclusão de um efeito aleatório no modelo, que induz a covariância entre as respostas repetidas. A intepretação pode ser marginal ou condicional, sendo mais usual condicionar no efeito aleatório. Se o objetivo do estudo for fazer predição da resposta, o modelo de transição será o mais adequado. Nessa classe de modelos as respostas passadas entram no modelo como preditores adicionais, modelando, assim, a correlação entre as respostas.

Capítulo 5

Exemplos numéricos

Neste capítulo os modelos longitudinais apresentados anteriormente serão utilizados em duas aplicações com dados reais. Apresentamos na Seção 5.1 os resultados de um estudo epidemiológico cujo objetivo foi avaliar o impacto da implantação dos sistemas de captação de águas da chuva na saúde das crianças residentes em 2 municípios do médio vale do Jequitinhonha em Minas Gerais. Na Seção 5.2 discutimos os resultados de um estudo sobre analgesia no parto, desenvolvido pela Faculdade de Medicina da UFMG e pelo Hospital Odilon Berhrens, com objetivo de comparar duas técnicas de analgesia para a dor no trabalho de parto.

5.1 Exemplo 1: Estudo sobre sistemas de captação de chuva na saúde da criança

Os dados nesta aplicação provêm de um estudo epidemiológico, do tipo coorte prospectiva, desenvolvido com 664 crianças de até 5 anos, selecionadas e acompanhadas pelo período de um ano (2009 a 2010). O objetivo principal do estudo foi avaliar o impacto da implantação dos sistemas de captação de água da chuva, construídas em sua maioria pelo P1MC (Programa 1 milhão de cisternas), na saúde das crianças de famílias rurais, residentes em 2 municípios (Chapada do Norte e Berilo) do Médio Vale do Jequitinhonha em Minas Gerais. A seleção das crianças participantes foi realizada com base em uma amostragem não probabilística, sendo tal decisão tomada devido à existência de um pequeno número de crianças que atendiam aos critérios para integrar o grupo exposto à intervenção. Na primeira etapa do estudo metade das crianças analisadas possuíam acesso às cisternas para armazenamento de água da chuva (Grupo 1), e a outra metade dependia de outras fontes de água (mananciais sem proteção sanitária, como rio, barragem, mina e cacimba), ou seja, não possuíam o sistema de captação de água em cisternas (Grupo

2). Das 664 crianças amostradas, 248 eram do município de Berilo, e 416 de Chapada do Norte. No estudo, dois indicadores de saúde foram analisados: a ocorrência de diarréia e a presença de parasitas intestinais (protozoários comensais: Endolimax nana, Entamoeba coli e Iodamoeba butschlii; protozoários patogênicos: Entamoeba histolytica/dispar e Giardia lamblia; helmintos: ancilostomídeos, Ascaris lumbricoides, Hymenolepis nana, Enterobius vermicularis, Strongyloides stercoralis e Trichuris trichiura) investigados em 3 etapas longitudinais do estudo.

Neste trabalho a resposta de interesse é a carga parasitária (1: criança poliinfectada; 2: criança monoinfectada; 3: criança não infectada), avaliada via exame de fezes, para a i-ésima criança, na etapa j, $j = 1, 2, 3$ do estudo ($j = 1$: início do estudo; $j = 2$: 6 meses após o início do estudo; $j = 3$: 1 ano após o início do estudo). As variáveis preditoras consideradas para ajuste dos modelos são: grupo (1: com cisterna; 0: sem cisterna), frequência de banho nas crianças (1: >1 vez ao dia; 0: uma vez ao dia), sexo (0: Feminino; 1: Masculino) e idade das crianças em meses.

A Tabela 5.1 apresenta uma descrição das variáveis preditoras em relação à variável de grupo no início do estudo. Nos dois grupos a proporção maior é de crianças do sexo feminino, que tomam mais do que um banho por dia, e cuja idade é superior a doze meses. Na Tabela 5.2 apresentamos uma descrição das variáveis preditoras na primeira etapa do estudo em relação à carga parasitária da criança. Observa-se que a proporção de crianças poliinfectadas no grupo com cisterna é maior do que no grupo sem cisterna. 7,2% das crianças que tomam mais do que um banho por dia são poliinfectadas, e das crianças que possuem um ano ou mais de vida, apenas 7,6% são poliinfectadas. É válido ressaltar que apesar do estudo ter sido conduzido com 664 crianças, o número de informações varia entre as variáveis consideradas no estudo, em decorrência de perda de informação.

Tabela 5.1: Descrição das variáveis preditoras por grupo no início do estudo

variáveis	n	grupo	
		com cisterna	sem cisterna
		n (%)	n (%)
sexo			
M	349	171 (51,5)	178 (53,6)
F	315	161 (48,5)	154 (46,4)
frequência de banho			
1 x ao dia	103	64 (19,3)	39 (11,8)
> 1 x ao dia	559	268 (80,7)	291 (88,2)
idade da criança			
< 12 meses	120	56 (16,9)	64 (19,3)
≥ 12 meses	542	275 (83,1)	267 (80,6)

Tabela 5.2: Descrição das variáveis preditoras no início do estudo em relação à carga parasitária da criança

variáveis	n	carga parasitária da criança		
		não infectada	monoinfectada	poliinfectada
		n (%)	n (%)	n (%)
sexo				
M	313	241 (77,0)	52 (16,6)	20 (6,4)
F	263	191 (72,6)	54 (20,6)	18 (6,8)
frequência de banho				
1 x ao dia	89	70 (78,6)	16 (18,0)	3 (3,4)
$>$ 1 x ao dia	485	360 (74,2)	90 (18,6)	35 (7,2)
idade da criança				
$<$ 12 meses	76	67 (88,2)	9 (11,8)	0 (0,0)
$\geq$ 12 meses	498	365 (73,3)	95 (19,1)	38 (7,6)
grupo				
com cisterna	292	224 (76,7)	46 (15,8)	22 (7,5)
sem cisterna	284	208 (73,3)	60 (21,1)	16 (5,6)

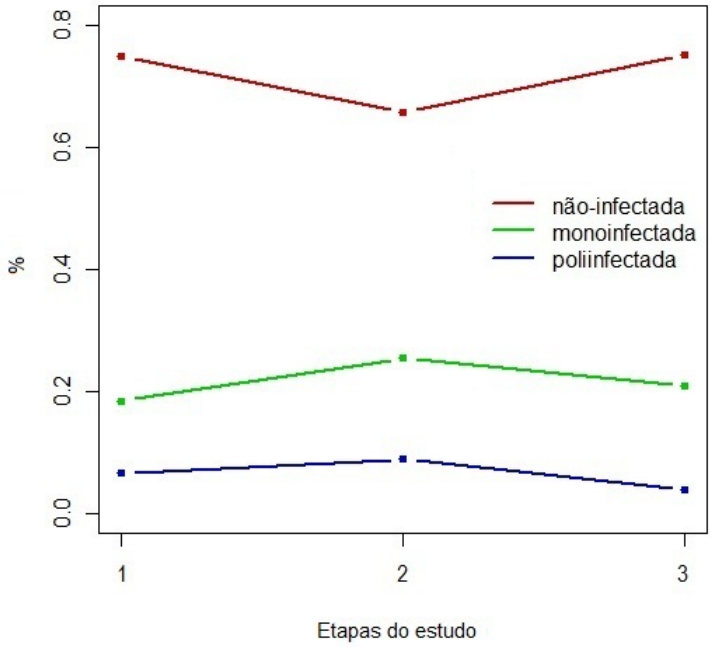

Figura 5.1: Perfil da carga parasitária da criança em cada etapa do estudo.

A Figura 5.1 apresenta o perfil médio das crianças quanto à carga parasitária em cada etapa do estudo. Nela observamos que a proporção de crianças poliinfectadas no início do estudo é de aproximadamente 7%, aumentando para cerca de 10% na etapa seguinte, e voltando a cair no final do estudo. Esse mesmo comportamento pode ser observado em relação à monoinfecção. Na Figura 5.2 há uma descrição da proporção de crianças

poliinfectadas, bem como poli ou monoinfectadas em cada grupo nas 3 etapas do estudo. É possível observar que o grupo com cisterna apresenta uma maior proporção de crianças poliinfectadas em todas as etapas do estudo (Figura 5.2 (a)). Ao avaliar a Figura 5.2 (b) nota-se que a proporção de crianças infectadas (com mono ou poliinfeção) é ligeiramente maior no grupo sem cisterna nas duas primeiras etapas do estudo, e apenas na etapa final essa proporção é maior no grupo com cisterna (maiores detalhes em Fonseca, 2012).

