UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

JOSÉ LUIZ PADILHA DA SILVA

TRATAMENTO DE DADOS PERDIDOS EM ESTUDOS LONGITUDINAIS COM RESPOSTAS ORDINAIS

Belo Horizonte
2015

JOSÉ LUIZ PADILHA DA SILVA

TRATAMENTO DE DADOS PERDIDOS EM ESTUDOS LONGITUDINAIS COM RESPOSTAS ORDINAIS

Tese apresentada ao Programa de Pós-Graduação em Estatística do Departamento de Estatística da Universidade Federal de Minas Gerais como parte dos requisitos para a obtenção do grau de Doutor em Estatística.

Orientador: Prof. Dr. Enrico A. Colosimo

Coorientador: Prof. Dr. Fábio N. Demarqui

Belo Horizonte

2015

À Viviane C. Ceschim,

razão da minha alegria!

AGRADECIMENTOS

Agradeço a Deus por ter me dado forças para alcançar mais esta conquista.

Aos meus pais e irmãos, por compreenderem a minha ausência em todos esses momentos, e pelo apoio incondicional.

Aos professores Enrico Colosimo e Fábio Demarqui, pela orientação, paciência e incentivo.

A todos os meus amigos que me incentivaram e me permitiram muitos momentos de alegria.

À FAPEMIG e à CAPES pelo apoio financeiro.

A todos que contribuíram direta ou indiretamente, meu muito obrigado!

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Esta tese tem por objetivo o tratamento de dados perdidos em estudos longitudinais com resposta ordinal. Em tais estudos a perda de dados na resposta e/ou covariáveis pode introduzir viés e levar a inferências enganosas sobre os parâmetros de regressão (Fitzmaurice *et al.*, 2004).

Neste trabalho, assumimos um modelo de razão de chances proporcionais para a resposta longitudinal ordinal e adotamos Equações de Estimação Generalizadas (GEE) (Liang & Zeger, 1986) para estimação dos parâmetros de regressão. O método GEE apresenta simplicidade computacional e objetiva estimar parâmetros fixos sem especificar a distribuição conjunta para as medidas repetidas. Contudo, na presença de dados ausentes, o estimador GEE padrão é consistente apenas sob a forte suposição de perda completamente ao acaso. Nós propomos um estimador GEE duplamente robusto para análise de dados longitudinais com perda intermitente na resposta e em uma covariável, ambas sujeitas a um mecanismo de perda ao acaso. O método proposto combina imputação múltipla e o método GEE ponderado e é atrativo no sentido de que, para consistência, requer a especificação correta de apenas um de seus modelos preditivos.

Num primeiro momento são assumidas equações de estimação independentes, o que simplifica o processo iterativo de estimação, e o foco é a comparação do viés do método GEE proposto frente a imputação múltipla e ao método GEE ponderado. Embora as correlações entre as respostas longitudinais sejam usualmente tratadas como parâmetros de perturbação, sabe-se que com covariáveis dependentes do tempo pode-se melhorar a precisão das estimativas especificando a estrutura de dependência. Com este objetivo, o estimador proposto foi estendido para acomodar a modelagem da estrutura de associação por meio do coeficiente de correlação e também de razão de chances locais. Dois conjuntos de dados reais oriundos da área médica ilustram a aplicação dos métodos.

Palavras-chave: Coeficiente de Correlação, Equações de Estimação Generalizadas, Estimadores Duplamente Robustos, Perda ao Acaso, Razão de Chances Locais.

ABSTRACT

The main goal of this thesis is the treatment of missing data in longitudinal studies with ordinal response. In such studies missing data in the response and/or covariates can introduce bias and lead to misleading inferences about the regression parameters (Fitzmaurice *et al.*, 2004).

In this work, we adopt a proportional odds model for the longitudinal ordinal response and make use of Generalized Estimating Equations (Liang & Zeger, 1986) (GEE) to estimate the regression parameters. The GEE method presents computational simplicity and is intended to estimate fixed parameters without specifying the joint distribution for repeated measures. Nevertheless, in the presence of missing data, the standard GEE estimator is consistent only under the strong assumption of missing completely at random data. We propose a doubly robust GEE estimator for analysis of longitudinal ordinal data with intermittent missing response and covariate, both subject to a missing at random mechanism. The proposed method combines ideas of multiple imputation and the weighted GEE and it is attractive in the sense that, for consistency, it requires only the correct specification of one of its predictive models, but not necessarily both.

First, independent estimating equations are assumed, which simplifies the iterative estimation process, and we focus on the bias of the proposed method compared to multiple imputation and the weighted GEE. Although the correlations between the longitudinal responses are usually treated as nuisance parameters, it is well known that in the presence of time-varying covariates the efficiency of the estimates can be improved by specifying the dependence structure. For this purpose, the proposed estimator is extended to accommodate the modeling of the association structure by means of the correlation coefficient as well as local odds ratio. Two real data sets from the medical field are used to illustrate the methods.

Keywords: Correlation Coefficient, Doubly Robust Estimators, Generalized Estimating Equations, Local Odds Ratio, Missing at Random.

CAPÍTULO 1

INTRODUÇÃO

1.1 Introdução

Dados longitudinais surgem quando cada indivíduo é avaliado repetidamente ao longo do tempo. Estas medidas repetidas para um dado indivíduo formam um conglomerado e se espera que as respostas dentro de um conglomerado apresentem correlação positiva (Fitzmaurice *et al.*, 2004). A dependência entre as observações de um mesmo conglomerado e a ordenação temporal devem ser incorporadas na análise.

Dados longitudinais ordinais surgem naturalmente em muitos cenários clínicos. Por exemplo, em ensaios clínicos, é comum a avaliação da qualidade de vida do paciente por meio de uma escala do tipo Likert (Donneau *et al.*, 2015a). O uso de respostas ordinais tem ganhado espaço em estudos de indicadores de condição de saúde, e também da gravidade de certas doenças. Em ensaios clínicos, respostas numa escala ordinal são muitas vezes utilizadas para quantificar os sintomas ou condição do paciente, bem como avaliar a eficiência de procedimentos pós-operatórios (Parsons *et al.*, 2009).

Em tais estudos longitudinais, contudo, os indivíduos podem abandonar o estudo prematuramente enquanto outros podem perder uma ou outra ocasião de medida. É bastante comum, também, que haja informação incompleta em algumas covariáveis que possam explicar a evolução longitudinal. Estes dados ausentes apresentam importantes implicações para a análise, como perda de eficiência e viés nas estimativas dos parâmetros de interesse (Fitzmaurice *et al.*, 2004).

Esta tese tem como objetivo estudar o impacto dos dados ausentes na modelagem de dados longitudinais com resposta ordinal, quando a perda dos dados ocorre com padrão arbitrário tanto na resposta como na covariável. O propósito das seções que se seguem é introduzir, de maneira sucinta, alguns tópicos encontrados na análise de dados longitudinais com resposta ordinal e a teoria básica para o tratamento de dados ausentes.

1.1.1 Modelo de Chances Proporcionais

Considere uma amostra de n indivíduos para os quais uma resposta ordinal O com J categorias é avaliada em T_i ($T_i \leq T$) ocasiões para cada indivíduo. Então O_{it} denota a avaliação da resposta ordinal O para o i-ésimo indivíduo ($i = 1, \ldots, n$) na t-ésima ocasião ($t = 1, \ldots, T_i$). Associado à resposta há um vetor $p \times 1$ de covariáveis, $\boldsymbol{X}_{it}$, medido no tempo t. A natureza ordinal da variável resposta pode ser levada em conta considerando probabilidades acumuladas $Pr(O_{it} \leq j),\ j = 1, \ldots, J$. A função de ligação logito cumulativo, que é geralmente aplicada para relacionar as probabilidades marginais da resposta ao vetor de covariáveis $\boldsymbol{X}$ (McCullagh, 1980), é dado por

$$\text{logit}\left[Pr(O_{it} \leq j|\boldsymbol{X}_i)\right] = \beta_{0j} + \boldsymbol{X}_{it}^T\boldsymbol{\beta}_x, \quad j = 1, \ldots, J-1, \tag{1.1}$$

em que $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0,J-1})^T$ são os parâmetros de intercepto e $\boldsymbol{\beta}_x = (\beta_1, \ldots, \beta_p)^T$ é o vetor de coeficientes associados com $\boldsymbol{X}_{it}$. Nesta tese, a função de ligação logito é escolhida pela sua simplicidade e facilidade de interpretação. A formulação em (1.1) implica um modelo de chances proporcionais (McCullagh, 1980). Neste modelo, o efeito de $\boldsymbol{\beta}_x$ é independente da categoria (é invariante à combinação de categorias) e o exponencial dos parâmetros pode ser interpretado como uma razão de chances (Agresti, 2013).

A estimação dos parâmetros da regressão pode ser realizada por meio de métodos baseados em verossimilhança. Uma dificuldade, contudo, é a complexidade da relação entre os parâmetros do modelo e as probabilidades conjuntas que definem a verossimilhança (Donneau *et al.*, 2015a). Uma alternativa comumente adotada são as chamadas Equações de Estimação Generalizadas (GEE), método proposto por Liang & Zeger (1986), e que tem se tornado bastante popular na análise de dados correlacionados não gaussianos. Tal popularidade deve-se ao fato de os parâmetros estimados terem interpretação marginal. O método evita a especificação da distribuição conjunta das medidas repetidas ao permitir adotar uma estrutura de correlação de 'trabalho', e requer apenas que as distribuições marginais sejam especificadas. Uma característica atrativa do método GEE é que os parâmetros de associação são tratados como parâmetros de perturbação e,

diferentemente do que ocorre com métodos baseados em verossimilhança, as estimativas dos parâmetros de regressão do GEE permanecem válidas mesmo quando a estrutura de correlação é mal especificada (Liang & Zeger, 1986). Outra importante vantagem do GEE é sua simplicidade computacional.

1.1.2 O Método GEE

Como o modelo de chances proporcionais não é parte da família regular de modelos lineares generalizados, são necessárias algumas transformações antes de se aplicar o método GEE. Para cada indivíduo, em cada tempo, define-se um vetor expandido de variáveis binárias $\boldsymbol{Y}_{it} = (Y_{it1}, \ldots, Y_{it(J-1)})^T$ de dimensão $J-1$. Há três opções para escolher as variáveis binárias. A primeira define $Y_{itj} = I(O_{it} = j)$ (Lipsitz *et al.* (1994), Touloumis *et al.* (2013)), a segunda define $Y_{itj} = I(O_{it} > j)$ (Heagerty & Zeger (1996)) e a terceira define $Y_{itj} = I(O_{it} \leq j)$ (Parsons *et al.* (2006)). Em todas as opções $I(\cdot)$ é a função indicadora que vale um se o argumento é verdadeiro e zero, caso contrário. Por conveniência, a primeira opção será adotada neste trabalho. As diferentes codificações das variáveis binárias não devem, em princípio, levar a estimativas diferentes dos parâmetros fixos. Consequentemente, o vetor $\boldsymbol{\mu}_i = E(\boldsymbol{Y}_i|\boldsymbol{X}_i)$ é a média de todas as variáveis binárias $\boldsymbol{Y}_i = (Y_{i1}^T, \ldots, Y_{iT_i}^T)^T$. O vetor de parâmetros de regressão $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_x^T)^T$ pode ser estimado por meio do método GEE resolvendo o seguinte sistema de equações

$$\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}, \tag{1.2}$$

em que $\boldsymbol{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T}$, e $\boldsymbol{V}_i = \boldsymbol{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ é uma matriz de pesos ou matriz de covariâncias de $\boldsymbol{Y}_i$. O vetor $\boldsymbol{\alpha}$ expressa uma suposição de 'trabalho' para a estrutura de correlação/associação entre as medidas repetidas.

Liang & Zeger (1986) provaram que, dada qualquer parametrização da matriz $\boldsymbol{V}_i$ e a correta especificação do modelo marginal em (1.1), a solução $\hat{\boldsymbol{\beta}}$ de (1.2), é um estimador consistente de $\boldsymbol{\beta}$ e $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ tem uma distribuição assintótica normal multivariada com vetor de médias $\boldsymbol{0}$ e matriz de covariância $\boldsymbol{V}_{\boldsymbol{\beta}} = \lim_{n\to\infty} n\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{-1}$, em que

$\boldsymbol{\Sigma}_0 = \sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} \boldsymbol{D}_i^T$, e $\boldsymbol{\Sigma}_1 = \sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} \text{Cov}(\boldsymbol{Y}_i) \boldsymbol{V}_i^{-1} \boldsymbol{D}_i^T$. Na prática, a matriz de covariâncias do tipo "sanduíche" $\boldsymbol{V}_{\boldsymbol{\beta}}$ é calculada ignorando o limite e substituindo $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ e $\text{Cov}(\boldsymbol{Y}_i)$ por $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ e $(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)^T$, respectivamente (Touloumis *et al.*, 2013).

1.1.3 Estimação do vetor de associação e da matriz de covariâncias

A forma mais simples de determinar estimativas GEE é assumir independência entre os pares da resposta de cada indivíduo, ou seja, restringir $\boldsymbol{\alpha} = \mathbf{0}$. As Equações de Estimação Independentes (IEE) são eficientes apenas quando as covariáveis são constantes no tempo ou quando a estrutura de independência é verdadeira (Lipsitz *et al.*, 1994). Por outro lado, quando existe dependência entre as medidas repetidas ou covariáveis dependentes do tempo, é possível ganhar eficiência modelando a estrutura de associação (Touloumis *et al.*, 2013).

Várias propostas têm sido formuladas para a estimação do vetor de associação $\boldsymbol{\alpha}$. Lipsitz *et al.* (1994) definiram $\boldsymbol{\alpha}$ como o coeficiente de correlação de Pearson e sugeriram o uso do método dos momentos para a estimação de várias estruturas de correlação. Ainda considerando o coeficiente de correlação, Parsons *et al.* (2006) propuseram um enfoque que estima o vetor de asociação $\boldsymbol{\alpha}$ minimizando uma função objetivo $Q(\boldsymbol{\alpha}|\boldsymbol{\beta}, \boldsymbol{Y})$. Entre outros, Lipsitz *et al.* (1991) e Carey *et al.* (1993) demonstraram que o coeficiente de correlação entre duas respostas binárias é restrito pelas probabilidades marginais. Como alternativa ao coeficiente de correlação, vários autores adotaram a razão de chances como medida de associação. Lumley (1996) propôs o uso de razão de chances globais. Heagerty & Zeger (1996) estenderam o método da regressão logística alternada de Carey *et al.* (1993) para dados ordinais introduzindo um segundo conjunto de equações de estimação para estimar as razões de chances globais. Recentemente, Touloumis *et al.* (2013) utilizaram razões de chances locais para estimar os parâmetros de associação na matriz $\boldsymbol{V}_i$. Os autores usaram modelos de associação (Goodman, 1985) para reparametrizar as razões de chances locais e, então, estimaram os parâmetros usando o procedimento IPFP (Iterative Proportional Fitting Procedure, Deming & Stephan (1940)).

Neste trabalho, confrontaremos as abordagens de correlação de Lipsitz *et al.* (1994) e

de razões de chances locais de Touloumis *et al.* (2013). Tal comparação é inédita para o caso de dados ausentes. As duas parametrizações são introduzidas a seguir para o caso de dados completos.

1.1.3.1 Coeficiente de Correlação

Lipsitz *et al.* (1994) sugeriram um método que restringe as correlações entre os pares de respostas nos diferentes tempos. Nesta abordagem, a matriz de covariâncias $\boldsymbol{V}_i$ é decomposta na forma $\boldsymbol{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{F}_i^{1/2}(\boldsymbol{\beta})\boldsymbol{C}_i(\boldsymbol{\alpha})\boldsymbol{F}_i^{1/2}(\boldsymbol{\beta})$, em que $\boldsymbol{C}_i$ é a matriz de correlação marginal e $\boldsymbol{F}_i$ é uma matriz diagonal contendo as variâncias marginais, $\boldsymbol{F}_{it}$, dadas por

$$\boldsymbol{F}_{it} = \text{diag}\left[\mu_{it1}(1-\mu_{it1}), \ldots, \mu_{it,J-1}(1-\mu_{it,J-1})\right].$$

Os blocos diagonais de $\boldsymbol{C}_i(\boldsymbol{\alpha})$ são $\boldsymbol{F}_{it}^{-1/2}\boldsymbol{V}_{it}\boldsymbol{F}_{it}^{-1/2}$, com $\boldsymbol{V}_{it} = \text{diag}(\boldsymbol{\mu}_{it}) - \boldsymbol{\mu}_{it}\boldsymbol{\mu}_{it}^T$; e os blocos fora da diagonal de $\boldsymbol{C}_i(\boldsymbol{\alpha})$ são $\boldsymbol{\rho}_{itt'} = \boldsymbol{\rho}_{itt'}(\boldsymbol{\alpha}) = Corr(\boldsymbol{Y}_{it}, \boldsymbol{Y}_{it'})$, que representam a matriz de correlação entre os vetores $\boldsymbol{Y}_{it}$ e $\boldsymbol{Y}_{it'}$, $t \neq t'$. Assim, o vetor $\boldsymbol{\alpha}$ é um vetor de parâmetros associado com um modelo para $\boldsymbol{\rho}_{itt'}$. Defina o vetor de resíduos padronizados de Pearson $\boldsymbol{e}_{it}$ como

$$\boldsymbol{e}_{it} = \boldsymbol{F}_{it}^{-1/2}\left(\boldsymbol{Y}_{it} - \boldsymbol{\mu}_{it}\right).$$

Logo, segue que

$$\boldsymbol{C}_{itt'}(\boldsymbol{\alpha}) = Corr(\boldsymbol{Y}_{it}, \boldsymbol{Y}_{it'}) = E(\boldsymbol{e}_{it}\boldsymbol{e}_{it'}^T).$$

Para reduzir a dimensão do vetor de correlações, Lipsitz *et al.* (1994) assumiram uma estrutura de correlação constante para todos os indivíduos e sugeriram o uso do método de momentos para estimação de uma variedade de estruturas para as matrizes de correlação. Exemplos incluem *independência*, *simetria composta* e *não estruturada*.

1.1.3.2 Razão de Chances Locais

Ao invés de usar o coeficiente de correlação, alguns autores modelaram a matriz $\boldsymbol{V}_{itt'}$ por meio de probabilidades conjuntas entre $\boldsymbol{Y}_{it}$ and $\boldsymbol{Y}_{it'}$, para $t \neq t'$. A covariância entre

Y_{itj} e $Y_{itj'}$ pode ser escrita como

$$\begin{aligned} Cov(Y_{itj}, Y_{it'j'}|\boldsymbol{X}, \boldsymbol{\beta}) &= E(Y_{itj}, Y_{it'j'}|\boldsymbol{X}, \boldsymbol{\beta}) - E(Y_{itj}|\boldsymbol{X}, \boldsymbol{\beta})E(Y_{it'j'}|\boldsymbol{X}, \boldsymbol{\beta}) \\ &= \mu_{itjt'j'} - \mu_{itj}\mu_{it'j'}. \end{aligned}$$

O produto dos primeiros momentos pode ser facilmente calculado pela especificação do modelo marginal em (1.1). Já as probabilidades conjuntas $\mu_{itjt'j'} = P(Y_{itj} = 1, Y_{it'j'} = 1|\boldsymbol{X}, \boldsymbol{\beta})$ podem ser modeladas através de um vetor paramétrico que descreve a estrutura de associação entre os níveis da resposta ordinal nos diferentes tempos. A estimação destas probabilidades conjuntas pode ser realizada através de razões de chances globais (veja, por exemplo, Lumley (1996) e Heagerty & Zeger (1996)), ou por meio de razões de chances locais (Touloumis *et al.*, 2013). Touloumis *et al.* (2013) argumentam que as razões de chances locais são a melhor parametrização no sentido de serem independentes da variação de $\boldsymbol{\beta}$ e produzirem probabilidades positivas únicas e válidas.

Para um par de respostas, considere uma tabela $J \times J$ no qual a célula (j, j') contém a probabilidade de resposta na linha j e na coluna j'. Seja $\theta_{tjt'j'}$ a razão de chances local para o ponto de corte (j, j') no par de tempos (t, t') e denote por $\boldsymbol{\alpha}$ o vetor de dimensão $L \times (J-1)^2$ consistindo das razões de chances locais, em que L é o número de pares de tempo. O vetor de associação $\boldsymbol{\alpha}$,

$$\boldsymbol{\alpha} = (\theta_{1121}, \ldots, \theta_{112(J-1)}, \ldots, \theta_{(T-1)1T1}, \ldots, \theta_{(T-1)(J-1)T(J-1)})^T,$$

pode ser estimado ajustando um modelo loglinear simultaneamente para as contagens obtidas de L tabelas de contingência marginalizadas e então calculando as razões de chances (Touloumis *et al.*, 2013). Como é de interesse apenas a estimação das razões de chances locais e não seus erros padrão, as tabelas são tratadas como independentes considerando distribuição Poisson para as contagens nas células.

Após impor restrições de identificabilidade, as razões de chances locais irrestritas obtidas do modelo de associação linha/coluna (Becker & Clogg, 1989) são determinadas por $\log(\theta_{tjt'j'}) = \varphi^{(t,t')}(\nu_j^{(t,t')} - \nu_{j+1}^{(t,t')})(\nu_{j'}^{(t,t')} - \nu_{j'+1}^{(t,t')})$, em que o parâmetro $\varphi^{(t,t')}$ mede a asso-

ciação média da tabela de contingência marginalizada e $\{\nu_j : j = 1, \ldots, J\}$ são os escores arbitrários para as J categorias da resposta. Para aumentar a parcimônia, escores com espaçamento de uma unidade são comumente adotados (isto é, $\nu_j^{(t,t')} = j$). Opções de estruturas de razões de chances locais incluem *uniforme*, *simetria composta* por tempo, *simetria composta* por categoria e *não estruturada*.

Condicionalmente a $\hat{\boldsymbol{\alpha}}$, e na especificação marginal em (1.1), as probabilidades conjuntas $\mu_{itjt'j'}$ podem ser estimadas com base na estrutura estimada das razões de chances locais usando o algoritmo IPFP (Iterative Proportional Fitting Procedure, Deming & Stephan (1940)). Touloumis *et al.* (2013) provaram que a solução do algoritmo IPFP preserva as razões de chance locais dos valores iniciais. Assim, é possível calcular a matriz de peso $\boldsymbol{V}_i$ e as equações de estimação em (1.2) podem ser obtidas com respeito a $\boldsymbol{\beta}$.

1.2 Dados Ausentes em Estudos Longitudinais

Embora a maioria dos estudos longitudinais seja delineada para coletar os dados de todos os indivíduos na amostra em cada instante de tempo, pode-se dizer que muitos estudos apresentam observações faltantes. Na área da saúde, por exemplo, os dados ausentes são uma regra, não uma exceção (Fitzmaurice *et al.*, 2004).

O perfil de observações incompletas em um estudo longitudinal pode exibir uma variedade de padrões. A perda pode ser *intermitente* (quando há uma ou mais perdas pontuais) ou *por abandono* (perda completa a partir de um instante de tempo). De acordo com Fitzmaurice *et al.* (2004), as implicações gerais para análise são: i) desbalanceamento no tempo, ii) perda de informação com consequente redução na eficiência ou um decréscimo na precisão com que mudanças na resposta média podem ser estimadas e, iii) introdução de viés nas estimativas dos parâmetros de interesse. Covariáveis ausentes causam complicações adicionais na análise.

Quando ocorrem observações ausentes, além da distribuição do processo de medida, o interesse também reside no mecanismo gerador dos dados faltantes. Rubin (1987) apresenta uma distinção entre os processos responsáveis pelos dados ausentes: *dados faltantes completamente ao acaso* (missing completely at random, MCAR), *dados faltantes ao acaso*

(missing at random, MAR), e *dados faltantes não aleatoriamente* (missing not at random, MNAR). Tais processos serão discutidos a seguir.

1.2.1 Mecanismos de Dados Ausentes

A determinação do mecanismo responsável pelos dados ausentes tem uma implicação decisiva na escolha e validade do método estatístico usado para analisar os dados (Donneau *et al.*, 2015a).

Supondo que os dados ausentes ocorrem apenas na resposta longitudinal, defina $\boldsymbol{R}_i$ como o vetor de indicadoras da resposta, $\boldsymbol{R}_i = (R_{i1}, R_{i2}, \ldots, R_{iT})^T$, com $R_{it} = 1$ se O_{it} é observada e $R_{it} = 0$ em caso contrário. Em geral, a distribuição de $\boldsymbol{R}$ pode estar relacionada com $\boldsymbol{O}$, e assim admitimos um modelo de probabilidade para $\boldsymbol{R}$, $Pr(\boldsymbol{R}|\boldsymbol{O}, \boldsymbol{X}, \boldsymbol{\psi})$, em que $\boldsymbol{\psi}$ denota os parâmetros do mecanismo gerador dos dados ausentes. Dado $\boldsymbol{R}_i$, o conjunto de respostas, $\boldsymbol{O}_i = (O_{i1}, \ldots, O_{iT})^T$, pode ser particionado em duas componentes $\boldsymbol{O}_i^o$ e $\boldsymbol{O}_i^m$, correspondendo às respostas observadas e aos dados ausentes, respectivamente. Uma hierarquia de três mecanismos de dados ausentes pode ser distinguida avaliando como $\boldsymbol{R}_i$ está relacionado com $\boldsymbol{O}_i$ (Rubin, 1987):

1. *Dados Faltantes Completamente ao Acaso* (MCAR): quando a não resposta é independente dos dados observados ou não observados, isto é:

$$Pr(\boldsymbol{R}|\boldsymbol{O}^o, \boldsymbol{O}^m, \boldsymbol{X}, \boldsymbol{\psi}) = Pr(\boldsymbol{R}|\boldsymbol{\psi}).$$

2. *Dados Faltantes ao Acaso* (MAR): quando a probabilidade de não resposta é independente dos dados não observados:

$$Pr(\boldsymbol{R}|\boldsymbol{O}^o, \boldsymbol{O}^m, \boldsymbol{X}, \boldsymbol{\psi}) = Pr(\boldsymbol{R}|\boldsymbol{O}^o, \boldsymbol{X}, \boldsymbol{\psi}).$$

3. *Dados Faltantes não Aleatoriamente* (MNAR): quando a probabilidade de não res-

posta depende dos dados não observados:

$$Pr(\boldsymbol{R}|\boldsymbol{O}^o, \boldsymbol{O}^m, \boldsymbol{X}, \boldsymbol{\psi}) = Pr(\boldsymbol{R}|\boldsymbol{O}^o, \boldsymbol{O}^m, \boldsymbol{X}, \boldsymbol{\psi}).$$

Quando a perda ocorre nas covariáveis, os mecanismos podem ser considerados de forma similar.

1.2.2 Implicações para a Análise

Considerando dados ausentes na variável resposta, Liang & Zeger (1986) discutem que as inferências obtidas pelo método GEE são corretas apenas sob perda MCAR, devido ao fato de que são baseadas em considerações frequentistas. Um exceção importante, mencionada pelos autores, ocorre quando a estrutura de correlação especificada é a correta, o que pode garantir a consistência do estimador GEE sob o mecanismo MAR. Contudo, isso não é verdadeiro quando a perda ocorre em uma covariável que é MAR dada a resposta (Carpenter & Kenward, 2013). Neste caso, mesmo métodos baseados em verossimilhança são inválidos e o mecanismo gerador dos dados ausentes deve ser modelado a fim de se obterem estimativas válidas dos parâmetros de regressão.

Quando os dados são MNAR, praticamente todos os métodos padrões de análise de dados longitudinais são inválidos (Fitzmaurice *et al.*, 2009). Para se obterem estimadores válidos é preciso a modelagem conjunta do vetor de respostas longitudinais e do mecanismo gerador dos dados ausentes.

1.3 Aplicações

A seguir são apresentados duas aplicações de dados reais que envolvem dados ausentes e que serão utilizados como ilustrações da metodologia proposta na tese. Na Subseção 1.3.1, apresentamos um estudo desenvolvido na Faculdade de Medicina da UFMG e no Hospital Odilon Berhrens, com o objetivo de comparar duas técnicas de analgesia no trabalho de parto. Tal estudo é caracterizado por um padrão de perda do tipo abandono total. Na Subseção 1.3.2, descrevemos um estudo sobre a evolução da insuficiência cardíaca em uma

coorte de pacientes com estenose mitral, uma doença caracterizada pelo estreitamento da válvula mitral, tratados no Hospital das Clínicas da UFMG. Para estes dados, um padrão de perda intermitente ocorre na resposta longitudinal.

1.3.1 Analgesia no Parto

Este estudo foi conduzido pela Faculdade de Medicina da UFMG e pelo Hospital Municipal Odilon Berhens, com o objetivo de comparar duas técnicas de analgesia no trabalho de parto. Um total de 49 pacientes foram acompanhadas durante todo o período até o parto. Uma avaliação em relação à intensidade da dor e medidas como pressão arterial, frequência cardíaca materna, infusão de ocitocina, nível de sedação, sinais de depressão respiratória, apnéia, dentre outras, foi feita inicialmente a cada 5 minutos, nos 30 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 (padrão-ouro), onde um analgésico local é utilizado. A segunda técnica, cuja eficiência será comparada àquela do padrão-ouro, é a infusão venosa contínua de remifentanil, um opióide que tem início de ação após 1 a 3 minutos. A intensidade da dor normalmente depende 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 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.

A resposta de interesse é a intensidade da dor, avaliada através de uma escala visual analógica (EVA), que consiste de uma linha reta, indicando-se em uma extremidade a marcação de "ausência de dor", e na outra, "pior dor imaginável". A resposta foi então codificada como: 1, dor leve e tolerável; 2: dor moderada e que causa desconforto; 3: dor intensa e insuportável). Para os objetivos de ilustração apenas três medidas $(0, 60, 90$ minutos) foram selecionadas para análise. As variáveis preditoras consideradas foram: grupo

(peridural ou remifentanil), idade da paciente, dilatação uterina, e infusão de ocitocina. A resposta e infusão de ocitocina não foram obtidos para 9 pacientes no tempo 60 e 18 pacientes no tempo 90. A não resposta se deve ao fato de o parto ter acontecido antes de 60 ou 90 minutos. Portanto, um mecanismo MAR parece uma suposição razoável para estes dados. As outras covariáveis foram completamente observadas.

1.3.2 Estenose Mitral

Os dados são oriundos de uma coorte de 164 pacientes com estenose mitral reumática que foram encaminhados para tratamento no Hospital das Clínicas da UFMG. Os pacientes foram incluídos antes da intervenção da válvula mitral e, em seguida, foram acompanhados no ambulatório a cada 4 meses de acordo com o seu estado clínico.