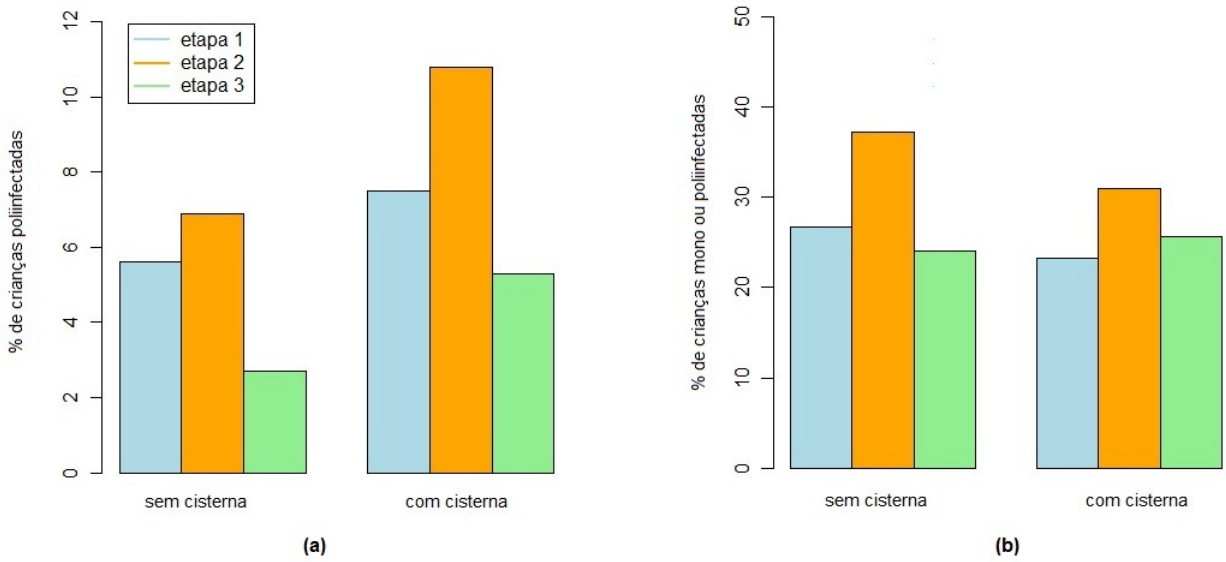

Figura 5.2: Descrição da proporção de crianças (a) poliinfectadas (b) poli ou monoinfectadas em cada grupo nas 3 etapas do estudo.

Para ajuste do modelo marginal com respostas ordinais (Seção 4.1) foram utilizadas as funções `ordegee`$(\cdot)$ da biblioteca `geepack` e `repolr`$(\cdot)$ da biblioteca `repolr`. Tais funções utilizam a versão da regressão logística alternada (ALR), estendida por Heagerty e Zeger (1996), para estimação dos parâmetros. Além disso, esta última função testa, através da estatística escore, se o pressuposto do modelo de *odds* proporcionais foi violado. Diferentes estruturas (simétrica composta, não estruturada, independente) foram utililizadas para a matriz de trabalho. Para o ajuste do GLMM (Seção 4.2) foram utilizadas as funções `clmm`$(\cdot)$ e `clmm2`$(\cdot)$ da biblioteca `ordinal`. O método de estimação utilizado foi o de máxima verossimilhança via quadratura gaussiana adaptativa, considerando-se 50 pontos para a quadratura. Para o GLMM foram ajustados 2 modelos: em um deles considerou-se apenas o intercepto como aleatório, e no outro intercepto e *slope* (coeficiente angular) para a covariável etapa foram considerados aleatórios. Para avaliar a significância da variância do efeito aleatório em ambos os modelos foi realizado o teste da razão

de verossimilhança (TRV), calculando-se também o coeficiente de correlação intra-classe (apenas para o modelo com o intercepto aleatório). Para o modelo de transição (Seção 4.3) foram ajustados dois modelos de 1ª ordem, onde em um deles considerou-se a interação entre a resposta no tempo $j-1$ e a variável grupo, e no outro um modelo sem interação. O modelo de transição foi ajustado usando a função `vglm`$(\cdot)$ da biblioteca `VGAM`. Todos os ajustes foram realizados no *software* R-2.15 (R, 2012).

Para os modelos marginal, misto (dados os efeitos aleatórios) e de transição, respectivamente, assumiu-se que o log da *odds* para uma das categorias da resposta ordinal em cada etapa do estudo segue os seguintes modelos de *odds* proporcionais:

- Modelo marginal

$$\log\left[\frac{\mathbb{P}(Y_{ij}\leq k|X_{ij})}{1-\mathbb{P}(Y_{ij}\leq k|X_{ij})}\right]=\gamma_k+\beta_1\text{grupo}_{ij}+\beta_2\text{sexo}_i+\beta_3\text{freq.banho}_i+\beta_4\text{etapa}_j+\beta_5\text{idade}_i$$

- GLMM

$$\log\left[\frac{\mathbb{P}(Y_{ij}\leq k|b_i)}{1-\mathbb{P}(Y_{ij}\leq k|bi)}\right]=\gamma_k+\beta_1\text{grupo}_{ij}+\beta_2\text{sexo}_i+\beta_3\text{freq.banho}_i+\beta_4\text{etapa}_j+\beta_5\text{idade}_i+b_i$$

$$\log\left[\frac{\mathbb{P}(Y_{ij}\leq k|b_i)}{1-\mathbb{P}(Y_{ij}\leq k|bi)}\right]=\gamma_k+\beta_1\text{grupo}_{ij}+\beta_2\text{sexo}_i+\beta_3\text{freq.banho}_i+\beta_4\text{etapa}_j+\beta_5\text{idade}_i+b_{1i}+b_{2i}\text{etapa}_j$$

- Modelo de transição

$$\log\left[\frac{\mathbb{P}(Y_{ij}\leq k|\mathcal{H}_{ij},X_{ij})}{1-\mathbb{P}(Y_{ij}\leq k|\mathcal{H}_{ij},X_{ij})}\right]=\gamma_k+\beta_1\text{grupo}_{ij}+\beta_2\text{sexo}_i+\beta_3\text{freq.banho}_i+\beta_4\text{etapa}_j+\beta_5\text{idade}_i+$$
$$+\alpha_1 y_{ij-1}+\alpha_2 y_{ij-1}\times\text{grupo}_{ij}$$

$$\log\left[\frac{\mathbb{P}(Y_{ij}\leq k|\mathcal{H}_{ij},X_{ij})}{1-\mathbb{P}(Y_{ij}\leq k|\mathcal{H}_{ij},X_{ij})}\right]=\gamma_k+\beta_1\text{grupo}_{ij}+\beta_2\text{sexo}_i+\beta_3\text{freq.banho}_i+\beta_4\text{etapa}_j+\beta_5\text{idade}_i+$$
$$+\alpha y_{ij-1}$$

Na abordagem marginal foram ajustados 3 modelos que diferem apenas em relação à estrutura da matriz de trabalho $R(\alpha)$ (assumimos matrizes do tipo independente, simétrica composta e não-estruturada). É possível observar na Tabela 5.3 que não há diferença nas estimativas dos parâmetros entre os modelos. Como mencionado na subseção

3.1.1, mesmo sob má especificação da matriz $R(\alpha)$, as estimativas para β são consistentes. Um critério para escolha da melhor estrutura é o QIC, contudo, para a função até então disponível no *software* R para ajuste do modelo marginal ordinal, esta medida não pode ser calculada. Um critério alternativo para escolha da referida estrutura seria avaliar a significância estatística do parâmetro α, que mede a associação entre as medidas repetidas. No modelo 1 assumiu-se uma estrutura do tipo independente para $R(\alpha)$, logo $R(\alpha) = I$. Para os modelos 2 e 3 nota-se que o componente α não foi estatisticamente significativo. Assim, por parcimônia, o modelo 2 poderia ser o escolhido, visto que apenas 1 parâmetro é estimado para a estrutura de associação. O pressuposto do modelo de *odds* proporcionais foi avaliado, e o teste apontou para sua violação (p-valor=0,029). As variáveis preditoras que violaram o pressuposto foram grupo, idade e etapa. Assim sendo, o mais adequado é considerar, por exemplo, o modelo de *odds* proporcionais parciais para ajuste, onde o efeito das variáveis preditoras que violaram o pressuposto de proporcionalidade das *odds* variará de acordo com a categoria da resposta. O ajuste deste modelo marginal é apenas possível no *software* SAS (SAS Institute Inc., 2002). Os resultados obtidos para o modelo de *odds* proporcionais parciais encontram-se na Tabela 5.4. O modelo final considera uma estrutura do tipo simétrica composta para a matriz de trabalho. É válido ressaltar que também foi ajustado um mode-lo considerando a matriz de trabalho do tipo não estruturada, contudo, os componentes estimados para a estrutura de associação não foram estatisticamente significativos. Assim, por parcimônia, optamos pela estrutura do tipo simétrica composta.

Tabela 5.3: Estimativas dos parâmetros para o modelo marginal no estudo sobre sistemas de captação de chuva na saúde da criança.

	$R(\alpha)$ independente			$R(\alpha)$ simétrica composta			$R(\alpha)$ não estruturada		
variável	$\widehat{\beta}$	ep*	p-valor	$\widehat{\beta}$	ep	p-valor	$\widehat{\beta}$	ep	p-valor
γ_1	-3,825	0,456	-	-3,828	0,456	-	-3,809	0,456	-
γ_2	-2,066	0,445	-	-2,069	0,444	-	-2,049	0,443	-
grupo (com cisterna)	0,352	0,187	0,064	0,353	0,186	0,059	0,353	0,188	0,065
sexo (M)	0,062	0,186	0,739	0,060	0,186	0,746	0,058	0,186	0,751
freq.banho (> 1x ao dia)	0,282	0,286	0,324	0,279	0,286	0,329	0,284	0,286	0,321
idade	0,031	0,007	<0,001	0,031	0,006	<0,001	0,031	0,007	<0,001
etapa	-0,135	0,098	0,172	-0,133	0,098	0,177	-0,146	0,098	0,137
$\widehat{\alpha}$**	-			0,291	0,398	0,465	-0,309	0,520	0,553
							0,438	0,435	0,314
							0,845	0,557	0,129

*erro-padrão baseado na variância do estimador sanduiche
**associação intra-indivíduo baseada no log da *odds* ratio

A partir dos resultados da Tabela 5.4 observamos que a chance de poliinfecção no estudo é 1,7 ($e^{0,535}$) vezes maior entre crianças que possuem cisterna quando comparadas àquelas que não possuem. Observamos ainda que há um aumento na chance de poliinfecção a cada um mês de acréscimo na idade da criança ($e^{0,037} = 1,04$). O mesmo ocorre

Tabela 5.4: Estimativas dos parâmetros para o modelo marginal via *odds* proporcionais parciais no estudo sobre sistemas de captação de chuva na saúde da criança.

variável	$\widehat{\beta}$	ep*	p-valor
γ_1	-2,083	0,336	-
γ_2	-1,923	0,258	-
grupo_1 (com cisterna)	0,535	0,191	0,002
grupo_2	-0,059	0,118	0,621
sexo (M)	0,056	0,118	0,633
freq.banho (> 1x ao dia)	0,248	0,182	0,172
idade_1	0,037	0,007	0,037
idade_2	0,023	0,004	<0,001
etapa_1	-0,315	0,112	0,101
etapa_2	-0,131	0,066	0,049
$\widehat{\alpha}$	0,776	0,130	<0,001

em relação à chance de mono ou poliinfecção ($e^{0,023} = 1,02$). Já em relação à variável etapa observa-se que há uma redução na chance de mono ou poliinfecção ($e^{-0,131} = 0,88$) no decorrer do estudo. Nesse modelo a associação entre as etapas do estudo foi estatisticamente significativa, indicando que a chance de uma criança com mono ou poliinfecção em uma etapa permanecer em tal condição na etapa seguinte é 2,2 ($e^{0,776}$). Essa mesma chance também vale para uma criança não infectada numa etapa, permanecer em tal condição na etapa seguinte do estudo.

Tabela 5.5: Estimativas dos parâmetros para o modelo misto no estudo sobre sistemas de captação de chuva na saúde da criança.

	GLM			GLMM- intercepto aleatório			GLMM- intercepto e slope aleatório		
variável	est	ep	p-valor	est	ep	p-valor	est	ep	p-valor
γ_1	-3,495	0,244	-	-3,551	0,297	-	-3,534	0,433	-
γ_2	-1,761	0,225	-	-1,808	0,275	-	-1,781	0,417	-
grupo (com cisterna)	-0,057	0,116	0,622	-0,054	0,118	0,648	-0,060	0,118	0,610
sexo (M)	-0,031	0,117	0,791	-0,030	0,117	0,797	-0,032	0,118	0,787
freq.banho (>1x ao dia)	0,311	0,166	0,060	0,313	0,168	0,062	0,313	0,168	0,063
idade	0,021	0,004	<0,001	0,021	0,004	<0,001	0,021	0,004	<0,001
Etapa				0,014	0,069	0,836	0,001	0,161	0,983
$g_{11} = Var(b_{1i})$				0,038			0,040		
$g_{22} = Var(b_{2i})$							0,042		
TRV*						0,47			0,01
ICC[1]				0,011					
AIC	2204,7			2208,6			2203,2		

est: estimativa dos parâmetros; ep: erro-padrão

* sob H_0 $\sigma^2_{b_{li}} = 0, \quad l = 1, 2$ (indica se o efeito é no intercepto ou slope)

[1] ICC avaliado em b_{1i}

No modelo misto (Tabela 5.5) observa-se não haver diferença entre os modelos ajustados, e isso pode decorrer do fato das crianças terem recebido medicação para combater a infecção no início do estudo. Ao compararmos o modelo sob independência com o GLMM assumindo apenas um intercepto aleatório, notamos, pelo TRV, que não haveria

necessidade de ajustar um modelo com efeito aleatório. Apesar do TRV apontar para uma não-rejeição da hipótese nula (p-valor=0,47), mantivemos o ajuste do modelo misto com um efeito aleatório, e dos três modelos ajustados, optamos pelo último, que considera dois efeitos aleatórios, em decorrência do TRV ter apontado para uma rejeição (p-valor=0,01) da hipótese nula, indicando, assim, a importância de manter o coeficiente angular (*slope*) no modelo. É válido ressaltar que a distribuição usual da estatística de teste é qui-quadrado com 1 grau de liberdade, contudo, utilizamos uma mistura de qui-quadrados $\left(\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2\right)$ como distribuição para a referida estatística, em decorrência da hipótese testada estar na fronteira do espaço paramétrico. Para o modelo final (modelo com 2 efeitos aleatórios) foi testado o pressuposto de proporcionalidade das *odds*, e o mesmo foi rejeitado (p-valor=0,003). Assim, o mais adequado é considerar um modelo de *odds* proporcionais parciais, onde o efeito para as variáveis preditoras grupo, etapa e idade, identificadas no teste como as que violaram o pressuposto, variará de acordo com a categoria da resposta. As demais covariáveis terão seu efeito constante. É válido ressaltar que o modelo final de odds proporcionais parciais considera que há efeito aleatório no intercepto e no *slope*.

Tabela 5.6: Estimativas dos parâmetros para o modelo misto via *odds* proporcionais parciais no estudo sobre sistemas de captação de chuva na saúde da criança.

variável	estimativa	ep
γ_1	-3,864	0,445
γ_2	-1,772	0,275
Grupo_1(com cisterna)	0,521	0,214*
Grupo_2(com cisterna)	-0,110	0,118
sexo (M)	-0,029	0,116
freq.banho (>1x ao dia)	0,304	0,167
idade.criança_1	0,033	0,008*
idade.criança_2	0,020	0,004*
Etapa_1	-0,173	0,134
Etapa_2	0,031	0,070
$g_{11} = \text{Var}(b_{1i})$	0,022	
$g_{22} = \text{Var}(b_{2i})$	0,042	

* p-valor$<$ 0,05

A partir do ajuste do modelo misto considerando *odds* proporcionais parciais (Tabela 5.6), observamos que a chance de poliinfecção aumenta a cada um mês acrescido na idade da criança ($e^{0,033} = 1,04$). O mesmo ocorre com a chance de mono ou poliinfecção ($e^{0,020} = 1,03$).

Por fim, o modelo de transição foi ajustado para predizer o risco de infecção

Tabela 5.7: Estimativas dos parâmetros para o modelo de transição no estudo sobre sistemas de captação de chuva na saúde da criança.

	modelo com interação			modelo sem interação		
variável	est	ep	p-valor	est	ep	p-valor
γ_1	-2,243	0,462	-	-2,726	0,384	-
γ_2	-0,593	0,445	-	-0,887	0,365	-
grupo (com cisterna)	-0,510	0,553	0,356	0,061	0,156	0,696
sexo (M)	0,077	0,156	0,622	0,078	0,156	0,616
freq.banho (>1x ao dia)	-0,214	0,221	0,333	-0,226	0,220	0,305
idade	0,013	0,006	0,023	0,013	0,006	0,022
Etapa 3	-0,555	0,159	<0,001	-0,553	0,158	<0,001
Y_{j-1}						
não- infectada	-0,426	0,411	0,299	-0,128	0,293	0,662
monoinfectada	-0,505	0,472	0,284	-0,095	0,312	0,767
Y_{j-1} por Grupo						
não- infectada	0,580	0,585	0,321			
monoinfectada	0,756	0,648	0,243			
AIC	1225.9			1223.4		

est: estimativa dos parâmetros
ep: erro-padrão

(mono ou poliinfecção) na criança controlando pela idade, grupo, frequência de banho e etapa. A Tabela 5.7 traz os resultados de dois ajustes: no primeiro deles foi avaliada a interação entre a resposta na etapa $j-1$ e a variável de grupo, que dentre todas consideradas no ajuste, é a única que muda de uma etapa para outra. O segundo modelo é sem interação. O pressuposto de proporcionalidade das *odds* foi testado e o mesmo não foi violado (p-valor=0,09) no modelo de transição. Assim, é posível observar que a interação não foi estatisticamente significativa, indicando que o efeito da variável grupo entre as etapas é similar e parece não afetar a carga parasitária da criança. Considerando, então, o modelo sem interação, observa-se que as variáveis preditoras idade e etapa foram estatisticamente significativas. A correlação entre as respostas Y_{ij} e Y_{ij-1} não foi estatisticamente significativa, indicando que o efeito de uma etapa a outra do estudo não muda. Analisando a idade, observa-se que a chance de uma criança ser poliinfectada no tempo j aumenta ($e^{0,013} = 1,03$) a cada um mês que acrescentamos na idade da criança. Devido ao pressuposto de proporcinaliade das *odds* a chance de poli ou mono infecção é a mesma. Observamos, ainda, que a chance de poliinfecção na etapa 3 e menor quando comparada à etapa 2 ($e^{-0,553} = 0,57$). O atrativo dos modelos de transição está justamente na possibilidade de predizer o comportamento da resposta, avaliando como a mesma se comporta quando há a transição de uma categoria para a outra no tempo.

Na Tabela 5.8 apresentamos um resumo dos efeitos que foram estatisticamente significativos nos três modelos finais (marginal e misto usando*odds* proporcionais parci-

ais, e transição considerando *odds* proporcionais) considerados nesta aplicação. Apesar dos modelos não serem comparáveis, observamos que os efeitos foram similares nas três abordagens consideradas.