A estenose mitral é um estreitamento da válvula mitral no coração causada pela doença reumática, o que restringe o fluxo de sangue através da válvula. A principal manifestação clínica da doença é falta de ar, classificada em quatro categorias com base em quanto os pacientes são limitados durante a atividade física. A Classificação Funcional da *New York Heart Association* (NYHA) fornece uma maneira simples de classificar o grau de falta de ar. Os pacientes sem sintomas e sem limitação de atividade física ordinária foram classificados na classe I; limitação leve da atividade física determina a classe II; limitação acentuada da atividade física determina a classe III; e pacientes com limitações severas que resultam em incapacidade para realizar qualquer atividade física sem desconforto pertencem à classe IV.

Todos os pacientes passaram por valvuloplastia mitral percutânea (PMV), que é um tratamento eficaz para desobstruir a válvula mitral estenosada. Este procedimento é feito através da inserção de um cateter com um balão na sua ponta para abrir a válvula mitral estreitada. Este procedimento leva à melhora da classe funcional na maioria dos pacientes.

O objetivo do estudo é relacionar a evolução longitudinal da classe funcional com possíveis preditores. As covariáveis medidas são: complacência ventricular, ritmo cardíaco, características morfológicas da válvula mitral expressa como um escore ecocardiográfico, área da válvula mitral, pressão da artéria pulmonar, calcificação valvar, e sucesso do pro-

cedimento para abrir a válvula mitral sem complicações. Alguns valores da resposta não são observados para alguns indivíduos, devido principalmente a complicações no procedimento. Uma covariável basal de particular interesse, a complacência ventricular, não foi observada para cerca de um terço dos pacientes. As razões de não observar tal preditor incluem características morfológicas da válvula mitral e calcificação valvar. Deste modo, um mecanismo MAR parece razoável para descrever o mecanismo de não resposta.

1.4 Métodos para Tratar Dados Ausentes

Dois métodos comumente usados para lidar com dados ausentes em estudos longitudinais são a imputação múltipla (MIGEE) (Little & Rubin, 1987) e o método GEE ponderado (WGEE) (Robins *et al.*, 1995). No método MIGEE, os dados ausentes são imputados múltiplas vezes e uma estimativa final combinada é obtida levando-se em consideração a incerteza inerente à imputação. O método WGEE envolve a ponderação de cada observação pelo inverso da probabilidade de o dado ser observado. Assim, são dados pesos grandes para as observações coletadas cujas probabilidades estimadas de serem observadas são pequenas. A fim de se obterem estimativas consistentes é necessária a correta especificação do modelo de pesos (para o WGEE) ou do modelo de imputação (para o MIGEE). Um terceiro enfoque que combina estes dois métodos tem a propriedade de dupla robustez. Para consistência, ele requer a especificação correta de pelo menos um de seus modelos preditivos. Estas abordagens são introduzidas nas seções a seguir.

1.4.1 Imputação Múltipla

A Imputação Múltipla foi formalmente introduzida por Rubin (1978) e desde então tem se tornado um enfoque popular na análise de dados incompletos. A ideia chave é substituir cada valor ausente com um conjunto de M valores plausíveis amostrados da distribuição dos dados ausentes condicionada nos dados observados. Esta distribuição condicional reflete a incerteza sobre o verdadeiro valor a ser imputado. Então, um método GEE padrão é usado para analisar cada um dos M conjuntos de dados imputados, o que produz

M diferentes conjuntos de estimativas de parâmetros e erros padrões. Essas estimativas são então combinadas para fornecer uma única estimativa dos parâmetros de interesse, junto com erros padrões que refletem a incerteza inerente à imputação dos dados não observados. Uma característica importante dos métodos de imputação múltipla é que os modelos de imputação e análise podem ser considerados, de certo modo, separadamente, embora deva haver alguma compatibilidade entre eles (Schafer, 1999).

Um método de imputação comumente usado para lidar com respostas não gaussianas é a imputação via *equações encadeadas* (van Buuren *et al.* (1999), van Buuren (2007)), método comumente referido como *especificação condicional completa* (FCS). Começando com uma imputação, o método FCS gera imputações iterando as densidades condicionais. Tal método é atrativo pois pode tratar dados de diferentes tipos, tais como binários ou ordinais, usando um modelo de regressão univariado apropriado para cada variável com valores não observados.

Denote por $\tilde{\boldsymbol{\beta}}_m$ e $\tilde{\boldsymbol{U}}_m$, respectivamente, as estimativas do vetor $\boldsymbol{\beta}$ e de sua matriz de covariância resultante da análise GEE do m-ésimo conjunto de dados imputado, ($m = 1, \ldots, M$). Segundo Rubin (1987), a estimativa pontual para o parâmetro de interesse $\boldsymbol{\beta}$ é simplesmente a média das M estimativas pontuais da imputação

$$\hat{\boldsymbol{\beta}}_{MI} = \frac{1}{M}\sum_{m=1}^{M}\tilde{\boldsymbol{\beta}}_m,$$

e uma estimativa da matriz de covariâncias de $\hat{\boldsymbol{\beta}}_{MI}$ é dada por

$$\widehat{\boldsymbol{U}}_{MI} = \tilde{\boldsymbol{W}} + \left(\frac{M+1}{M}\right)\tilde{\boldsymbol{B}},$$

em que

$$\tilde{\boldsymbol{W}} = \frac{1}{M}\sum_{m=1}^{M}\tilde{\boldsymbol{U}}_m \quad \text{e} \quad \tilde{\boldsymbol{B}} = \frac{1}{M-1}\sum_{m=1}^{M}(\tilde{\boldsymbol{\beta}}_m - \hat{\boldsymbol{\beta}}_{\boldsymbol{MI}})(\tilde{\boldsymbol{\beta}}_m - \hat{\boldsymbol{\beta}}_{\boldsymbol{MI}})^T.$$

Note que $\hat{\boldsymbol{U}}_{MI}$ é uma combinação entre a variabilidade intra imputação ($\tilde{\boldsymbol{W}}$) e entre as imputações ($\tilde{\boldsymbol{B}}$). Dado que o modelo da imputação seja correto, as estimativas resultantes

são consistentes (Schafer, 1997). Detalhes sobre o procedimento podem ser encontrados em: Schafer (1997), Little & Rubin (2002) e Carpenter & Kenward (2013).

No contexto de respostas longitudinais ordinais, um artigo recente de Donneau *et al.* (2015a) compara, via estudos de simulação, dois métodos de imputação (imputação sob um modelo normal multivariado e imputação sob um modelo ordinal) para respostas sujeitas a perda por abandono total. Em outro trabalho, os mesmos autores comparam a modelagem conjunta sob um modelo normal e o enfoque FCS para perda intermitente na resposta (Donneau *et al.*, 2015b).

1.4.2 GEE Ponderado

Robins *et al.* (1995) propuseram uma classe de estimadores GEE ponderados válidos sob perda MAR do tipo abandono total. A ideia principal é ponderar a contribuição de cada indivíduo no GEE pelo inverso da probabilidade de tal indivíduo ser completamente observado. Assim, cada indivíduo que permanece no estudo é representativo de si mesmo e também de indivíduos similares que já saíram do estudo. A incorporação dos pesos reduz o possível viés nas estimativas dos parâmetros de regressão (Beunckens *et al.*, 2008).

Seja π_{it} a probabilidade de o dado ser observado até (e incluindo) a ocasião t. O modelo para o abandono total é da forma $\pi_{it} = (1 - \lambda_{i1}) \times \ldots \times (1 - \lambda_{it})$, em que $\lambda_{it} = \lambda_{it}(\boldsymbol{\psi}) = P(R_{it} = 0 | \overline{\boldsymbol{O}}_{i,t-1}, \boldsymbol{X}_i, R_{i,t-1} = 1)$ é o risco de abandono total no tempo t, e $\overline{\boldsymbol{O}}$ é o histórico da resposta ordinal até o tempo $t-1$. Geralmente, assume-se que não há observações perdidas no primeiro tempo e se modela λ_{it} por um modelo logístico. Os preditores em tal modelo podem incluir funções dos históricos das respostas observadas assim como qualquer informação adicional de covariáveis que ajude a predizer as respostas não observadas e tornem a suposição MAR plausível.

As equações de estimação generalizadas ponderadas (WGEE) para $\boldsymbol{\beta}$ são dadas por

$$\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\psi}) = \sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} \boldsymbol{\Delta}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}, \tag{1.3}$$

em que $\boldsymbol{\Delta}_i$ é uma matriz diagonal com elementos R_{it}/π_{it}. Assim, cada componente do

vetor de resíduos observado $(\boldsymbol{Y}_{it} - \boldsymbol{\mu}_{it})$ é ponderado por π_{it}^{-1}. Sob perda MAR e dada a correta especificação do modelo para os dados ausentes, uma estimativa consistente para $\boldsymbol{\beta}$ pode ser obtida resolvendo (1.3) (Robins *et al.*, 1995).

No contexto de respostas longitudinais ordinais, Toledano & Gatsonis (1999) desenvolveram um método GEE ponderado para acomodar padrões arbitrários de perda MCAR na resposta e perda em uma covariável sujeita ao mecanismo MAR. No caso de respostas binárias, Chen & Zhou (2011) estenderam o método de Robins *et al.* (1995) para acomodar um padrão arbitrário de perdas tanto na resposta como em uma covariável.

Uma desvantagem dos estimadores ponderados é que eles podem ser instáveis em situações em que alguns indivíduos recebem grandes pesos. Isso ocorre quando algumas covariáveis são altamente preditivas das indicadoras de não resposta (Vansteelandt *et al.*, 2010). Já que no WGEE todos os indivíduos recebem pesos, qualquer especificação incorreta do modelo de dados ausentes afetará todos os indivíduos incluídos na análise (Beunckens *et al.*, 2008). Por outro lado, uma especificação errada no modelo de imputação afetará apenas a porção imputada dos dados (Schafer, 1997). Comparações entre imputação múltipla e métodos de ponderação inversa podem ser encontradas em Carpenter *et al.* (2006), Beunckens *et al.* (2008), e Birhanu *et al.* (2011).

1.4.3 Estimadores GEE Duplamente Robustos

Alguns autores (veja, por exemplo, Scharfstein *et al.* (1999), Tsiatis (2006)) notaram que ao adicionar um termo de esperança zero, digamos $\phi(\cdot)$, aos estimadores de probabilidade inversa ainda resultaria em estimativas consistentes sob um mecanismo MAR. As soluções dessas equações de estimação aumentadas dão origem aos chamados estimadores *duplamente robustos*. A escolha mais eficiente de $\phi(\cdot)$ depende do padrão da não resposta e envolve uma esperança com respeito à distribuição das quantidades ausentes condicional às quantidades observadas. Dado que o mecanismo de perda seja MAR, os estimadores são duplamente robustos no sentido de que são consistentes para $\boldsymbol{\beta}$ se o modelo de não resposta ou o modelo para a esperança condicional (ou ambos) estejam corretamente especificados (Tsiatis, 2006).

O uso de estimadores duplamente robustos em cenários longitudinais é tratado em Bang & Robins (2005), Seaman & Copas (2009), Chen & Zhou (2011) e Birhanu *et al.* (2011). Tais desenvolvimentos, contudo, têm focado primariamente em respostas binárias e, até onde temos conhecimento, não foram investigados com dados longitudinais ordinais.

1.5 Contribuições e Organização da Tese

A presente tese contribui com a literatura de dados ausentes em estudos longitudinais quando a resposta de interesse é ordinal. Um estimador GEE duplamente robusto é proposto para situações em que há perda intermitente na resposta e também em uma covariável, ambas sujeitas ao mecanismo MAR. O caso abordado de padrão arbitrário de perda na resposta engloba a situação comumente discutida de perda por abandono do estudo. A associação entre as medidas repetidas é avaliada estendendo os enfoques de correlação de Lipsitz *et al.* (1994) e de razões de chances locais de Touloumis *et al.* (2013) para o caso de dados ausentes. A incorporação de uma covariável potencialmente ausente nas análises é de particular interesse, pois permite a recuperação de informação valiosa que seria descartada por métodos usuais de análise.

O Capítulo 2 apresenta o artigo *Doubly Robust-Based Generalized Estimating Equations for the Analysis of Longitudinal Ordinal Missing Data*. Nesse trabalho, um estimador duplamente robusto é proposto para análise de dados longitudinais com resposta ordinal quando tanto a resposta quanto uma covariável podem apresentar um padrão arbitrário de não resposta sujeita a um mecanismo MAR. Um modelo de chances proporcionais é adotado para a resposta longitudinal e equações de estimação independentes são usadas para estimação dos parâmetros de interesse. Focando no viés das estimativas, o desempenho do estimador proposto é comparado com o enfoque FCS de imputação múltipla e com um estimador ponderado, adaptado do trabalho de Chen & Zhou (2011) para o caso ordinal. Resultados de simulação indicam, em geral, para os cenários considerados, melhor desempenho do método proposto em comparação com os concorrentes MIGEE e WGEE. O método é aplicado a um conjunto de dados reais oriundo do estudo de Analgesia no Parto, cujo objetivo era comparar duas técnicas de analgesia com relação à intensidade

da dor do trabalho de parto em 49 pacientes de Minas Gerais, Brasil.

O Capítulo 3 apresenta o artigo *Modeling the Association Structure in DRGEE for Longitudinal Ordinal Missing Data*. O foco desse artigo está na eficiência dos estimadores ao se modelar a associação entre as medidas repetidas. São consideradas duas parametrizações para a matriz de covariância: a abordagem de Lipsitz *et al.* (1994) que faz uso do coeficiente de correlação e o enfoque de razões de chances locais de Touloumis *et al.* (2013). As diferenças na implementação e consistência das estimativas sob perda MAR são discutidas. Resultados de simulação indicam que, para covariáveis dependentes do tempo, pode-se melhorar consideravelmente a eficiência das estimativas com a modelagem da estrutura de dependência das medidas repetidas. Para os cenários considerados, os resultados também indicam superioridade da abordagem de razões de chances locais em termos de convergência do algoritmo e viés das estimativas em pequenas amostras. Os métodos são aplicados a um banco de dados reais de uma coorte de 164 pacientes com estenose mitral, doença cardíaca valvar caracterizada pelo estreitamento do orifício da válvula mitral do coração. A resposta de interesse é a classe funcional, maior determinante de qualidade de vida e sobrevida do indivíduo. O objetivo do estudo é caracterizar a extensão da insuficiência cardíaca ao longo do tempo.

REFERÊNCIAS BIBLIOGRÁFICAS

Agresti, Alan. 2013. *Categorical Data Analysis*. 3 edn. John Wiley & Sons.

Bang, Heejung, & Robins, James M. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61**(4), 962–973.

Becker, Mark P, & Clogg, Clifford C. 1989. Analysis of sets of two-way contingency tables using association models. *Journal of the American Statistical Association*, **84**(405), 142–151.

Beunckens, Caroline, Sotto, Cristina, & Molenberghs, Geert. 2008. A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational Statistics & Data Analysis*, **52**(3), 1533–1548.

Birhanu, Teshome, Molenberghs, Geert, Sotto, Cristina, & Kenward, Michael G. 2011. Doubly robust and multiple-imputation-based generalized estimating equations. *Journal of Biopharmaceutical Statistics*, **21**(2), 202–225.

Carey, Vincent, Zeger, Scott L, & Diggle, Peter. 1993. Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, **80**(3), 517–526.