Tabela 5.8: Estimativas dos parâmetros em termos de razões de chance para os modelos finais no estudo sobre sistemas de captação de chuva na saúde da criança.

variável	modelo marginal	modelo misto	modelo de transição
grupo_1 (com cisterna)	$e^{0,535} = 1,70$	$e^{0,521} = 1,68$	-
grupo_2 (com cisterna)	-	-	
sexo (Masc)	-	-	-
freq.banho (> 1x ao dia)	-	-	-
idade_1	$e^{0,037} = 1,04$	$e^{0,033} = 1,03$	$e^{0,013} = 1,01$
idade_2	$e^{0,023} = 1,02$	$e^{0,020} = 1,02$	
etapa_1	-	-	
etapa_2	$e^{-0,131} = 0,88$	-	$e^{-0,553} = 0,57$

OR: odds ratio

Os modelos ajustados nessa aplicação possuem propósitos diferentes na análise de dados longitudinais. A escolha da melhor abordagem irá depender da pergunta de interesse no estudo. Se o objetivo for inferir sobre o comportamento médio da população estudada em relação ao impacto dos sistemas de captação de água da chuva na saúde das crianças, o modelo marginal é o mais adequado. Caso o alvo da pesquisa seja o indivíduo, o modelo misto é o mais indicado. Contudo, se existe interesse em predizer o comportamento futuro da resposta, avaliando o impacto das respostas passadas sob a atual, o modelo indicado será o de transição.

5.2 Exemplo 2: Estudo sobre analgesia no parto

Os dados nesta aplicação provêm de um estudo comparativo conduzido pela Faculdade de Medicina da UFMG e pelo Hospital municipal Odilon Berhens, cujo objetivo foi comparar duas técnicas de analgesia para a dor do trabalho de parto. No estudo 49 pacientes foram acompanhadas por um profissional treinado durante todo o período até o parto. Foi feita uma avaliação em relação à intensidade da dor e medidas como pressão arterial, frequência cardíaca materna, consumo de ocitocina, nível de sedação, sinais de depressão respiratória, apnéia, dentre outras, inicialmente a cada 5 minutos, nos trinta primeiros minutos após o início da anestesia, e após isso a cada 30 minutos até o parto. A primeira técnica utilizada para comparação foi a analgesia epidural, considerada como padrão-ouro, onde um analgésico local é utilizado. A segunda técnica, cuja eficiência será comparada ao padrão-ouro, é a infusão venosa contínua de remifentanil, um opióide que tem início de ação muito rápido (de 1 a 3 minutos).

A intensidade da dor é normalmente dependente do grau de dilatação do colo uterino, sendo, em geral, de leve intensidade, e é do tipo cólica na fase inicial quando a dilatação do colo é menor do que 3 cm, e com a progressão do trabalho de parto os segmentos espinhais adjacentes são estimulados e a dor torna-se mais intensa. A analgesia de parto bloqueia parcial ou completamente os efeitos deletérios da dor e promove conforto à parturiente por controlar de modo efetivo a dor associada às contrações, devendo ser iniciada no momento em que a dor tornar-se incômoda para a parturiente, independente do grau de dilatação do colo uterino e havendo a confirmação do diagnóstico de trabalho de parto (dilatação do colo de 2 a 3 cm com contrações rítmicas na frequência de 3 a 5 cm, intervalo de 10 minutos) (maiores detalhes em Soares, 2013).

Consideraremos como resposta de interesse a intensidade da dor, avaliada através de uma escala visual analógica (EVA) (1: dor leve e tolerável; 2: dor moderada e que causa desconforto; 3: dor intensa e insuportável), tendo sido medida a cada 5 minutos (5,10,15,20,25,30 minutos) após a anestesia, e após isso a cada 30 minutos até o parto (60, 90, 120, ...). O último tempo anotado foi de 360 minutos. No entanto, para as análises apresentadas aqui considerou-se o tempo até 90 minutos para reduzir o número de medidas por paciente. Neste caso j denota os tempos onde a intensidade da dor foi avaliada. Assim, $j = 0, 5, 10, 15, 20, 25, 30, 60, 90$ minutos. As variáveis preditoras consideradas para ajuste dos modelos são: grupo (0:peridural; 1:remifentanil), idade da paciente, doença atual (0:não; 1: sim), uso de medicamentos (0:não; 1: sim), tipo de parto (0:normal; 1:cesáreo), frequência respiratória (FR), consumo de ocitocina, dilatação uterina (DU).

Analisando os dados, verificou-se que a idade média das pacientes acompanhadas era de 22 anos. Até os 15 primeiros minutos após injeção da anestesia 100% das mulheres

ainda não haviam tido bebê. Das 49 mulheres amostradas, 37% tiveram o bebê em até 90 minutos após a injeção da anestesia. Verificou-se, ainda, que antes da anestesia ter sido injetada nas pacientes, 48,9% delas classificaram sua dor como intensa, e após os 5 primeiros minutos de efeito da anestesia esse percentual diminuiu para 36,7%. A maior parte (67,3%) das pacientes acompanhadas no estudo optaram pela anestesia padrão (Peridural). Ao avaliar duas medidas que interferem na evolução do trabalho de parto (dilatação uterina e consumo de ocitocina), é possível notar que a variação da dilatação uterina é maior entre as pacientes que classificaram sua dor como leve. Nota-se, ainda, que essa variação parece não diferir entre as pacientes cuja intensidade da dor é moderada ou intensa (Figura 5.3 (a)). Em relação ao consumo de ocitocina, aparentemente também não há diferença entre as categorias da resposta (Figura 5.3 (b)). Analisando essas medidas em relação ao tipo de anestesia que as pacientes usaram, verifica-se que há uma maior variabilidade em relação à dilatação uterina dentre as pacientes que usaram a anestesia epidural, sendo a mediana da dilatação para esse grupo em torno de 6cm. Já para as pacientes que usaram remifentanil, a mediana da dilatação é de 5cm (Figura 5.4 (a)). Para o consumo de ocitocina, nota-se uma maior variabilidade dentre as pacientes que usaram remifentanil, contudo, o consumo mediano é o mesmo para os 2 grupos (Figura 5.4 (b)).

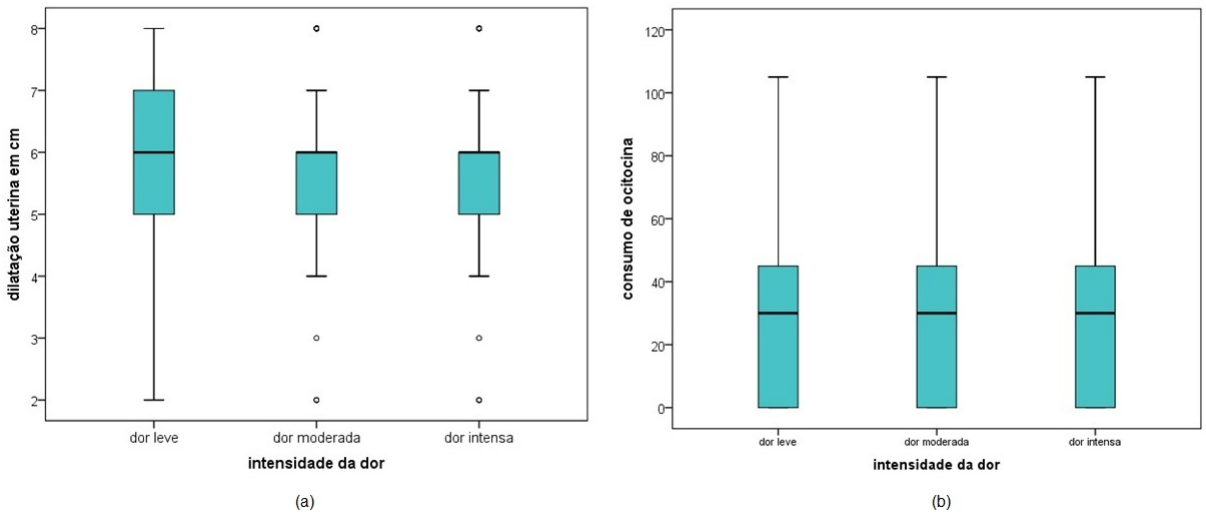

Figura 5.3: Boxplots para (a) dilatação uterina e (b) consumo de ocitocina em relação à intensidade da dor após os 5 minutos de anestesia

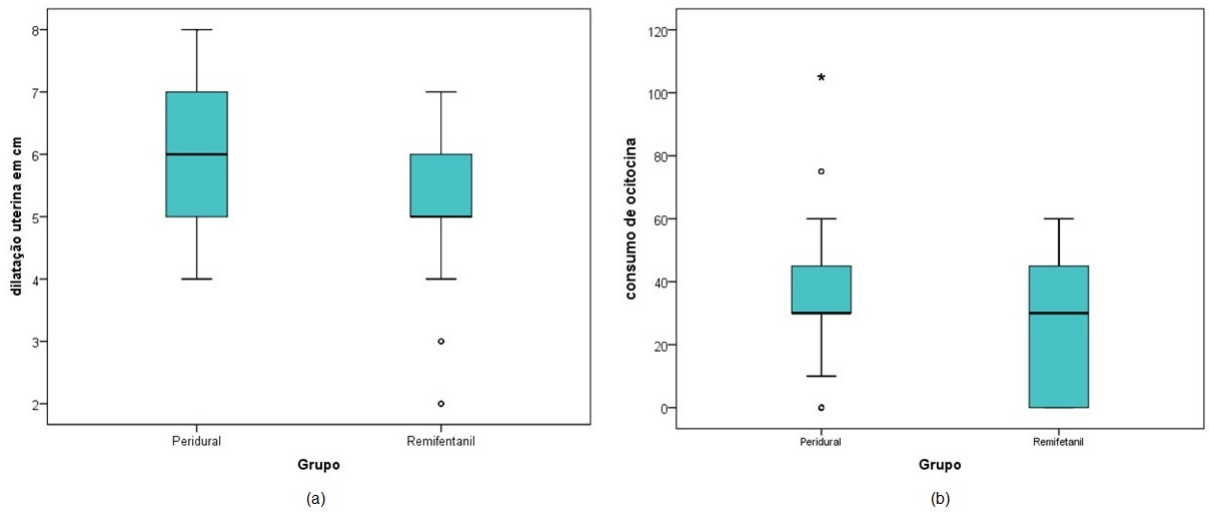

Figura 5.4: Boxplots para (a) dilatação uterina e (b) consumo de ocitocina em relação ao tipo de anestesia.

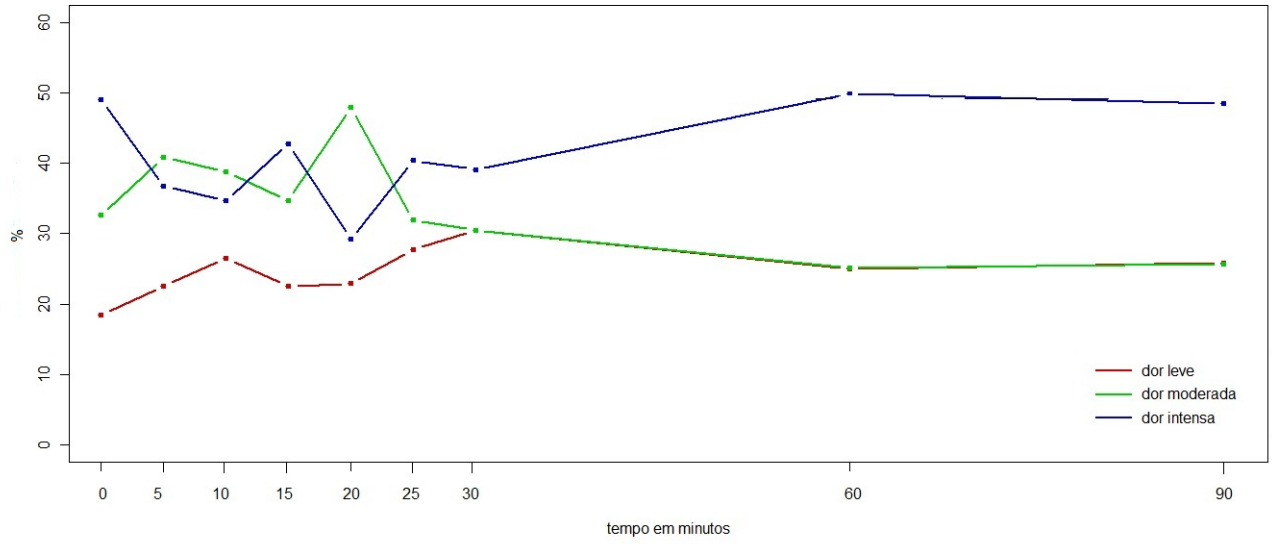

Figura 5.5: Perfil para a intensidade da dor até a hora do parto.

Na Figura 5.5 é possível observar o perfil médio da intensidade da dor durante os 90 minutos em que as pacientes foram acompanhadas. Nota-se que durante os primeiros 30 minutos após o início da anestesia há grande oscilação em relação à intensidade da dor entre as pacientes. Antes das pacientes receberem a anestesia (tempo 0), a proporção de pacientes com dor intensa era de 49%, e essa proporção cai para 36,7% após os 5 primeiros

minutos de anestesia. Já para as pacientes cuja dor era moderada ou leve essa proporção aumenta do tempo 0 para o 5. Observa-se, ainda, que após os 30 primeiros minutos da aplicação da anestesia, a proporção cai para as mulheres cuja dor é leve ou moderada, mas aumenta entre aquelas com dor intensa (entre 30 e 60 minutos), voltando a cair entre 60 e 90 minutos.

Após a realização de uma análise univariada, foram levadas para o modelo multivariado todas as covariáves com valor-p $< 0,20$ (grupo, idade, consumo de ocitocina, frequência respiratória, dilatação uterina e tempo). Nesta aplicação são ajustados apenas os modelos marginal e misto, visto que os modelos de transição não são aconselháveis quando o número de medidas repetidas por indivíduo é muito superior a 4, além disso, nesse estudo o interesse principal não é fazer predição. Assim, assumimos que o log da *odds* para uma das categorias da resposta ordinal em cada tempo j segue os seguintes modelos de *odds* proporcionais marginais e mistos (um modelo considerando efeito aleatório apenas no intercepto, e outro que considera no intercepto e no *slope* para a variável tempo), respectivamente. É válido ressaltar que os modelos foram ajustados a partir das funções do *software* R (R, 2012) mencionadas na Seção 5.1.

- Modelo marginal

$$\log\left[\frac{\mathbb{P}(Y_{ij} \leq k|X_{ij})}{1-\mathbb{P}(Y_{ij} \leq k|X_{ij})}\right] = \gamma_k + \beta_1 \text{grupo}_i + \beta_2 \text{idade}_i + \beta_3 \text{ocitocina}_i + \beta_4 \text{FR}_{ij} + \beta_5 \text{DU}_{ij} + \beta_6 \text{tempo}_{ij}$$

- GLMM

$$\log\left[\frac{\mathbb{P}(Y_{ij} \leq k|b_i)}{1-\mathbb{P}(Y_{ij} \leq k|bi)}\right] = \gamma_k + \beta_1 \text{grupo}_i + \beta_2 \text{idade}_i + \beta_3 \text{ocitocina}_i + \beta_4 \text{FR}_{ij} + \beta_5 \text{DU}_{ij} + \beta_6 \text{tempo}_{ij} + b_i$$

$$\log\left[\frac{\mathbb{P}(Y_{ij} \leq k|b_i)}{1-\mathbb{P}(Y_{ij} \leq k|bi)}\right] = \gamma_k + \beta_1 \text{grupo}_i + \beta_2 \text{idade}_i + \beta_3 \text{ocitocina}_i + \beta_4 \text{FR}_{ij} + \beta_5 \text{DU}_{ij} + \beta_6 \text{tempo}_{ij} + b_{1i} + b_{2i} tempo_{ij}$$

No modelo marginal (Tabela 5.9) três diferentes estruturas para a matriz de trabalho foram consideradas, e os resultados obtidos nos 3 ajustes foram bem similares, o que corrobora a teoria do modelo no sentido de que independente da matriz de trabalho escolhida, as estimativas para o coeficiente β são consistentes. Ressalta-se, contudo, que se o interesse for também a estimação do parâmetro α, é então importante escolher corretamente a matriz de trabalho. Nesse sentido, o modelo 1 seria descartado, pois a matriz

do tipo independente assume não haver associação intra-indivíduo, e portanto nenhum parâmetro é estimado para avaliar a associação. Para o modelo 3, 36 combinações entre as 9 medidas do tempo são consideradas, e com isso tem-se 36 estimativas para o parâmetro α. Devido à impossibilidade de escolher a melhor estrutura para a matriz $R(\alpha)$ fazendo uso do QIC, pode-se usar o critério da parcimônia, e optar pelo modelo que estime menos parâmetros de associação. No modelo 3 muitos dos α's estimados foram estatisticamente significativos, indicando haver mudança no log da chance de um tempo para outro, contudo, muitas medidas foram estimadas. Assim, por parcimônia, pode-se escolher o modelo 2, cuja medida de associação também foi estatisticamente significativa e um único α foi estimado. O pressuposto de proporcionalidade das *odds* foi testado e rejeitado, indicando que o mais adequado é considerar outro modelo para respostas ordinais. Em tal caso foi ajustado no *software* SAS 9.0 o modelo de *odds* proporcionais parciais considerando as variáveis preditoras grupo, frequência respiratória e dilatação uterina com efeitos variando de acordo com a categoria da resposta (Tabela 5.10). Além disso, por parcimônia, a matriz de trabalho considerada foi a simétrica composta, pois no ajuste considerando a matriz do tipo não estruturada a maioria dos efeitos estimados para a estrutura da associação não foram estatisticamente significativos.

Tabela 5.9: Estimativas dos parâmetros para o modelo marginal no estudo sobre analgesia do parto.