Carpenter, James R., & Kenward, Michael G. 2013. *Multiple Imputation and its Application*. Wiley & Sons.

Carpenter, James R, Kenward, Michael G, & Vansteelandt, Stijn. 2006. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **169**(3), 571–584.

Chen, Baojiang, & Zhou, Xiao-Hua. 2011. Doubly robust estimates for binary longitudinal data analysis with missing response and missing covariates. *Biometrics*, **67**(3), 830–842.

Deming, W Edwards, & Stephan, Frederick F. 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, **11**(4), 427–444.

Donneau, Anne-Françoise, Mauer, Murielle, Molenberghs, Geert, & Albert, Adelin. 2015a. A simulation study comparing multiple imputation methods for incomplete longitudinal ordinal data. *Communications in Statistics-Simulation and Computation*, **44**(5), 1311–1338.

Donneau, Anne-Françoise, Mauer, Murielle, Lambert, Philippe, Molenberghs, Geert, & Albert, Adelin. 2015b. Simulation-based study comparing multiple imputation methods for non-monotone missing ordinal data in longitudinal settings. *Journal of Biopharmaceutical Statistics*, **25**(3), 570–601.

Fitzmaurice, G., Davidian, M., Molenberghs, G., & Verbeke, G. 2009. *Longitudinal Data Analysis*. Handbooks of Modern Statistical Methods. Chapman & Hall/CRC.

Fitzmaurice, Garrett M., Laird, M., & Ware, James H. 2004. *Applied Longitudinal Analysis*. Wiley-Interscience.

Goodman, Leo A. 1985. The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *The Annals of Statistics*, **13**(1), 10–69.

Heagerty, Patrick J, & Zeger, Scott L. 1996. Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association*, **91**(435), 1024–1036.

Li, Lingling, Shen, Changyu, Li, Xiaochun, & Robins, James M. 2013. On weighting approaches for missing data. *Statistical Methods in Medical Research*, **22**(1), 14–30.

Liang, Kung-Yee, & Zeger, Scott L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22.

Lipsitz, Stuart R, Laird, Nan M, & Harrington, David P. 1991. Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, **78**(1), 153–160.

Lipsitz, Stuart R, Kim, Kyungmann, & Zhao, Lueping. 1994. Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine*, **13**(11), 1149–1163.

Little, Roderick JA, & Rubin, Donald B. 1987. *Statistical Analysis with Missing Data*. 1 edn. Wiley New York.

Little, Roderick JA, & Rubin, Donald B. 2002. *Statistical Analysis with Missing Data*. 2 edn. Wiley New York.

Lumley, Thomas. 1996. Generalized estimating equations for ordinal data: a note on working correlation structures. *Biometrics*, **52**(1), 354–361.

McCullagh, Peter. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, **42**(2), 109–142.

Molenberghs, Geert, & Kenward, Michael G. 2010. Semi-parametric marginal models for hierarchical data and their corresponding full models. *Computational Statistics & Data Analysis*, **54**(2), 585–597.

Molenberghs, Geert, & Verbeke, Geert. 2005. *Models for discrete longitudinal data*. 1 edn. Springer Series in Statistics. Springer-Verlag New York.

Nooraee, Nazanin, Molenberghs, Geert, & van den Heuvel, Edwin R. 2014. GEE for longitudinal ordinal data: Comparing R-geepack, R-multgee, R-repolr, SAS-GENMOD, SPSS-GENLIN. *Computational Statistics & Data Analysis*, **77**, 70–83.

Parsons, Nicholas R, Edmondson, RN, & Gilmour, SG. 2006. A generalized estimating equation method for fitting autocorrelated ordinal score data with an application in horticultural research. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **55**(4), 507–524.

Parsons, Nick R, Costa, Matthew L, Achten, Juul, & Stallard, Nigel. 2009. Repeated measures proportional odds logistic regression analysis of ordinal score data in the statistical software package R. *Computational Statistics & Data Analysis*, **53**(3), 632–641.

Peterson, Bercedis, & Harrell Jr, Frank E. 1990. Partial proportional odds models for ordinal response variables. *Applied Statistics*, **39**(2), 205–217.

Pierce, Donald A. 1982. The asymptotic effect of substituting estimators for parameters in certain types of statistics. *The Annals of Statistics*, **10**(2), 475–478.

Poleto, Frederico Z, Singer, Julio M, & Paulino, Carlos Daniel. 2014. A product-multinomial framework for categorical data analysis with missing responses. *Brazilian Journal of Probability and Statistics*, **28**(1), 109–139.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robins, James M, Rotnitzky, Andrea, & Zhao, Lue Ping. 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**(429), 106–121.

Rubin, Donald B. 1976. Inference and missing data. *Biometrika*, **63**(3), 581–592.

Rubin, Donald B. 1978. Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. *Pages 20–34 of: Proceedings of the survey research methods section of the American Statistical Association*, vol. 1. American Statistical Association.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. John Wiley New York.

Schafer, Joseph L. 1997. *Analysis of incomplete multivariate data*. Monographs on Statistics & Applied Probability. Chapman and Hall/CRC.

Schafer, Joseph L. 1999. Multiple imputation: a primer. *Statistical Methods in Medical Research*, **8**(1), 3–15.

Scharfstein, Daniel O, Rotnitzky, Andrea, & Robins, James M. 1999. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, **94**(448), 1096–1120.

Seaman, Shaun, & Copas, Andrew. 2009. Doubly robust generalized estimating equations for longitudinal data. *Statistics in Medicine*, **28**(6), 937–955.

Toledano, Alicia Y, & Gatsonis, Constantine. 1999. Generalized estimating equations for ordinal categorical data: arbitrary patterns of missing responses and missingness in a key covariate. *Biometrics*, **55**(2), 488–496.

Touloumis, Anestis. 2015. *SimCorMultRes: Simulates Correlated Multinomial Responses*. R package version 1.3.0.

Touloumis, Anestis, Agresti, Alan, & Kateri, Maria. 2013. GEE for multinomial responses using a local odds ratios parameterization. *Biometrics*, **69**(3), 633–640.

Tsiatis, Anastasios A. 2006. *Semiparametric theory and missing data*. Springer Series in Statistics. Springer New York.

van Buuren, S., Boshuizen, H. C., & Knook, D. L. 1999. Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. *Statistics in Medicine*, **18**(6), 681–694.

van Buuren, Stef. 2007. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, **16**(3), 219–242.

van Buuren, Stef, & Groothuis-Oudshoorn, Karin. 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, **45**(3), 1–67.

Vansteelandt, Stijn, Carpenter, James, & Kenward, Michael G. 2010. Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, **6**(1), 37–48.

CHAPTER 2

ARTIGO 1: DOUBLY ROBUST-BASED GENERALIZED ESTIMATING EQUATIONS FOR THE ANALYSIS OF LONGITUDINAL ORDINAL MISSING DATA

José Luiz P. da Silva, Enrico A. Colosimo, and Fábio N. Demarqui

Abstract: *Generalized Estimation Equations (GEE) are a well-known method for the analysis of non-Gaussian longitudinal data. This method has computational simplicity and parameters with a population-averaged interpretation. However, in the presence of missing data, this estimator is only consistent under the strong assumption of missing completely at random (MCAR). Some corrections can be done when the missing data mechanism is missing at random (MAR): inverse probability weighting (WGEE) and multiple imputation (MIGEE). In order to obtain consistent estimates, it is necessary the correct specification of the weight model for WGEE or the imputation model for the MIGEE. A recent method combining ideas of these two approaches has the doubly robust property. For consistency, it requires only the weight or the imputation model to be correct. In this work, it is assumed a proportional odds model and it is proposed a doubly robust estimator for the analysis of ordinal longitudinal data with intermittently missing response and covariate under the MAR mechanism. Simulation results revealed better performance of the proposed method compared to WGEE and MIGEE. The method is applied to a data set related to Analgesia in Childbirth study.*

Keywords: *Missing at random; Multiple imputation; Proportional Odds Model; Weighted GEE.*

2.1 Introduction

The Generalized Estimating Equation (GEE) method (Liang & Zeger, 1986) is one of the most popular approaches for the analysis of non-Gaussian correlated data. Its main advantage resides in the fact that one is only required to correctly specify the mean structure of the response in order to ensure that the parameter estimator is consistent and asymptotically normal. In its basic formulation, the association parameters among repeated measures were taken as nuisance parameters. The GEE method has computational simplicity (there is no need to deal with complex, and in some cases, intractable likelihoods) and it further allows a population-averaged interpretation of the parameters of interest.

It is very common for sets of longitudinal data to be incomplete, in the sense that not all planned observations are actually observed. This problem is pervasive in longitudinal data because nonresponse can occur at any time from the beginning of the study. Two patterns of missing data can be observed for the response: (1) dropout, when a subject leaves the study prematurely for reasons beyond the control of the investigator, leading to a monotone pattern of nonresponse, or (2) intermittent nonresponse, in which a subject returns to the study after some occasions of nonresponse. Covariates may also be missing, leading to limitations in data analysis. In the presence of missing data, three issues are of main concern: (1) potential serious bias due to systematic differences between the observed data and the missing data, (2) complications in data handling and statistical inferences, and (3) loss of efficiency. Therefore, in order to make correct inferences it is fundamental to know the missing data mechanism generating the nonresponse and how to handle it.

Little & Rubin (1987) provided a formal framework to deal with missing data by defining the commonly adopted taxonomy of missing data mechanisms. A nonresponse process is said to be *missing completely at random* (MCAR) if missingness is independent of both unobserved and observed data, and *missing at random* (MAR) if, conditional on the observed data, the missingness is independent of the unobserved data. When the nonresponse process depends on unobserved quantities it is said to be *missing not at*

random (MNAR).

When data are incomplete, GEE suffers from its frequentist nature and is, in its basic form, the estimator is consistent only under MCAR mechanisms (Liang & Zeger, 1986). The first effort to make GEE applicable to the more realistic MAR scenario was via Multiple Imputation (MIGEE), proposed by Little & Rubin (1987), in which the missing portions of data are multiply imputed taking into account the uncertainty associated with the predicted values. The completed data sets are analyzed by standard methods for complete data, and estimates are combined in a final analysis. Multiple imputation is detailed in the books by Schafer (1997), Little & Rubin (2002) and Carpenter & Kenward (2013). Later, Robins *et al.* (1995) proposed the Weighted Generalized Estimating Equations (WGEE), which consists in weighting each observation by the inverse of the probability of the data being observed. This method produces consistent estimates provided the weight model is correctly specified.

Doubly robust estimators (DRGEE) arise as a third generalization of ordinary GEE to deal with data subject to MAR mechanisms. Doubly robust methods have received increasing attention in the literature in the last decade (see Carpenter *et al.* (2006), Bang & Robins (2005), Tsiatis (2006), Seaman & Copas (2009), Chen & Zhou (2011)). The main idea is to supplement the WGEE with a predictive model for the missing quantities conditionally on the observed ones. Doubly Robust methods only require that the dropout or the conditional model to be correctly specified in order to provide consistent estimates. In the analysis of longitudinal binary data, doubly robust methods have been applied by Seaman & Copas (2009), Birhanu *et al.* (2011), for missing responses and by Chen & Zhou (2011) for intermittently missing response and a single missing covariate.

The literature on GEE for missing data is comparatively scarce for longitudinal ordinal response. With the objective of analyzing categorical data with MAR and MNAR response mechanisms, Poleto *et al.* (2014) considered a product-multinomial framework. In Toledano & Gatsonis (1999), the authors used a weighted GEE method to accommodate arbitrary patterns of a MCAR missing response and missingness in a key covariate subject to a MAR mechanism. A recent paper from Donneau *et al.* (2015a) compared

two multiple imputation methods (multivariate normal imputation and ordinal imputation regression) for longitudinal ordinal data subject to dropout. In another paper, the same authors compared the joint modeling and fully conditional specification approaches for non-monotone missingness (Donneau *et al.*, 2015b). The above mentioned papers used single robust versions of GEE and they have treated only a missing MAR baseline covariate or missing MAR response. Thus the use of a doubly robust GEE method for ordinal data with simultaneously intermittently missing response and missing covariate has been in need of further development.

This work was motivated by the Analgesia in Childbirth study conducted in Minas Gerais state, Brazil. The main objective of that study was to compare two techniques of analgesia for labor pain in 49 patients. The response, pain intensity, was subjectively assessed by each patient, and various clinical covariates were observed until delivery. Response and a particular covariate (infusion of oxytocin) were missing for some patients and the MAR mechanism seems to be a reasonable assumption for this data.

In the current paper, it is proposed a doubly robust approach for the analysis of longitudinal ordinal data with intermittently missing response and covariate that is MAR. To our knowledge, the proposed methodology is new to the GEE literature of missing ordinal data, can be used for handling arbitrary patterns of missing data in the ordinal response and missingness in a key covariate, as those frequently arising in medical studies.

The paper is organized as follows. In Section 2.2 is defined the notation for GEE with fully observed data and missing data mechanisms. Section 2.3 outlines the WGEE and MIGEE approaches. The proposed methodology is detailed in Section 2.4. A simulation study is presented in Section 2.5, in which the finite-sample biases and standard errors obtained via the standard GEE, MIGEE, WGEE and doubly robust versions are compared. Data arising from the Analgesia in Childbirth study are analyzed in Section 2.6. The paper ends with a discussion and future research directions in Section 2.7.

2.2 GEE for Complete Data and Missing Data Assumptions

In this section it is introduced the generalized estimating equations for the analysis of fully observed ordinal data. Subsection 2.2.1 establishes the model and notation for longitudinal ordinal data. Subsection 2.2.2 presents a series of assumptions related to the missing data mechanisms that must be considered in order to build consistent estimators.

2.2.1 GEE for Longitudinal Ordinal Response

Let $O_{it} \in \{1, 2, \ldots, J\}$ be the ordinal response for subject i $(i = 1, \ldots, n)$ at time t $(t = 1, \ldots, T_i,\ T_i \leq T)$. As the response has J levels it can be defined $Y_{itj} = I(O_{it} = j)$ for $j = 1, \ldots, J$, where $I(A)$ denotes the indicator function. Y_{itj} is converted into the equivalent $(J-1)$-variate vector $\boldsymbol{Y}_{it} = (Y_{it1}, \ldots, Y_{it(J-1)})^T$ and let $\boldsymbol{Y}_i = (Y_{i1}^T, \ldots, Y_{iT_i}^T)^T$ be the stacked response vector. When $J = 2$ the response is binary and $\boldsymbol{Y}_{it}$ is a scalar. Let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iT_i})^T$ denote the covariate vector that may be missing and $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}^T, \ldots, \boldsymbol{Z}_{iT_i}^T)^T$ the covariate vector that is always observed, where $\boldsymbol{Z}_{it}$ is the covariate vector for subject i at time t.

The marginal distribution of $\boldsymbol{Y}_{it}$ is assumed to be multinomial (with sample size $\sum_{j=1}^{J} Y_{itj} = 1$), that is

$$f(\boldsymbol{Y}_{it}|X_{it}, \boldsymbol{Z}_{it}; \boldsymbol{\beta}) = \prod_{j=1}^{J} \mu_{itj}^{y_{itj}}, \tag{2.1}$$

where $\mu_{itj} = \mu_{itj}(\boldsymbol{\beta}) = E(Y_{itj}|X_{it}, \boldsymbol{Z}_{it}; \boldsymbol{\beta}) = Pr(O_{it} = j|X_{it}, \boldsymbol{Z}_{it}; \boldsymbol{\beta})$, is the probability of response j at time t for individual i, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters. In this work, a cumulative logit link is used for modeling μ_{itj}, that is,

$$\text{logit}\left[Pr(O_{it} \leq j|X_{it}, \boldsymbol{Z}_{it}, \boldsymbol{\beta})\right] = \beta_{0j} + X_{it}\beta_x + \boldsymbol{Z}_{it}^T\boldsymbol{\beta}_z, \quad j = 1, \ldots, J-1, \tag{2.2}$$

which implies a proportional odds model (McCullagh, 1980). In such model the interpretation of $\boldsymbol{\beta}$ is the same regardless of the number of categories (i.e., it is invariant to the combination of categories). A desired feature of this model is that the exponential of the parameters is interpreted as an odds ratio (Agresti, 2013).

The main interest is to make inferences related to the regression parameters $\boldsymbol{\beta} = (\beta_{01}, \ldots, \beta_{0,J-1}, \beta_x, \boldsymbol{\beta}_z^T)^T$ associated with the $(J-1) \times 1$ marginal probability vectors

$$E(\boldsymbol{Y}_{it}|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) = \boldsymbol{\mu}_{it}(\boldsymbol{\beta}) = (\mu_{it1}, \ldots, \mu_{it(J-1)})^T.$$

These marginal probability vectors are grouped to form a vector $E(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) = \boldsymbol{\mu}_i = (\boldsymbol{\mu}_{i1}^T, \ldots, \boldsymbol{\mu}_{iT_i}^T)^T$ with the same dimension of $\boldsymbol{Y}_i$.

In order to estimate $\boldsymbol{\beta}$, the generalized estimation equations are used (Liang & Zeger (1986), Lipsitz *et al.* (1994)), which takes the form

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}, \tag{2.3}$$

where $\boldsymbol{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T}$ and $\boldsymbol{V}_i = \boldsymbol{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is a $T_i(J-1) \times T_i(J-1)$ "working covariance" matrix, usually decomposed into the form $\boldsymbol{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{F}_i^{1/2}(\boldsymbol{\beta}) \boldsymbol{C}_i(\boldsymbol{\alpha}) \boldsymbol{F}_i^{1/2}(\boldsymbol{\beta})$, where $\boldsymbol{F}_i$ is a matrix containing the marginal variances,

$$\boldsymbol{F}_{it} = \operatorname{diag}\left[\mu_{it1}(1-\mu_{it1}), \ldots, \mu_{it,J-1}(1-\mu_{it,J-1})\right],$$

and $\boldsymbol{C}_i$ the marginal correlation matrix. The $(J-1) \times (J-1)$ diagonal blocks of $\boldsymbol{C}_i(\boldsymbol{\alpha})$ are $\boldsymbol{F}_{it}^{-1/2} \boldsymbol{V}_{it} \boldsymbol{F}_{it}^{-1/2}$, with $\boldsymbol{V}_{it} = \operatorname{diag}(\boldsymbol{\mu}_{it}) - \boldsymbol{\mu}_{it}\boldsymbol{\mu}_{it}^T$; and the $(J-1) \times (J-1)$ off-diagonal blocks of $\boldsymbol{C}_i(\boldsymbol{\alpha})$ are $\boldsymbol{\alpha}_{itt'}$, which represents the correlation matrix between $\boldsymbol{Y}_{it}$ and $\boldsymbol{Y}_{it'}$, $t \neq t'$ (Lipsitz *et al.*, 1994).

Under mild regularity conditions and the correct specification of the marginal mean model in (2.2), Liang and Zeger (1986) proved that the estimator $\hat{\boldsymbol{\beta}}$, obtained by solving (2.3), is consistent and $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a p-variate normal distribution with mean $\boldsymbol{0}$ and covariance matrix

$$\mathrm{V}_{\boldsymbol{\beta}} = \lim_{n \to \infty} n \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_0^{-1}, \tag{2.4}$$

where $\boldsymbol{\Sigma}_0 = \sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} \boldsymbol{D}_i^T$, and $\boldsymbol{\Sigma}_1 = \sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} \mathrm{Cov}(\boldsymbol{Y}_i) \boldsymbol{V}_i^{-1} \boldsymbol{D}_i^T$. In practice, the

"sandwich" covariance matrix $V_{\boldsymbol{\beta}}$ in (2.4) is calculated by ignoring the limit and replacing $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and $\text{Cov}(\boldsymbol{Y}_i)$ by $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ and $(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)^T$, respectively (Touloumis *et al.*, 2013).

2.2.2 Missing Data Framework

For each occasion t it can be defined $R_{it} = 0$ if O_{it} and X_{it} are missing, $R_{it} = 1$ if O_{it} is missing and X_{it} is observed, $R_{it} = 2$ if O_{it} is observed and X_{it} is missing, and $R_{it} = 3$ if O_{it} and X_{it} are both observed. Let $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{iT_i})^T$ and $\bar{\boldsymbol{R}}_{it} = (R_{i1}, \ldots, R_{i,t-1})$.

By specifying conditional models of the form $Pr(R_{it} = r_{it}|\bar{\boldsymbol{R}}_{it}, \boldsymbol{O}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ it can be obtained $Pr(\boldsymbol{R}_i = \boldsymbol{r}_i|\boldsymbol{O}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ through $\prod_{i=2}^{T_i} Pr(R_{it} = r_{it}|\bar{\boldsymbol{R}}_{it}, \boldsymbol{O}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) Pr(R_{i1} = r_{i1}|\boldsymbol{O}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$. Let $\lambda_{itk} = Pr(R_{it} = k|\bar{\boldsymbol{R}}_{it}, \boldsymbol{O}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$, for $k = 0, 1, 2, 3$. This general formulation encompasses MCAR, MAR and MNAR mechanisms. In particular, the MAR mechanism requires

$$Pr(\boldsymbol{R}_i = \boldsymbol{r}_i|\boldsymbol{O}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) = Pr(\boldsymbol{R}_i = \boldsymbol{r}_i|\boldsymbol{O}_i^o, \boldsymbol{X}_i^o, \boldsymbol{Z}_i), \tag{2.5}$$

where $\boldsymbol{O}_i^o$ and $\boldsymbol{X}_i^o$ denote the observed components of $\boldsymbol{O}_i$ and $\boldsymbol{X}_i$, respectively. Because $\boldsymbol{R}_i$ is modeled through a product of conditional models, it is natural to make the following additional assumption (Chen & Zhou, 2011):

$$Pr(R_{it} = r_{it}|\bar{\boldsymbol{R}}_{it}, \boldsymbol{O}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) = Pr(R_{it} = r_{it}|\bar{\boldsymbol{R}}_{it}, \bar{\boldsymbol{O}}_{it}^o, \bar{\boldsymbol{X}}_{it}^o, \boldsymbol{Z}_i), \tag{2.6}$$

for each time t, where $\bar{\boldsymbol{O}}_{it}^o$ and $\bar{\boldsymbol{X}}_{it}^o$ are the histories of observed responses and covariates up to time $t-1$.

Let $\pi_{it} = Pr(R_{it} = 3|\boldsymbol{O}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ be the marginal probability of observing both $\boldsymbol{O}_i$ and $\boldsymbol{X}_i$ at time t, given the entire vectors of responses and covariates. Then, π_{it} is expressed as

$$\pi_{it} = \sum_{r_{i1}, \ldots, r_{i,t-1}} Pr(R_{it} = 3, R_{i,t-1} = r_{i,t-1}, \ldots, R_{i1} = r_{i1}|\boldsymbol{O}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i).$$

This marginal probability can be expressed in terms of the conditional probabilities λ_{itk}'s.

Throughout this paper the so-called *positivity assumption* is required, that is, π_{it} must be bounded away from zero. This condition is needed in order to guarantee the existence of $\sqrt{n}$-consistent estimators of $\boldsymbol{\beta}$ (Robins *et al.*, 1995).

2.3 Available Approaches for Missing Data

Multiple imputation and weighted generalized estimation equations are two commonly used methods available for missing data under the MAR mechanism. These methods are presented in Subsections 2.3.1 and 2.3.2, respectively. They serve as the basis for the construction of the doubly robust estimator, presented in Section 2.4.

2.3.1 Multiple Imputation Generalized Estimating Equations

An imputation model commonly used to handle intermittently missing response and covariate, is the imputation using chained equations (van Buuren *et al.* (1999), van Buuren (2007)), which is more commonly referred to as full conditional specification (FCS). This approach specifies conditional distributions for each incomplete variable, conditionally on all others variables in the imputation model. Starting from an initial imputation, FCS draws imputations by iterating over the conditional densities.

Denote by $\tilde{\boldsymbol{\beta}}_m$ and $\tilde{\boldsymbol{U}}_m$, respectively, the estimate of $\boldsymbol{\beta}$ and its covariance matrix from the GEE analysis of the m-th completed data set, $(m = 1, \ldots, M)$. Following Rubin (1987), the combined point estimate for the parameter of interest $\boldsymbol{\beta}$ from the multiple imputation is simply the average of the M complete-data point estimates

$$\hat{\boldsymbol{\beta}}_{MI} = \frac{1}{M} \sum_{m=1}^{M} \tilde{\boldsymbol{\beta}}_m,$$

and an estimate of the covariance matrix of $\hat{\boldsymbol{\beta}}_{MI}$ is given by

$$\widehat{\boldsymbol{U}}_{MI} = \tilde{\boldsymbol{W}} + \left(\frac{M+1}{M}\right) \tilde{\boldsymbol{B}},$$

where

$$\tilde{\boldsymbol{W}} = \frac{1}{M}\sum_{m=1}^{M}\tilde{\boldsymbol{U}}_m \quad \text{and} \quad \tilde{\boldsymbol{B}} = \frac{1}{M-1}\sum_{m=1}^{M}(\tilde{\boldsymbol{\beta}}_m - \hat{\boldsymbol{\beta}}_{\boldsymbol{MI}})(\tilde{\boldsymbol{\beta}}_m - \hat{\boldsymbol{\beta}}_{\boldsymbol{MI}})^T.$$

2.3.2 Weighted Generalized Estimating Equations

Robins *et al.* (1995) proposed a class of weighted estimating equations to account for MAR mechanism. In a binary longitudinal data setup, Chen & Zhou (2011) extended the method to accommodate arbitrary patterns of missing response and missing covariate. Their method was adapted here for longitudinal ordinal responses.

Define a weight matrix $\boldsymbol{\Delta}_i = [\delta_{itt'}]_{T_i(J_i-1)\times T_i(J_i-1)}$, $t = 1, \ldots, T_i, t' = 1, \ldots, T_i$, where $\delta_{itt'} = \{I(R_{it} = 1, R_{it'} = 3) + I(R_{it} = 3, R_{it'} = 3)\}/\pi_{itt'}$ for $t \neq t'$, $\delta_{itt} = I(R_{it} = 3)/\pi_{it}$, and $\pi_{itt'} = Pr(R_{it} = 1, R_{it'} = 3|\boldsymbol{O}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) + Pr(R_{it} = 3, R_{it'} = 3|\boldsymbol{O}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$. Let $\boldsymbol{M}_i = \boldsymbol{F}_i^{-1/2}(\boldsymbol{C}_i^{-1} \cdot \boldsymbol{\Delta}_i)\boldsymbol{F}_i^{-1/2}$ where $\boldsymbol{A} \cdot \boldsymbol{B} = [a_{it} \cdot b_{it}]$ denotes the Hadamard product of the matrices $\boldsymbol{A} = [a_{it}]$ and $\boldsymbol{B} = [b_{it}]$.

The weighted generalized estimating equations (WGEE) for $\boldsymbol{\beta}$ are given by

$$\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\psi}) = \sum_{i=1}^{n}\boldsymbol{U}_i(\boldsymbol{\beta}, \boldsymbol{\psi}) = \boldsymbol{0}, \tag{2.7}$$

where $\boldsymbol{U}_i(\boldsymbol{\beta}, \boldsymbol{\psi}) = \boldsymbol{D}_i\boldsymbol{M}_i(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$. A consistent estimate for $\boldsymbol{\beta}$ can be obtained by solving (2.7), under the correct specification of the missing data model.

To model λ_{itk} a politomic logistic regression is adopted, with λ_{it0} as being the reference category, that is

$$\log\left(\frac{\lambda_{itk}}{\lambda_{it0}}\right) = \boldsymbol{u}_{itk}^T\psi_k, \quad k = 1, 2, 3, \tag{2.8}$$

where the covariates $\boldsymbol{u}_{itk}$ may be a subset of $\left\{\bar{\boldsymbol{R}}_{it}, \bar{\boldsymbol{O}}_{it}^o, \bar{\boldsymbol{X}}_{it}^o, \boldsymbol{Z}_i\right\}$.

2.4 Doubly Robust GEE for Longitudinal Ordinal Data

Some authors (e.g., Scharfstein *et al.* (1999), Tsiatis (2006)) noted that adding a term of expectation zero, say $\phi(\cdot)$, to the inverse probability weighted estimators would still

result in consistent estimates under a MAR mechanism. The solutions of these augmented estimating equations give rise to the so-called *doubly robust* estimators.

Chen & Zhou (2011) showed that the optimal ϕ_{opt} for missing response and covariate is $\phi_{opt} = E_{(\boldsymbol{Y}_i^m, \boldsymbol{X}_i^m | \boldsymbol{Y}_i^o, \boldsymbol{X}_i^o, \boldsymbol{Z}_i, \boldsymbol{R}_i)} \{\boldsymbol{D}_i \boldsymbol{N}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)\}$, with $\boldsymbol{N}_i = \boldsymbol{F}_i^{-1/2} \left\{\boldsymbol{C}_i^{-1} \cdot (\mathbf{1}\mathbf{1}^T - \boldsymbol{\Delta}_i)\right\} \boldsymbol{F}_i^{-1/2}$, where $\mathbf{1}$ is a vector of 1's of length $T_i(J-1)$, and $\boldsymbol{Y}_i^m$ and $\boldsymbol{X}_i^m$ denote the missing components of $\boldsymbol{Y}_i$ and $\boldsymbol{X}_i$, respectively.

An improved estimate for $\boldsymbol{\beta}$ can then be obtained by solving the estimating equations

$$\boldsymbol{S}_1(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{S}_{1i}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left[\boldsymbol{D}_i \boldsymbol{M}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) + E_{(\boldsymbol{Y}_i^m, \boldsymbol{X}_i^m | \boldsymbol{Y}_i^o, \boldsymbol{X}_i^o, \boldsymbol{Z}_i, \boldsymbol{R}_i)} \{\boldsymbol{D}_i \boldsymbol{N}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)\}\right] = \mathbf{0}. \tag{2.9}$$

The estimator for $\boldsymbol{\beta}$ in (2.9) is doubly-robust in the sense that it is consistent if *at least one* of the missing data model or the covariate model is correctly specified.

Applications of doubly robust estimators in longitudinal settings include Bang & Robins (2005), Seaman & Copas (2009), Chen & Zhou (2011) and Birhanu *et al.* (2011). Those developments focus mainly on binary response and have not been, to our knowledge, investigated with ordinal longitudinal data. In this work, it is considered a longitudinal response measured on a ordinal scale.