	$R(\alpha)$ independente			$R(\alpha)$ simétrica composta			$R(\alpha)$ não estruturada^{+}		
variável	$\widehat{\beta}$	ep*	p-valor	$\widehat{\beta}$	ep	p-valor	$\widehat{\beta}$	ep	p-valor
γ_1	1,351	1,218	-	1,804	1,287	-	1,669	1,308	-
γ_2	3,008	1,218	-	3,443	1,284	-	3,304	1,296	-
grupo (Remifetanil)	-1,629	0,369	<0,001	-1,793	0,408	<0,001	-1,829	0,044	<0,001
Tempo	-0,001	0,004	0,827	-0,002	0,004	0,608	-0,002	0,004	0,544
idade	-0,037	0,035	0,299	-0,039	0,037	0,303	-0,039	0,037	0,286
ocitocina	-0,001	0,008	0,866	-0,003	0,009	0,775	-0,001	0,008	0,846
DU	-0,173	0,130	0,185	-0,197	0,144	0,172	-0,173	0,130	0,185
FR	-0,016	0,033	0,620	-0,028	0,035	0,427	-0,021	0,035	0,548
$\widehat{\alpha}^{**}$	-			1,043	0,246	<0,001	$\widehat{\alpha}_{0,5} = 1,767$	0,709	0,012
							$\widehat{\alpha}_{0,10} = 0,948$	0,596	0,112
							$\vdots$		
							$\widehat{\alpha}_{60,90} = 1,188$	0,722	0,099

*erro-padrão baseado no estimador sanduiche da variância
**associação intra-indivíduo baseada no log da *odds* ratio
$^{+}$ Para a estrutura de associação são 36 combinações entre os tempos, ou seja 36 α's foram estimados

A partir dos resultados obtidos para o modelo marginal usando *odds* proporcionais parciais (Tabela 5.10) observa-se que apenas os efeitos de grupo e idade foram estatisticamente significativos. Nesse sentido, a chance de uma parturiente do estudo sentir dor leve será menor dentre aquelas que tomaram remifentanil ($e^{-1,963} = 0,14$) em comparação às que tomaram peridural. A chance da parturiente sentir dor leve ou

Tabela 5.10: Estimativas dos parâmetros para o modelo marginal usando *odds* proporcionais parciais no estudo sobre analgesia do parto.

variável	$\widehat{\beta}$	ep	p-valor
γ_1	1,817	1,199	-
γ_2	3,099	0,822	-
Grupo_1 (Remi)	-1,963	0,445	0,029
Grupo_2	-0,997	0,244	<0,001
tempo	-0,004	0,003	0,242
idade	-0,039	0,015	0,010
ocitocina	0,002	0,004	0,624
DU_1	-0,288	0,143	0,133
DU_2	-0,073	0,096	0,448
FR_1	0,003	0,047	0,171
FR_2	-0,061	0,029	0,036
$\widehat{\alpha}$	0,807	0,183	<0,001

moderada($e^{-0,997} = 0,37$) é também menor entre aquelas que tomaram remifentanil em comparação às que optaram por peridural. Já em relação à idade da paciente, a chance de uma paciente do estudo sentir dor (pouca, pouca ou moderada, intensa) diminui a cada 1 ano que é acrescido na idade da mesma ($e^{-0,039} = 0,96$). Como a variável idade não violou o pressuposto de proporcionalidade das *odds* essa chance será a mesma qualquer que seja a categoria da resposta. Já a chance da parturiente sentir dor leve ou moderada também reduz ($e^{-0,061} = 0,94$) com o aumento da frequência respiratória. A associação entre resposta e tempo de acompanhamento foi estatisticamente significativa, indicando que há mudança na chance de uma paciente permanecer em certa categoria da resposta de um tempo para o outro. Em tal caso, se não houvesse tantos tempos de acompanhamento a serem estimados, o aconselhável seria considerar a estrutura de associação do modelo 3.

Tabela 5.11: Estimativas dos parâmetros para o modelo misto no estudo sobre analgesia do parto.

	GLM			GLMM- intercepto aleatório			GLMM- intercepto e slope aleatórios		
variável	est	ep	p-valor	est	ep	p-valor	est	ep	p-valor
γ_1	1,462	0,735	-	2,634	1,250	-	2,624	1,294	-
γ_2	3,091	0,748	-	4,570	1,269	-	4,555	1,313	-
grupo (Remilfetanil)	1,246	0,230	<0,001	-1,519	0,418	<0,001	-1,517	0,410	<0,001
Tempo				-0,009	0,005	0,055	-0,009	0,005	0,060
idade	0,040	0,018	0,027	-0,050	0,033	0,127	-0,049	0,032	0,119
ocitocina	-0,003	0,004	0,545	0,005	0,007	0,501	0,005	0,007	0,504
DU	0,161	0,086	0,063	-0,218	0,165	0,184	-0,218	0,158	0,167
FR	0,030	0,026	0,246	-0,063	0,036	0,083	-0,063	0,038	0,103
$g_{11} = Var(b_{1i})$				0,946			0,912		
$g_{22} = Var(b_{2i})$							$2,37\times10^{-10}$		
TRV*						<0,0001			0,49
ICC					0,223				
AIC		850,67			821,75			824,54	

est: estimativa dos parâmetros; ep: erro-padrão
* sob H_0 $\sigma^2_{li} = 0, \quad l = 1, 2$
[1] ICC avaliado em b_{1i}

Para o modelo misto (Tabela 5.11) observa-se que o TRV comparando o primeiro e segundo modelo foi estatisticamente significativo ($< 0,0001$), indicando que a variância do efeito aleatório é diferente de 0. O mesmo não foi verificado (p-valor=0,49) para o segundo e terceiro modelo, indicando não haver necessidade de um efeito aleatório no coeficiente angular (*slope*). É válido ressaltar que para esta comparação a distribuição da estatística de teste é uma mistura de qui-quadrados $\left(\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2\right)$. Pelo AIC o segundo modelo é o melhor dentre os três ajustados. O pressuposto de proporcionalidade das *odds* foi testado, e o mesmo foi rejeitado (p-valor=0,003). As variáveis preditoras que violaram tal pressuposto foram o grupo, frequência respiratória e dilatação uterina. Foi ajustado um modelo de *odds* proporcionais parciais com efeito aleatório apenas no intercepto, e os resultados encontram-se na Tabela 5.12.

Tabela 5.12: Estimativas dos parâmetros para o modelo misto usando *odds* proporcionais parciais no estudo sobre analgesia do parto.

variável	estimativa	ep
γ_1	2,702	1,213
γ_2	4,245	1,001
grupo_1 (Remi)	-2,389	0,528*
grupo_2	-1,236	0,429*
tempo	-0,009	0,004*
idade	-0,049	0,036
ocitocina	0,004	0,008
DU_1	-0,406	0,163*
DU_2	-0,082	0,174
FR_1	0,004	0,044
FR_2	-0,096	0,035*
$g_{11} = \text{Var}(b_{1i})$	0,961	
ICC	0,226	

* p-valor$<$ 0,05

A partir do ajuste do modelo misto de *odds* proporcionais parciais (Tabela 5.12) observamos que dado duas parturientes que possuem o mesmo efeito aleatório, a chance de sentir dor leve é menor ($e^{-2,839} = 0,06$) para a parturiente que tomou remifentanil em comparação àquela que optou por peridural. Esse mesmo comportamento é observado em relação à dor leve ou moderada ($e^{-1,236} = 0,29$). Além disso, a cada 5 minutos acrescentados no tempo de acompanhamento, a chance de uma parturiente sentir dor (leve, moderada ou severa) diminui ($e^{-0,009} = 0,99$). É válido ressaltar que como a variável tempo não violou o pressuposto de proporcionalidade das *odds*, a referida chance vale para qualquer dos logitos. Observamos ainda que a cada 1 centímetro de aumento da dilatação uterina da parturiente, a chance dela sentir dor leve diminui ($e^{-0,406} = 0,67$). Observa-se, ainda, que há uma redução ($e^{-0,096} = 0,91$) na chance da parturiente sentir dor leve ou moderada quando a frequência respiratória aumenta. Para esse modelo 22,6% da variabilidade não explicada é atribuída à variação entre os tempos de acompanhamento.

Na Tabela 5.13 apresentamos um resumo dos efeitos que foram estatisticamente significativos nos dois modelos finais (marginal e misto usando *odds* proporcionais parciais) considerados nesta aplicação.

Tabela 5.13: Estimativas dos parâmetros em termos de razões de chance para os modelos finais no estudo sobre analgesia do parto.

variável	modelo marginal	modelo misto
grupo_1 (Remi)	$e^{-1,963} = 0,14$	$e^{-2,389} = 0,09$
grupo_2 (Remi)	$e^{-0,997} = 0,37$	$e^{-1,236} = 0,29$
tempo	-	$e^{-0,009} = 0,99$
idade	$e^{-0,039} = 0,96$	-
ocitocina	-	-
DU_1	-	$e^{-0,406} = 0,67$
DU_2	-	-
FR_1	-	-
FR_2	$e^{-0,061} = 0,94$	$e^{-0,096} = 0,91$

Nesse estudo o objetivo era comparar duas técnicas para analgesia da dor na hora do parto. A partir da modelagem marginal é possível fazer inferências sobre a média populacional a partir do comportamento da variável de grupo. Já no modelo misto o efeito da variável de grupo na intensidade da dor é avaliado a nível da paciente, condicional ao tempo de acompanhamento. Portanto, a escolha do modelo mais adequado para avaliação da analgesia da dor dependerá da pergunta que o pesquisador deseja responder ao comparar o efeito das duas técnicas na analgesia da dor. Assim sendo, se ao comparar as técnicas o esperado é que haja uma redução na intensidade da dor para a população em estudo que optou pelo uso de remifentanil, então o modelo marginal seria a abordagem mais adequada. Contudo, se o esperado no estudo é que condicional ao tempo de acompanhamento, a intensidade da dor, a nível da paciente, seja menor entre aquelas que optaram pelo uso de remifentanil, então o modelo misto é o mais adequado para a análise.