The referred expectation in the second part of (2.9) is with respect to the conditional distribution of $(\boldsymbol{Y}_i^m, \boldsymbol{X}_i^m | \boldsymbol{Y}_i^o, \boldsymbol{X}_i^o, \boldsymbol{Z}_i, \boldsymbol{R}_i)$, which can be written as

$$\begin{aligned} P(\boldsymbol{Y}_i^m = \boldsymbol{y}_i^m, \boldsymbol{X}_i^m = \boldsymbol{x}_i^m | \boldsymbol{Y}_i^o, \boldsymbol{X}_i^o, \boldsymbol{Z}_i, \boldsymbol{R}_i; \boldsymbol{\beta}^*, \boldsymbol{\gamma}) &= P(\boldsymbol{Y}_i^m = \boldsymbol{y}_i^m, \boldsymbol{X}_i^m = \boldsymbol{x}_i^m | \boldsymbol{Y}_i^o, \boldsymbol{X}_i^o, \boldsymbol{Z}_i; \boldsymbol{\beta}^*, \boldsymbol{\gamma}) \\ &= P(\boldsymbol{Y}_i^m = \boldsymbol{y}_i^m | \boldsymbol{Y}_i^o, \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}^*) \\ &\quad \times P(\boldsymbol{X}_i^m = \boldsymbol{x}_i^m | \boldsymbol{Y}_i^o, \boldsymbol{X}_i^o, \boldsymbol{Z}_i; \boldsymbol{\gamma}). \end{aligned}$$

The multivariate distribution $P(\boldsymbol{Y}_i^m = \boldsymbol{y}_i^m | \boldsymbol{Y}_i^o, \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}^*)$ is expressed through a product of univariate ordinal models.

2.4.1 Estimation for the Nuisance Parameters

The method of maximum likelihood is employed to estimate $\boldsymbol{\psi}$. The log likelihood of the politomic logistic model for $\boldsymbol{\psi}$ has the form

$$l(\boldsymbol{\psi}) = \sum_{i=1}^{n} l_i(\boldsymbol{\psi}) = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \sum_{k=0}^{3} I(R_{it} = k) \log(\lambda_{itk}),$$

with corresponding score function given by

$$\boldsymbol{S}_2(\boldsymbol{\psi}) = \sum_{i=1}^{n} \boldsymbol{S}_{2i}(\boldsymbol{\psi}) = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \sum_{k=0}^{3} \frac{I(R_{it} = k)}{\lambda_{itk}} \frac{\partial \lambda_{itk}}{\partial \boldsymbol{\psi}^T}.$$

The maximum likelihood estimator $\hat{\boldsymbol{\psi}}$, is obtained by solving $\boldsymbol{S}_2(\boldsymbol{\psi}) = \mathbf{0}$.

For the missing covariate model, the observed likelihood function for $\boldsymbol{\gamma}$ is

$$L_3(\boldsymbol{\gamma}) = \prod_{i=1}^{n} \int Pr(\boldsymbol{X}_i | \boldsymbol{Z}_i, \boldsymbol{Y}_i) d\boldsymbol{X}_i^m,$$

with score function $\boldsymbol{S}_3(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \boldsymbol{S}_{3i}(\boldsymbol{\gamma})$, where $\boldsymbol{S}_{3i}(\boldsymbol{\gamma}) = \partial \log \int Pr(\boldsymbol{X}_i | \boldsymbol{Z}_i, \boldsymbol{Y}_i)) d\boldsymbol{X}_i^m / \partial \boldsymbol{\gamma}^T$. Similarly, a consistent estimator of $\boldsymbol{\gamma}$ can be obtained by solving $\boldsymbol{S}_3(\boldsymbol{\gamma}) = \mathbf{0}$.

2.4.2 Estimation and Inference for the Doubly Robust Method

Let's denote the vector of all parameters as $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\psi}^T, \boldsymbol{\gamma}^T)^T$. Our primary interest lies in estimating $\boldsymbol{\beta}$. This task can be accomplished by plugging in the estimates $\hat{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\gamma}}$ in (2.9) and solving the estimating equations for $\boldsymbol{\beta}$, that is,

$$\boldsymbol{S}_1(\boldsymbol{\beta}, \hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} \boldsymbol{S}_{1i}(\boldsymbol{\beta}, \hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\gamma}}) = 0. \tag{2.10}$$

Second term of $\boldsymbol{S}_{1i}$ in (2.9) can be written, for $\boldsymbol{X}$ discrete, as

$$E_{(\boldsymbol{Y}_i^m, \boldsymbol{X}_i^m | \boldsymbol{Y}_i^o, \boldsymbol{X}_i^o, \boldsymbol{Z}_i, \boldsymbol{R}_i)} \left\{ \boldsymbol{D}_i \boldsymbol{N}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \right\} = \sum_{(\boldsymbol{y}_i^m, \boldsymbol{x}_i^m)} w_{ixy} \left\{ \boldsymbol{D}_i \boldsymbol{N}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \right\},$$

where the weight w_{ixy} is given by

$$w_{ixy} = P(\boldsymbol{Y}_i^m = \boldsymbol{y}_i^m | \boldsymbol{Y}_i^o, \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}^*) \times P(\boldsymbol{X}_i^m = \boldsymbol{x}_i^m | \boldsymbol{Y}_i^o, \boldsymbol{X}_i^o, \boldsymbol{Z}_i; \hat{\boldsymbol{\gamma}}).$$

In the case of $\boldsymbol{X}$ continuous, the second term in $\boldsymbol{S}_{1i}$ takes the form

$$E_{(\boldsymbol{Y}_i^m, \boldsymbol{X}_i^m | \boldsymbol{Y}_i^o, \boldsymbol{X}_i^o, \boldsymbol{Z}_i, \boldsymbol{R}_i)} \left\{ \boldsymbol{D}_i \boldsymbol{N}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \right\} = \int_{(\boldsymbol{Y}_i^m, \boldsymbol{X}_i^m)} w_{ixy} \left\{ \boldsymbol{D}_i \boldsymbol{N}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \right\} d\boldsymbol{Y}_i^m \boldsymbol{X}_i^m,$$

with conditional probability

$$w_{ixy} = P(\boldsymbol{Y}_i^m = \boldsymbol{y}_i^m | \boldsymbol{Y}_i^o, \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}^*) \times P(\boldsymbol{X}_i^m = \boldsymbol{x}_i^m | \boldsymbol{Y}_i^o, \boldsymbol{X}_i^o, \boldsymbol{Z}_i; \hat{\boldsymbol{\gamma}}).$$

The expectation in (2.9) can be cumbersome, depending on the missing data pattern. In such case, instead of using numerical integration, a Monte Carlo method can be applied to approximate the corresponding integral.

In this work, it is assumed independence working correlation and it is adopted a sandwich standard error as given in the Appendix A.

2.5 Simulation Study

A small simulation study taking into account different sample sizes was conducted in order to quantify the bias and precision under misspecification of the predictive models. It is considered a study with $T_i = T = 3$ repeated ordinal measures (with three categories) and two covariates (quantitative and qualitative). The true marginal model is

$$logit\ Pr(O_{itj} \leq j | X_{it}, Z_{it}) = \beta_{0j} + \beta_1 X_{it} + \beta_2 Z_{it}, \quad j = 1, 2. \tag{2.11}$$

where Z_{it} is normal with unit variance and mean 0, 0.5 or 1 for $t = 1, 2, 3$, respectively.

The binary covariate X_{it} may be missing at some time points and is generated according to

$$logit\ Pr(X_{it} = 1 | \bar{X}_{it}, Z_{it}) = \gamma_0 + \gamma_1 X_{i,t-1} + \gamma_2 Z_{it}. \tag{2.12}$$

It is assumed $\beta_{01} = -0.4$, $\beta_{02} = 1.2$, $\beta_1 = -0.5$, $\beta_2 = 0.5$, $\gamma_0 = log(1)$, $\gamma_1 = 2$ and $\gamma_2 = 2$. The correlated ordinal responses were generated according to the algorithm proposed by Touloumis (2015) with constant correlation between the latent vectors equals to $\rho = 0.9$.

As independent estimating equations were fitted, R_{it} can be defined as the indicator of observing both O_{it} and X_{it}, and

$$log\left(\frac{Pr(R_{it}=1)}{Pr(R_{it}=0)}\right) = \psi_{0t} + \psi_1 I(R_{i,t-1}=1) + \psi_2 O^*_{i,t-1} + \psi_3 X^*_{i,t-1} + \psi_4 Z_{it}, \;\; t = 2, 3, \quad (2.13)$$

where $O^*_{i,t-1} = O_{i,t-1}$, if $O_{i,t-1}$ is observed and 0 otherwise, and $X^*_{i,t-1} = X_{i,t-1}$if $X_{i,t-1}$ is observed and 0 otherwise. The true values are $\psi_{02} = 6.6$, $\psi_{03} = 6$, $\psi_1 = 2$, $\psi_2 = -2$, $\psi_3 = -2$ and $\psi_4 = 2$. About 24% of the observations were missing under this setup.

For comparison purposes, it was considered ordinary GEE for the complete and available data, respectively, weighted GEE (WGEE), multiple imputation (MIGEE) by chained equations with $M = 10$, and the proposed doubly robust version (DRGEE). In order to investigate robustness of these methods, the predicted models were also misspecified by omitting the covariate X_{t-1} from the covariate model (2.12) or the missing data model (2.13). Correctly specified models are indicated with a *plus* sign ('x^+' and 'r^+', for the covariate and missing data model, respectively) and incorrectly specified models are indicated with a *minus* sign ('x^-' and 'r^-', for the covariate and missing data model, respectively). All methods were implemented in the software *R* (R Core Team, 2015). For the generation of the multiple imputed values, the package *mice* (van Buuren & Groothuis-Oudshoorn, 2011) was used.

Results are summarized in Table 2.1. In each of the $S = 1000$ Monte Carlo replications it was obtained the relative bias percentage for each parameter, defined as $100 \times (\hat{\beta} - \beta)/\beta$, its standard deviation obtained through the sandwich estimator, and the coverage probability as a nominal level of 95%.

Specifically, the MAR missingness impact over the response and the covariate is observed for all regression parameters, the largest relative bias occurs in the binary covariate X. This comes in addition to the natural increase of parameter uncertainty. Bias in the

Table 2.1: Relative bias percentage, standard deviation and empirical coverage for 1000 simulations of incomplete covariate and response data.

	Relative Bias				Standard Deviation				Empirical Coverage			
	β_{01}	β_{02}	β_1	β_2	β_{01}	β_{02}	β_1	β_2	β_{01}	β_{02}	β_1	β_2
						$n = 50$						
Complete	-4.60	7.89	15.07	7.38	0.379	0.410	0.458	0.179	0.94	0.94	0.97	0.95
Available	-19.81	15.37	-8.36	-11.66	0.386	0.420	0.473	0.190	0.94	0.92	0.97	0.94
WGEE(r^+)	-12.33	12.02	19.96	1.74	0.406	0.440	0.513	0.204	0.94	0.94	0.97	0.95
WGEE(r^-)	1.50	7.43	1.31	0.51	0.413	0.450	0.508	0.201	0.95	0.93	0.96	0.95
MIGEE(x^+)	22.17	-3.44	-38.26	-4.73	0.386	0.417	0.480	0.186	0.96	0.96	0.97	0.96
MIGEE(x^-)	15.81	-0.13	-26.18	-2.54	0.387	0.417	0.481	0.187	0.95	0.96	0.97	0.96
DRGEE(x^+, r^+)	-8.86	9.67	22.20	8.99	0.410	0.442	0.513	0.199	0.95	0.94	0.97	0.95
DRGEE(x^-, r^+)	-7.77	9.44	19.33	7.91	0.471	0.507	0.579	0.204	0.95	0.94	0.97	0.95
DRGEE(x^+, r^-)	-7.95	10.05	20.58	8.26	0.433	0.473	0.540	0.197	0.95	0.93	0.96	0.95
DRGEE(x^-, r^-)	2.86	6.42	-0.28	2.99	0.414	0.450	0.529	0.200	0.95	0.93	0.97	0.95
						$n = 150$						
Complete	-1.36	1.53	5.13	1.99	0.220	0.236	0.264	0.103	0.94	0.94	0.96	0.95
Available	-17.50	8.94	-18.41	-17.15	0.224	0.242	0.273	0.109	0.91	0.92	0.94	0.91
WGEE(r^+)	-5.84	3.41	13.47	1.36	0.247	0.267	0.317	0.126	0.93	0.93	0.95	0.94
WGEE(r^-)	10.20	-2.04	-12.91	-2.59	0.253	0.279	0.307	0.120	0.94	0.93	0.93	0.94
MIGEE(x^+)	8.22	-2.57	-14.36	-2.40	0.228	0.244	0.283	0.108	0.94	0.94	0.96	0.95
MIGEE(x^-)	9.08	-1.97	-16.33	-3.60	0.226	0.242	0.280	0.108	0.93	0.93	0.95	0.94
DRGEE(x^+, r^+)	-4.21	2.64	11.90	4.12	0.249	0.269	0.317	0.117	0.94	0.94	0.96	0.95
DRGEE(x^-, r^+)	-5.02	3.07	12.26	3.63	0.320	0.345	0.396	0.119	0.94	0.94	0.95	0.95
DRGEE(x^+, r^-)	-1.86	1.99	7.26	2.95	0.249	0.275	0.313	0.113	0.94	0.94	0.95	0.94
DRGEE(x^-, r^-)	9.88	-1.86	-15.60	-2.86	0.246	0.269	0.317	0.116	0.94	0.93	0.94	0.94
						$n = 300$						
Complete	1.42	0.19	-0.04	0.85	0.156	0.167	0.187	0.072	0.96	0.94	0.94	0.95
Available	-14.82	7.67	-23.24	-18.14	0.159	0.171	0.194	0.077	0.93	0.91	0.92	0.87
WGEE(r^+)	-0.24	0.75	5.61	1.23	0.182	0.198	0.235	0.095	0.95	0.94	0.94	0.93
WGEE(r^-)	15.64	-4.50	-22.17	-3.89	0.184	0.205	0.222	0.088	0.95	0.94	0.92	0.94
MIGEE(x^+)	6.23	-1.86	-9.95	-1.44	0.162	0.174	0.201	0.076	0.95	0.94	0.94	0.95
MIGEE(x^-)	10.27	-2.70	-18.22	-3.89	0.160	0.172	0.200	0.076	0.94	0.92	0.92	0.94
DRGEE(x^+, r^+)	-0.06	0.83	2.82	1.28	0.185	0.202	0.242	0.089	0.96	0.95	0.95	0.94
DRGEE(x^-, r^+)	-0.71	1.08	3.84	1.43	0.194	0.213	0.251	0.088	0.96	0.95	0.95	0.94
DRGEE(x^+, r^-)	2.53	0.03	-2.11	0.12	0.178	0.198	0.224	0.082	0.96	0.96	0.94	0.94
DRGEE(x^-, r^-)	13.86	-3.69	-24.29	-5.51	0.176	0.194	0.228	0.084	0.95	0.94	0.93	0.94
						$n = 600$						
Complete	-1.01	0.57	1.93	0.82	0.110	0.118	0.132	0.051	0.95	0.93	0.94	0.94
Available	-16.78	7.85	-22.20	-18.43	0.112	0.121	0.137	0.054	0.91	0.88	0.92	0.78
WGEE(r^+)	-1.00	0.28	5.43	1.25	0.132	0.144	0.170	0.070	0.95	0.94	0.95	0.95
WGEE(r^-)	14.87	-4.99	-21.53	-3.03	0.133	0.149	0.158	0.064	0.94	0.94	0.92	0.93
MIGEE(x^+)	1.76	-0.66	-3.63	-0.39	0.115	0.123	0.143	0.054	0.95	0.94	0.95	0.94
MIGEE(x^-)	6.96	-2.01	-14.18	-3.38	0.113	0.122	0.141	0.054	0.94	0.93	0.93	0.93
DRGEE(x^+, r^+)	-1.49	0.59	3.19	0.98	0.132	0.144	0.174	0.062	0.96	0.94	0.94	0.94
DRGEE(x^-, r^+)	-1.39	0.56	2.96	0.91	0.136	0.149	0.178	0.062	0.96	0.94	0.96	0.95
DRGEE(x^+, r^-)	0.42	0.08	-0.12	0.47	0.126	0.140	0.159	0.058	0.96	0.95	0.95	0.93
DRGEE(x^-, r^-)	12.24	-3.82	-23.18	-5.35	0.124	0.138	0.161	0.060	0.93	0.93	0.92	0.93

"+" indicates correctly specified model and "−" indicates misspecified model omitting the X_t predictor

intercept coefficients imply incorrect predicted probabilities for the levels of the response, whereas bias for parameter estimates associated with the regression covariates may erroneously attenuate or highlight an effect, thus leading to misinterpretations related to the importance of a given predictor on the longitudinal dynamics of the ordinal response.

It can be observed that for small sample sizes even the GEE for complete data presents a certain degree of bias. Increasing the sample size allows to clarify the performance distinctions among the compared methods. WGEE and MIGEE estimators are consistent when the model for the weight or the imputation model, respectively, are correctly specified. In this case, it is noted that both methods give good results for large sample sizes, the main distinction between them is the greater variability of the estimates for the weighted estimator.

The doubly robust method requires the simultaneous specification of two predictive models. When at least one of them is correctly specified the resulting estimator is still consistent. Estimates are, on average, closer to those obtained with fully observed data compared to WGEE or MIGEE. This behavior is systematic and it can be observed for all parameters. By increasing the sample size, the estimates from DRGEE present empirical bias in general smaller than their single robust competitors. This is specially true for the parameter associated with the incomplete binary variable X. Regarding the uncertainty of the parameter estimates, it is noted that the variability in DRGEE is greater than in the multiple imputation, but of the same order as the weighted method. Further, the efficiency of the doubly robust estimates appears relatively more sensitive to misspecification of the weight model than the covariate model. Empirical coverage rates were acceptable for correctly specified WGEE and MIGEE as well as for DRGEE when at least one of the predictive models are correctly specified.

Figure 2.1 shows boxplots of the percentage relative bias for the methods expected to be valid. As the degree of bias is different for the parameter estimates, for ease of visualization, they are represented in different scales. As can be seen, the estimates with larger variability were those associated with the first intercept and the covariate with incomplete values. The proposed method presents median relative bias close to zero, very

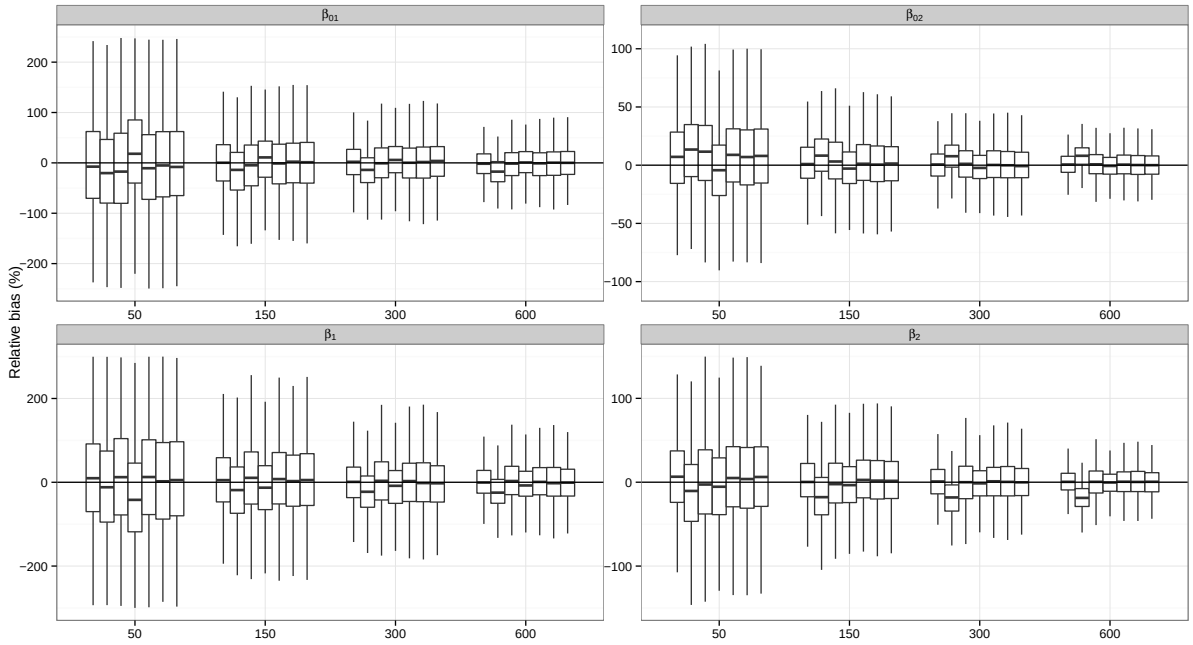

Figure 2.1: Boxplot of the relative bias for parameter estimates under correctly specified models. The boxes represent, respectively, GEE for complete and available data, WGEE(r^+), MIGEE(x^+), DRGEE(x^+, r^+), DRGEE(x^-, r^+) and DRGEE(x^+, r^-).

similar to those observed with complete data for all sample sizes. Variability of the doubly robust estimators is slightly larger than multiple imputation but of the same order as the weighted estimator. As expected, for all methods it can be noticed that the relative bias decreases as the sample size increases, reflecting their theoretical asymptotic consistency.

Figure 2.2 allows the comparison of incorrectly specified methods. When a key covariate is omitted from the weight model and/or from the imputation one, all methods are expected to be biased. This is specially true for the first intercept and for the incomplete binary covariate. Bias for MIGEE(x^-) seemed to get smaller than WGEE(r^-) as sample sizes increases. The bias of DRGEE(x^-, r^-) was comparable to WGEE(r^-) and slightly higher than that of multiple imputation for large sample sizes. In terms of variability of the estimates, the same pattern as those for correctly specified models is observed. That is, the multiple imputation is more efficient, followed by the doubly robust estimator and the weighted estimator.

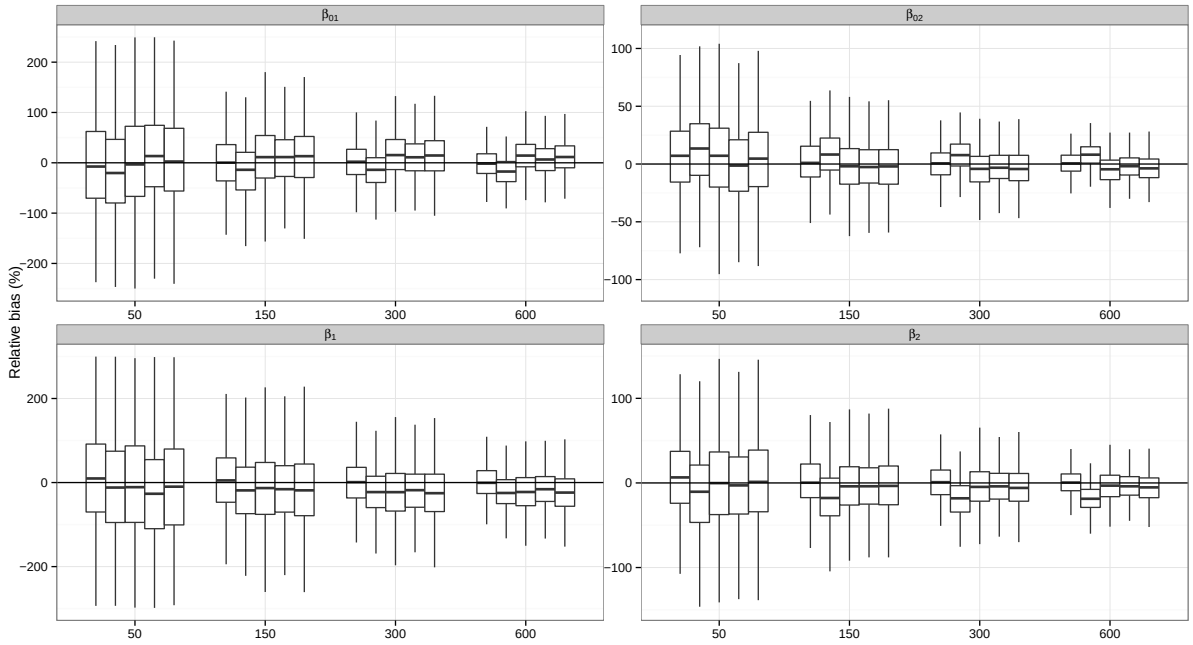

Figure 2.2: Boxplot of the relative bias for the estimates with incorrectly specified models. The boxes represent, respectively, GEE for complete and available data, WGEE(r^-), MIGEE(x^-), and DRGEE(x^-, r^-).

2.6 Data Analysis: Analgesia in Childbirth

This study was conducted in Minas Gerais state, Brazil, in order to compare two techniques of analgesia for labor pain. There were 49 patients monitored during their entire labor period until childbirth. Pain intensity was subjectively assessed by the patient, and measurements of blood pressure, maternal heart rate, infusion of oxytocin, sedation level, signs of respiratory depression, apnea, and other variables were recorded. One of the techniques used was epidural analgesia (the gold standard), which is a local anesthetic. The other one, whose efficiency was to be compared to the gold standard, involved continuous intravenous infusion of remifentanil, an opioid that has very rapid onset of action (1-3 minutes).

The response of interest is the intensity of pain as measured by a Visual Analog Scale (VAS), that consists of a straight line with one end meaning "no pain" and the other end meaning the "worst pain imaginable". The response, PAIN, was then coded as: 1: tolerable and mild pain; 2: moderate pain that causes discomfort; 3: intense and unbearable pain. Three measurements $(0, 60, 90$ minutes) were selected for data

analysis. Four predictor variables were considered: treatment GROUP (0: peridural; 1: remifentanil), AGE (in years), DU (uterine dilatation, in cm), and OXYT (infusion of oxytocin, a hormone that stimulates contractions during labor and birth, in mL/h). The OXYT is a time-varying ordinal covariate, coded as 1, if no infusion, 2, if infusion equals to 10 mL/h or 30 mL/h, and 3, if infusion equals or above 45 mL/h. These other covariates were chosen after a previous exploratory analysis.

The average age of patients was 22 years. Before the patient receive the anesthesia (at time 0), 49.0% of them rated their pain as intense, this percentage remained approximately constant in the other time (50% and 48.4% at times 60 and 90 minutes, respectively). Most patients (67.3%) opted for standard anesthesia. Figure 2.3 shows the observed longitudinal profile of pain intensity by group. The observed proportion of patients with mild pain was higher in the remifentanil group.

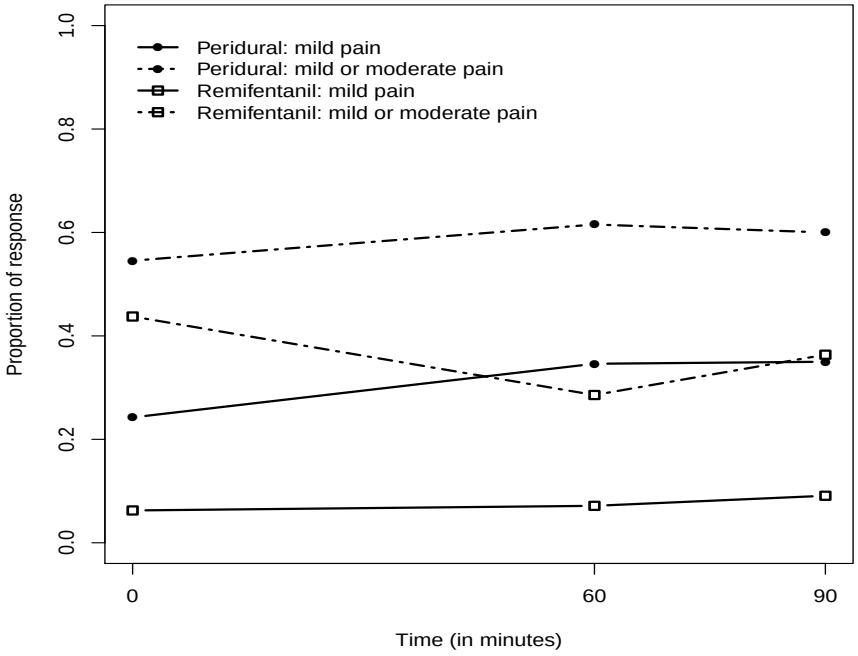

Figure 2.3: Observed longitudinal profile of pain intensity.

The response and oxytocin infusion was missing for 9 patients at time 60 and for 18 patients at time 90. Missingness is due to the fact that childbirth happened before 60 or 90 minutes. Therefore a MAR mechanism seems to be a reasonable assumption for this

data set. The other covariates in the analysis were fully observed.

For the ordinal response the following proportional odds model was adopted

$$logit\ Pr(PAIN_{itj} \leq j|\boldsymbol{u_{it}}) = \beta_{0j} + \boldsymbol{u}_{it}^T\boldsymbol{\beta}, \quad j = 1, 2, \quad t = 1, 2, 3, \tag{2.14}$$

where $\boldsymbol{u}_{it}$ is the covariate vector at time t, and it is formed by TIME, GROUP, AGE, DU and OXIT. In the response model, the TIME predictor was expressed in hours rather than minutes.

When using WGEE or DRGGE, it is necessary to correctly model π_i in order to obtain consistent estimates for $\boldsymbol{\beta}$. For the missing data process, R_{it} was defined as the indicator of observing both $PAIN_{it}$ and $OXIT_{it}$, and it takes the following form

$$log\left(\frac{Pr(R_{it} = 1)}{Pr(R_{it} = 0)}\right) = \psi_{0t} + \boldsymbol{w}_{it}^T\boldsymbol{\psi}, \quad t = 2, 3, \tag{2.15}$$

where $\boldsymbol{w}_{it}$ includes GROUP, AGE, DU, histories of OXYT and PAIN, and the previous indicator of missing data.

The distribution of the missing covariate OXYT also needs to be specified in a predictive model. With this aim, it was assumed a proportional odds model of the form

$$logit\ Pr(OXYT_{itj} \leq j|\boldsymbol{v_{it}}) = \gamma_{0j} + \boldsymbol{v}_{it}^T\boldsymbol{\gamma}, \quad j = 1, 2, \quad t = 2, 3 \tag{2.16}$$

where $\boldsymbol{v}_{it}$ includes main effects for GROUP, AGE, and DU. Note that the estimate of $\boldsymbol{\gamma}$ is not of interest, however it is necessary to model the missing mechanism related to the covariate as close as possible to truth in order to obtain consistent estimates for $\boldsymbol{\beta}$. The same is true for the missing data process. All predictors in this model process were maintained since an overspecification is better than a underspecification.