Capítulo 6

Considerações Finais

Nessa dissertação a modelagem de respostas politômicas ordinais foi discutida sob a ótica dos modelos de regressão para dados longitudinais. A especificação e interpretação dos modelos marginais, mistos e de transição para respostas ordinais foi ilustrada com a aplicação destas metodologias para a análise de 2 conjuntos de dados reais.

A modelagem de respostas ordinais tem sido alvo de crescente interesse nos últimos anos, e muitas metodologias foram propostas. O modelo de logitos cumulativos ou de *odds* proporcionais (McCullagh, 1986) é o mais conhecido, e assume que as *odds* em qualquer categoria da resposta ordinal são sempre as mesmas. Ao contrário do modelo para respostas politômicas nominais, a interpretação dos parâmetros não é obtida de forma direta, considerando uma das categorias da resposta como referência. No modelo de *odds* proporcionais a categoria de referência muda de acordo com o logito que está sendo analisado, e isso deve-se ao fato das probabilidades calculadas serem acumuladas. É importante nesse modelo que o pressuposto de proporcionalidade das *odds* seja testado, de forma que se possa garantir que as estimativas obtidas são confiáveis, e o modelo final possa ser considerado. Nas situações em que este pressuposto é violado, o aconselhado é que se busque outra metodologia adequada para respostas ordinais. Dentre as existentes há o modelo de *odds* proporcionais parciais (Peterson e Harrell, 1990), que nada mais é do que uma generalização do modelo de logitos cumulativos onde se permite que um subconjunto de variáveis preditoras que tiveram o pressuposto violado entre no modelo tendo seu efeito variando de acordo com a categoria da resposta ordinal, além de inserir o subconjunto de variáveis preditoras que não violaram o referido pressuposto.

Os trabalhos que discutem a modelagem de respostas ordinais restringem-se, em sua marioria, aos aspectos teóricos do modelo, em especial aos do modelo de *odds* proporcionais, talvez pelo mesmo se encontrar disponível na maioria dos *softwares* estatísticos. Contudo, pouca atenção é dada à interpretação do modelo ajustado, de forma a esclarecer como a mesma pode ser feita. Em decorrência disso a resposta ordinal é muitas vezes recategorizada e trabalhada como sendo uma variável binária, cujos modelos são mais

discutidos na literatura, e sua interpretação é mais direta.

Os modelos para respostas ordinais usados em estudos transversais também podem ser utilizados no contexto longitudinal. Em estudos longitudinais os modelos usuais (modelos marginais, mistos e de transição) podem ser estendidos de forma a acomodar a natureza ordinal da resposta, sendo importante ressaltar que a escolha da metodologia mais adequada dependerá do tipo de pergunta que o pesquisador deseja responder. Assim sendo, tais modelos não são comparáveis, pois o foco da análise e a forma como os mesmos são interpretados não são as mesmas. O modelo marginal, por exemplo, possui como foco da análise inferir sobre a média populacional, sendo uma das metodologias mais populares na modelagem longitudinal por basear-se exclusivamente em suposições sobre a resposta média. Já o modelo linear generalizado misto tem como foco o indivíduo, e a introdução de um efeito aleatório no modelo, a nível do indivíduo, tem importantes implicações nas estimativas dos coeficiente de regressão. O modelo de transição, por sua vez, é dentre os três o menos discutido na literatura longitudinal, sendo utilizado quando o objetivo é predizer o comportamento das respostas futuras a partir das passadas e de um conjunto de preditores, sendo desaconselhável seu uso quando o número de medidas repetidas por indivíduo for muito superior a 4.

Como o objetivo do trabalho foi apresentar a modelagem de respostas politômicas ordinais em estudos longitudinais, focando especialmente na especificação e interpretação dos modelos trabalhados, foram realizadas duas aplicações com conjuntos de dados reais, de forma a se obter um melhor entendimento da metodologia apresentada. As três abordagens discutidas no Capítulo 4 foram consideradas para análise dos dados da aplicação 1, e para a aplicação 2 apenas o modelo de transição não foi considerado por haver um número de medidas muito superior a 4. Em ambas as aplicações o pressuposto de proporcionalidade das *odds* foi violado para os modelos marginais e misto, e o modelo de *odds* proporcionais parciais foi considerado como modelo final. Nas duas aplicações usadas para ilustrar a metodologia estudada nessa dissertação ressaltamos que a escolha do melhor modelo dependerá do objetivo do estudo.

Ainda há muito o que ser estudado e discutido sobre a modelagem de respostas ordinais em estudos longitudinais. Uma proposta para trabalhos futuros seria explorar técnicas de diagnóstico para respostas ordinais, um tópico ainda pouco explorado em estudos com dados dessa natureza. Além disso, outros modelos podem também ser considerados para análise de respostas ordinais em estudos longitudinais, sendo um deles a classe de modelos conhecida como modelos de transição marginalizados (Azzalini, 1994; Lee e Daniels, 2007), ou ainda a modelagem conjunta de respostas ordinais longitudinais e dados de sobrevivência (Li et al., 2010, Chakrabort, 2010).

Apêndice A

Códigos usados nos *software* R e SAS

Nesse apêndice são apresentados os códigos em R - 2.15 e SAS 9.0, utilizados nos exemplos citados no texto.

1. **Código R usado na Tabela 2.2** (Seção 2.1):

   require (VGAM)
   mod<- vglm (cbind(ind, grupo, sala) ~ factor(escola)+factor(periodo), multinomial, data)

   Obs.: A última categoria da resposta é a cosiderada no R como sendo a referência. Para mudá-la bastar usar 'multinomial (reflevel=categoria)'.

2. **Código R usado na Tabela 2.3** (Seção 2.2.1):

   require (ordinal)
   mod<-clm(Poliparasit ~ Sexo+Freq.banho+Grupo+idade, data)

3. **Código R usado na Tabela 2.4** (Seção 2.2.2):

   require (VGAM)
   mopp<-vglm(Poliparasit ~ Sexo+Freq.banho+Grupo+idade, family=cumulative (parallel=FALSE ~ Grupo+idade),na.action = na.omit)

   Obs.: Para testar o pressuposto de proporcionaliade das *odds* foi usado o seguinte comando:

mop<-vglm(ordered(Poliparasit) ~ Sexo+Freq.banho+Grupo+idade, family=cumulative (parallel=T),na.action = na.omit)

mopp<-vglm(Poliparasit ~ Sexo+Freq.banho+Grupo+idade, family=cumulative (parallel=F),na.action = na.omit)

teste<-pchisq(deviance(mop)-deviance(mopp), df=df.residual(mop)-df.residual(mopp), lower.tail=F) (teste global para avaliar o pressuposto)

Obs. 2: Em sendo violado o pressuposto do modelo avalia-se individualmente cada covariável, de forma similar ao ajuste anterior.

4. **Código R usado na Tabela 5.3** (Capítulo 5):

require (geepack)
mod1<-ordgee(ordered(Poliparasit) ~ Grupo+Sexo+Freq.banho+idade+Etapa,id=Id, mean.link="logit", corstr="exchangeable",int.const = TRUE, rev=TRUE)

Obs.: Para mudar a estrutura da matriz $R_i(\alpha)$ basta mexer em 'corstr=', que para a função 'ordgee' pode ser: 'independence', 'unstructured', 'exchangeable' ou 'userdefined' (onde montamos a matriz).

Obs.2: Para testar o pressuposto de proporcionalidade das *odds* no modelo marginal usamos o comando:

require (gee)
require (repolr)
a<-repolr(resp ~ factor(grupo)+factor(sexo)+factor(banho)+idade+Etapa, subjects="Id", data, categories=3,times=c(1,2,3), corstr = "uniform", po.test = TRUE,fixed=FALSE)
summary(a[["gee"]])

5. **Código SAS usado na Tabela 5.4** (Capítulo 5):

libname a
proc print data=data;
run;
data; set a.data;
do; if resp.poli=1 then new.resp=1;

```
else new.resp=0; logtype=1; output; end;
do; if resp.poli=1 or resp.poli=2 then new.resp=1;
else resp.poli=0; logtype=2; output; end;
run;


proc genmod descending order=data;
class Id grupo sexo freq.banho logtype;
model new.resp= sexo freq.banho grupo idade Etapa logtype
logtype*grupo logtype*idade logtype*Etapa/
link=logit dist=bin type3;
repeated subject=Id / logor=exch;
run;
```

6. **Código R usado na Tabela 5.5** (Capítulo 5):

```
require (ordinal) g3<-clmm(ordered(Poliparasit) ~ Grupo+Sexo+Freq.banho+idade+
Etapa+(1|Id), nAGQ = 50, Hess = TRUE, threshold="flexible")
```

Obs.: Para ajustar o modelo assumindo independência, basta usar a função descrita no item 2. Para ajustar o modelo com intercepto aleatório usa-se (1|Id), somada à função. Já o modelo contendo intercepto e slope aleatórios, usa-se (1|Id)+(1|Etapa).

Obs. 2: O pressuposto de proporcionalidade das *odds* é testado usando:

```
g5<-clmm2(ordered(Poliparasit) 1, nominal = ~ Grupo+Sexo+Freq.banho+idade+
Etapa, data = infec,link = "logistic",nAGQ = 50, random = factor(Id), Hess =
TRUE, threshold="flexible")

g6<-clmm2(ordered(Poliparasit) ~ Grupo+Sexo+Freq.banho+idade +Etapa,random
= factor(Id), nAGQ = 50, Hess = TRUE, threshold="flexible")

anova(g6,g5)
```

7. **Código R usado na Tabela 5.6** (Capítulo 5):

   require(ordinal)
   gfinal<-clmm2(ordered(Poliparasit) ∼ Sexo+Freq.banho, nominal = ∼ Grupo+idade+ Etapa, data = infec,link = "logistic",nAGQ = 50, random = factor(Id), Hess = TRUE, threshold="flexible")

8. **Código R usado na Tabela 5.7** (Capítulo 5):

   require(VGAM)
   mod1<-vglm(resp.atual ∼ Sexo+Freq.banho+Grupo+idades+resp.pas+factor(Etapa), family=cumulative(parallel=T),na.action = na.omit)

Obs.: Os comandos em R para as tabelas 5.8, 5.10 e 5.11, respectivamente, são analógos aos itens 4, 6 e 7. O mesmo vale para o comando em SAS da tabela 5.9.

Referências Bibliográficas

[1] ABREU, M.S., SIQUEIRA, A. L., CAIAFFA, W. T. Regressão logística em estudos epidemiológicos. *Revista de Saúde Pública*, **43**, 183−194, 2009.

[2] AGRESTI, A. Categorical Data Analysis. John Wiley: New York, 2002.

[3] ANANTH, C.V, KLEINBAUM, D.G. Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology*, **26**, 1323−1333, 1997.

[4] AZZALINI, A. Logistic regression for autocorrelated data with application to repeated measures. *Biometrika*, **81**, 767−775, 1994.

[5] BRESLOW, N.E., CLAYTON, D.G. Approximate inference in generalized linear mixed models. *JASA*, **88**, 9−25, 1993.

[6] CAREY, V. J., ZEGER, S. L.,DIGGLE, Peter J. Modelling Multivariate Binary Data With Logistic Regressions. *Biometrika*, **80**, 517−526, 1993.

[7] CHAKRABORT, A., DAS, K. Inferences for joint modelling of repeated ordinal scores and time to event data. *Computational and Mathematical Methods in Medicine*, **11**, 281−295, 2010.

[8] CLAYTON, D. Repeated Ordinal Measurements: A generalized estimating equation approach. *Technical Report - Medical Research Council Biostatistics unit, Cambridge*, 1992.

[9] CORDEIRO, G.M., LIMA NETO, E.A. Modelos Paramétricos, 2006.

[10] DIGGLE, Peter J., HEAGERTY, P., LIANG, K-Y., ZEGER, S. L. Analysis of Longitudinal Data. OXFORD University Press - second edition, 2002.

[11] FITZMAURICE, G., DAVIDIAN, M., VERBEKE, G., MOLENBERGHS, G. Longitudinal Data Analysis (Handbooks of Modern Statistical Methods). Chapman & Hall/CRC, 2009.

[12] FITZMAURICE, G., LAIRD, N.M., WARE, J.H. Applied Longitudinal Analysis. Wiley Series in Probability and Statistics - second edition, 2011.

[13] FONSECA, J. E. Implantação de cisternas para armazenamento de água de chuva e seus impactos na saúde infantil: Um estudo de coorte em Berilo e Chapada do Norte, MG. *Dissertação de mestrado*, 2012.

[14] GANGE, S.J., LINTON, K.L.P., SCOTT, A.J., DeMETS, D.L., KLEIN, R. Analysis of correlated ordinal measures with ophthalmic applications. *Technical Report 77, University of Wisconsin Department of Biostatistics*, 1993.

[15] GANJALI, M., REZAEE, Z. A transition model for analysis of repeated measure ordinal response data to identify the effects of different treatments. *Biostatistics & Analysis*, **41**, 527−534, 2007.

[16] HARTZEL, J., AGRESTI, A., CAFFO, B. Multinomial logit random effect models. *Statistical Modelling*, **1**, 81−102, 2001.

[17] HARVILLE, D.A. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320−340, 1977.

[18] HEAGERTY, P., ZEGER, S. L. Marginal regression models for clustered ordinal measurements. *JASA*, **91**, 1024−1036, 1996.

[19] HEAGERTY, P., ZEGER, S. L. Marginalized multilevel models and likelihood inference (with discussion). *Statistical Science*, **15**, 1−26, 2000.

[20] HEAGERTY, P. Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics*, **58**, 342−351, 2002.

[21] HEDEKER, D., GIBBONS, R. D. A random-effects ordinal regression model for multilevel analysis. *Biometrics*, **50**, 933−944, 1994.

[22] HEDEKER, D., MERMELSTEIN, R.J. A multilevel thresholds of change model for analysis of stages of change data. *Multivariate Behavioral Research*, **33**, 427−455, 1998.

[23] HEDEKER, D., MERMELSTEIN, R.J. Analysis of longitudinal substance use outcomes using ordinal random-effects regression models. *Addiction*, **95 - Supplement 3**, $S381-S394$, 2000.

[24] HEDEKER, D. A mixed-effects multinomial logistic regression model. *Statistics in Medicine*, **22**, 1433−1446, 2003.

[25] HEDEKER, D., GIBBONS, R. D. Longitudinal data analysis. John Wiley & Sons, 2006.

[26] HOSMER, D.W., LEMESHOW, S. Applied Logistic Regression. John Wiley & Sons - second edition, 2000.

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

[28] LEE, Y., NELDER, J.A. Hierarchical Generalized Linear Models. *Journal of the Royal Statistical Society, Series B*, **58**, 619−678, 1996.

[29] LEE, K., DANIELS, M. A class of Markov models for longitudinal ordinal data. *Biometrics*, **63**, 1060−1067, 2007.

[30] LI,N., ELASHOFF, R M., LI, G., SAVER, J. Joint modeling of longitudinal ordinal data and competing risks survival times and analysis of the NINDS rt-PA stroke trial. *Statistics in Medicine*, **29**, 546−557, 2010.

[31] LIANG, K.Y., ZEGER, S.L. Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13−22, 1986a.

[32] LIANG, K.Y., ZEGER, S.L. Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics*, **42**, 121−130, 1986b.

[33] LIPSITZ, S.R., LAIRD, N.M., HARRINGTON, D.P. Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, **78**, 153−160, 1991.

[34] LIPSITZ, S.R., KIM, K., ZHAO, L. Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine*, **13**, 1149−1163, 1994.

[35] McCULLAGH, P. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society, Series B*, **42**, 109−142, 1980.

[36] McCULLAGH, P., NELDER, J.A. Generalized linear models. Chapman and Hall, London, 1989.

[37] McCULLOCH, C.E., SEARLE, S.R. Generalized, Linear and Mixed Models. Wiley Series in Probability and Statistics, 2001.

[38] MILLER, M.E., DAVIS, C.S., LANDIS, J.R. The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares. *Biometrics*, **49**, 1033−1044, 1993.

[39] NELDER, J.A., WEDDERBURN, R.W.M. Generalized Linear Models. *Journal of the Royal Statistical Society, Series A*, **135**, 370−384, 1972.

[40] O' CONNELL, A.A. Logistic regression models for ordinal response variables. Sage Pubblications, 2006.

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

[42] PARSONS, N.R., COSTA, M.L., ACHTEN, J., STALLARD, N. Repeated measures proportional odds logistic regression analysis of ordinal score data in the statistical software package R. emphComputational Statistics & Data Analysis, **53**, 632−641, 2009.

[43] PATTERSON, H.D., THOMPSON, R. Recovery of inter-block information when blocks sizes are unequal. *Biometrika*, **58**, 545−554, 1971.

[44] PETERSON, B., HARRELL, Jr. F.E. Partial Proportional Odds Models for Ordinal Response Variables. *Journal of the Royal Statistical Society- Series C*, **39**, 205−217, 1990.

[45] PETTITT, A., HAYNES, M., TRAN, T., HAY, J. A model for longitudinal employment status of immigrants to Australia. *Queensland University of Technology, Brisbane*, 2002.

[46] PINHEIRO, J.C., BATES, D.M. Approximations to the log-likelihood function in the nonlinear mixed effects model. *Journal of computation and graphical statistics*, **4**, 12−35, 1995.

[47] PRENTICE, R.L. Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**, 1033−1048, 1988.

[48] PRENTICE, R.L., ZHAO, L.P. Estimating equations for parameters in means and covariates of multivariate discrete and continuous responses. *Biometrics*, **47**, 825−839, 1991.

[49] R: A Language and Environment for Statistical Computing. *http://www.R-project.org/*, Vienna-Austria, 2012.

[50] SAS Institute Inc. *SAS/STAT Software*, Cary, NC - 2002.

[51] SKRONDAL, A., RABE-HESKETH,S. Multilevel Modelling (volume 2-chapter 19). Sage Pubblications, 2010.

[52] SNIJDERS, T. A. B., BOSKER, R. J. Multilevel Analysis: An introduction to basic and advanced multilevel modeling. Sage Pubblications, 1999.

[53] SOARES, E. S. C. Remifetanil venoso vs analgesia peridural intermitente para o trabalho de parto - Estudo comparativo. *Dissertação de mestrado*, 2013.

[54] STOKES, M.E., DAVIS,C.S., KOCH, G.G. Categorical data analysis using the SAS® system - 2nd edition. SAS Publishing, 2000.

[55] WALKER, S., DUNCAN, D. Estimation of the Probability of an Event As a Function of Several Independent Variables. *Biometrika*, **54**, 167−179, 1967.

[56] WEDDERBURN, R.W.M. Quasi-likelihood function, generalized linear models and the Gauss-Newton method. *Biometrika*, **61**, 439−477, 1974.