Results from four methods are shown in Table 2.2. The first one is the usual GEE method using the available data; the second is the weighted method (WGEE) using model (2.15) for the weights; the third is the multiple imputation by chained equation (MIGEE) in the *R* package *mice*; and the fourth, labeled DRGEE, is the proposed doubly robust

Table 2.2: Regression Parameters for the Analgesia in Birth Data

	Available			WGEE			MIGEE			DRGEE		
Parameter	Est	SE	P	Est	SE	P	Est	SE	P	Est	SE	P
INTERCEPT1	1.796	1.308	0.170	1.827	1.267	0.149	1.599	1.237	0.196	1.649	1.315	0.210
INTERCEPT2	3.302	1.314	0.012	3.261	1.310	0.013	3.105	1.242	0.012	3.069	1.351	0.023
TIME	-0.182	0.286	0.524	-0.157	0.315	0.619	-0.110	0.273	0.687	-0.192	0.293	0.511
GROUP	-1.221	0.445	0.006	-1.244	0.439	0.005	-1.056	0.409	0.010	-1.008	0.390	0.010
AGE	-0.066	0.035	0.057	-0.072	0.033	0.027	-0.062	0.031	0.046	-0.071	0.035	0.045
DU	-0.362	0.158	0.022	-0.354	0.168	0.035	-0.372	0.156	0.017	-0.340	0.166	0.040
OXYT(=2)	0.727	0.554	0.189	0.712	0.554	0.199	0.954	0.542	0.078	0.821	0.531	0.122
OXYT(=3)	1.285	0.510	0.012	1.431	0.503	0.004	1.410	0.488	0.004	1.468	0.475	0.002

method using (2.15) and (2.16) for the weight and the covariate models, respectively. It was used an independent working correlation.

The TIME effect is not significant for all the four methods. All methods provide the same conclusion for effects of GROUP. The negative effect of GROUP means that the chance of a woman feel mild pain is lower within the group receiving the remifentanil compared to the epidural group (the estimated odds is $e^{-1.008} = 0.365$ (95% CI: $0.247 - 0.539$) in the doubly robust method). All methods also agree with respect to the effect of DU. That is, for each increase of 1 cm of uterine dilation the chance of the parturient feel mild pain decreases (in the DRGEE it is $e^{-0.340} = 0.712$ (95% CI: $0.603 - 0.840$), for example). It can be noticed that the p-value for AGE effect varies from a non-significant value of 0.057 in the standard GEE to a significant one in DRGEE, as well as for the other two missing data approaches. The conclusion is that older women have lower chance of experiencing mild pain than young women. For the OXIT covariate all methods reached the same conclusion, that is, women who received high doses of oxytocin presented greater chance of experiencing mild pain than those women who received low doses of oxytocin.

2.7 Discussion

When longitudinal ordinal data are of interest, doubly robust GEE is a nice alternative to MIGEE and WGEE. The doubly robust method combines ideas from weighting and imputation and has been applied elsewhere for estimation of means, causal inference, and in the longitudinal setting for binary response data (Bang & Robins (2005), Carpenter *et al.* (2006), Seaman & Copas (2009), Chen & Zhou (2011), Li *et al.* (2013)). However,

as far as we know, it has not been investigated for the longitudinal ordinal case. A doubly robust estimator is attractive in the sense that it needs only the correct specification of at least one of the models, but not necessarily both. Simulation results have indicated that, when at least the covariate model or missing data model is correct, the doubly robust estimators are consistent and has small-sample bias comparable to single robust alternatives MIGEE or WGEE. The proposed method presented good coverage probabilities, as well as its competitors but with a slightly larger variance than multiple imputation. Simulation results also indicates that the bias of doubly robust estimators, when both the covariate model and the missing data model are incorrect, was of the same magnitude as those from the misspecified WGEE or MIGEE. We expect that, in practical applications, none of the predictive models would be grossly misspecified and then the proposed estimator would have a great potential of reducing the bias if the MAR assumption is correct.

When the assumed independent working correlation structure differs from the true underlying structure, there is no price to pay in terms the consistency and asymptotic normality of $\boldsymbol{\beta}$, but such a poor choice may result in loss of efficiency (Molenberghs & Verbeke, 2005). However, modeling the association structure in the presence of missing data remains a challenge, specially with longitudinal ordinal data, because there is no direct way of modeling the association parameters. Future research involves the investigation of the impact of other association structures in the doubly robust estimates.

In the proposed doubly robust estimator, marginal means were modeled by cumulative logits. This implies a proportional odds model that in some cases may not be valid. Another possible extension of the proposed model is, therefore, to allow for non-proportional odds for a subset of the explanatory variables (Peterson & Harrell Jr, 1990).

BIBLIOGRAPHY

Agresti, Alan. 2013. *Categorical Data Analysis*. 3 edn. John Wiley & Sons.

Bang, Heejung, & Robins, James M. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61**(4), 962–973.

Becker, Mark P, & Clogg, Clifford C. 1989. Analysis of sets of two-way contingency tables using association models. *Journal of the American Statistical Association*, **84**(405), 142–151.

Beunckens, Caroline, Sotto, Cristina, & Molenberghs, Geert. 2008. A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational Statistics & Data Analysis*, **52**(3), 1533–1548.

Birhanu, Teshome, Molenberghs, Geert, Sotto, Cristina, & Kenward, Michael G. 2011. Doubly robust and multiple-imputation-based generalized estimating equations. *Journal of Biopharmaceutical Statistics*, **21**(2), 202–225.

Carey, Vincent, Zeger, Scott L, & Diggle, Peter. 1993. Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, **80**(3), 517–526.

Carpenter, James R., & Kenward, Michael G. 2013. *Multiple Imputation and its Application*. Wiley & Sons.

Carpenter, James R, Kenward, Michael G, & Vansteelandt, Stijn. 2006. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **169**(3), 571–584.

Chen, Baojiang, & Zhou, Xiao-Hua. 2011. Doubly robust estimates for binary longitudinal data analysis with missing response and missing covariates. *Biometrics*, **67**(3), 830–842.

Deming, W Edwards, & Stephan, Frederick F. 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, **11**(4), 427–444.

Donneau, Anne-Françoise, Mauer, Murielle, Molenberghs, Geert, & Albert, Adelin. 2015a. A simulation study comparing multiple imputation methods for incomplete longitudinal ordinal data. *Communications in Statistics-Simulation and Computation*, **44**(5), 1311–1338.

Donneau, Anne-Françoise, Mauer, Murielle, Lambert, Philippe, Molenberghs, Geert, & Albert, Adelin. 2015b. Simulation-based study comparing multiple imputation methods for non-monotone missing ordinal data in longitudinal settings. *Journal of Biopharmaceutical Statistics*, **25**(3), 570–601.

Fitzmaurice, G., Davidian, M., Molenberghs, G., & Verbeke, G. 2009. *Longitudinal Data Analysis*. Handbooks of Modern Statistical Methods. Chapman & Hall/CRC.

Fitzmaurice, Garrett M., Laird, M., & Ware, James H. 2004. *Applied Longitudinal Analysis*. Wiley-Interscience.

Goodman, Leo A. 1985. The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *The Annals of Statistics*, **13**(1), 10–69.

Heagerty, Patrick J, & Zeger, Scott L. 1996. Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association*, **91**(435), 1024–1036.

Li, Lingling, Shen, Changyu, Li, Xiaochun, & Robins, James M. 2013. On weighting approaches for missing data. *Statistical Methods in Medical Research*, **22**(1), 14–30.

Liang, Kung-Yee, & Zeger, Scott L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22.

Lipsitz, Stuart R, Laird, Nan M, & Harrington, David P. 1991. Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, **78**(1), 153–160.

Lipsitz, Stuart R, Kim, Kyungmann, & Zhao, Lueping. 1994. Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine*, **13**(11), 1149–1163.

Little, Roderick JA, & Rubin, Donald B. 1987. *Statistical Analysis with Missing Data*. 1 edn. Wiley New York.

Little, Roderick JA, & Rubin, Donald B. 2002. *Statistical Analysis with Missing Data*. 2 edn. Wiley New York.

Lumley, Thomas. 1996. Generalized estimating equations for ordinal data: a note on working correlation structures. *Biometrics*, **52**(1), 354–361.

McCullagh, Peter. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, **42**(2), 109–142.

Molenberghs, Geert, & Kenward, Michael G. 2010. Semi-parametric marginal models for hierarchical data and their corresponding full models. *Computational Statistics & Data Analysis*, **54**(2), 585–597.

Molenberghs, Geert, & Verbeke, Geert. 2005. *Models for discrete longitudinal data*. 1 edn. Springer Series in Statistics. Springer-Verlag New York.

Nooraee, Nazanin, Molenberghs, Geert, & van den Heuvel, Edwin R. 2014. GEE for longitudinal ordinal data: Comparing R-geepack, R-multgee, R-repolr, SAS-GENMOD, SPSS-GENLIN. *Computational Statistics & Data Analysis*, **77**, 70–83.

Parsons, Nicholas R, Edmondson, RN, & Gilmour, SG. 2006. A generalized estimating equation method for fitting autocorrelated ordinal score data with an application in horticultural research. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **55**(4), 507–524.

Parsons, Nick R, Costa, Matthew L, Achten, Juul, & Stallard, Nigel. 2009. Repeated measures proportional odds logistic regression analysis of ordinal score data in the statistical software package R. *Computational Statistics & Data Analysis*, **53**(3), 632–641.

Peterson, Bercedis, & Harrell Jr, Frank E. 1990. Partial proportional odds models for ordinal response variables. *Applied Statistics*, **39**(2), 205–217.

Pierce, Donald A. 1982. The asymptotic effect of substituting estimators for parameters in certain types of statistics. *The Annals of Statistics*, **10**(2), 475–478.

Poleto, Frederico Z, Singer, Julio M, & Paulino, Carlos Daniel. 2014. A product-multinomial framework for categorical data analysis with missing responses. *Brazilian Journal of Probability and Statistics*, **28**(1), 109–139.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robins, James M, Rotnitzky, Andrea, & Zhao, Lue Ping. 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**(429), 106–121.

Rubin, Donald B. 1976. Inference and missing data. *Biometrika*, **63**(3), 581–592.

Rubin, Donald B. 1978. Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. *Pages 20–34 of: Proceedings of the survey research methods section of the American Statistical Association*, vol. 1. American Statistical Association.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. John Wiley New York.

Schafer, Joseph L. 1997. *Analysis of incomplete multivariate data*. Monographs on Statistics & Applied Probability. Chapman and Hall/CRC.

Schafer, Joseph L. 1999. Multiple imputation: a primer. *Statistical Methods in Medical Research*, **8**(1), 3–15.

Scharfstein, Daniel O, Rotnitzky, Andrea, & Robins, James M. 1999. Adjusting for non-ignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, **94**(448), 1096–1120.

Seaman, Shaun, & Copas, Andrew. 2009. Doubly robust generalized estimating equations for longitudinal data. *Statistics in Medicine*, **28**(6), 937–955.

Toledano, Alicia Y, & Gatsonis, Constantine. 1999. Generalized estimating equations for ordinal categorical data: arbitrary patterns of missing responses and missingness in a key covariate. *Biometrics*, **55**(2), 488–496.

Touloumis, Anestis. 2015. *SimCorMultRes: Simulates Correlated Multinomial Responses*. R package version 1.3.0.

Touloumis, Anestis, Agresti, Alan, & Kateri, Maria. 2013. GEE for multinomial responses using a local odds ratios parameterization. *Biometrics*, **69**(3), 633–640.

Tsiatis, Anastasios A. 2006. *Semiparametric theory and missing data*. Springer Series in Statistics. Springer New York.

van Buuren, S., Boshuizen, H. C., & Knook, D. L. 1999. Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. *Statistics in Medicine*, **18**(6), 681–694.

van Buuren, Stef. 2007. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, **16**(3), 219–242.

van Buuren, Stef, & Groothuis-Oudshoorn, Karin. 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, **45**(3), 1–67.

Vansteelandt, Stijn, Carpenter, James, & Kenward, Michael G. 2010. Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, **6**(1), 37–48.

CHAPTER 3

ARTIGO 2: MODELING THE ASSOCIATION STRUCTURE IN DOUBLY ROBUST GEE FOR LONGITUDINAL ORDINAL MISSING DATA

José Luiz P. da Silva, Enrico A. Colosimo, and Fábio N. Demarqui

Abstract: *Generalized Estimation Equations (GEE) are a well-known method for the analysis of categorical longitudinal responses. The GEE method has computational simplicity and population parameter interpretation. In the presence of missing data, this estimator is only consistent under the strong assumption of missing completely at random. A doubly robust estimator (DRGEE) for correlated ordinal longitudinal data is a nice approach for handling intermittently missing response and covariate under the MAR mechanism. Independent working correlation is the standard way in DRGEE. However, when the covariate is not time stationary, efficiency can be gained using a structured association. The goal of this paper is to extend the DRGEE estimator to allow modeling the association structure by means of either the correlation coefficient or local odds ratio. Simulation results revealed better performance of the local odds ratio parametrization, specially for small samples. The method is applied to a data set related to Rheumatic Mitral Stenosis.*

Keywords: *Correlation coefficient; Doubly robust estimators; Generalized estimating equations; Local odds ratio; Missing at random.*

3.1 Introduction

Longitudinal data arise when each individual is measured repeatedly through time. These repeated responses form a cluster and it is expected a correlation between responses within each cluster. A popular approach for the analysis of longitudinal data is the Gener-

alized Estimating Equation (GEE) method, proposed by Liang & Zeger (1986). The goal of this procedure is to estimate fixed parameters without specifying a joint distribution for the data (Nooraee *et al.*, 2014). GEE method requires only the correct specification of the response's mean structure for the parameter estimator to be consistent and asymptotically normal. An attractive feature of GEE is that the association parameters among repeated measures are taken as 'nuisance' parameters and, unlike maximum likelihood methods, the mean parameter estimates are not sensible to the specification of the association structure. Furthermore, GEE method allows marginal interpretation of the parameter of interest and it has computational simplicity.

A simple way of analyzing such correlated data is to consider an independence working assumption for the repeated responses. However, when the covariate is not time-stationary it will lead to inefficient marginal regression estimates (Lipsitz *et al.*, 1994). In the presence of time-varying covariates, efficiency can be gained assuming a different correlation structure. Nevertheless, modeling the association in ordinal data is not a simple task. Different approaches have been proposed to estimate the association parameters for ordinal responses. Lipsitz *et al.* (1994) provided moment estimators for a variety of correlation matrices, while Parsons *et al.* (2006) proposed an approach which estimates the correlation vector by minimizing the logarithm of the determinant of the covariance matrix of the fixed parameters. Instead of using correlations, Lumley (1996) proposed to use a common global odds to reduce the number of association parameters. Heagerty & Zeger (1996) extended the alternating logistic regressions, method proposed by Carey *et al.* (1993), to ordinal responses using a second set of estimating equations for the global odds ratio. Recently, Touloumis *et al.* (2013) considered a family of association models to estimate local odds ratios as a measure of association. A comparison study of different working association structures can be found in Nooraee *et al.* (2014).

In the presence of missing data, inferences under the ordinary GEE estimates are consistent if the missingness mechanism is missing completely at random (MCAR), as defined by Rubin (1976). When data is MAR, one can adopt multiple imputation GEE (MIGEE) (Little & Rubin, 1987) or a weighted version (WGEE) (Robins *et al.*, 1995).

These single robust versions of GEE for incomplete data require the correct specification of the weight model of GEE (WGEE) or the imputation model (MIGEE). Doubly robust estimators (DRGEE) (Carpenter *et al.* (2006), Tsiatis (2006), Seaman & Copas (2009), Chen & Zhou (2011)) combine ideas from these two approaches. For consistency, it requires only that the weight or the imputation model to be correctly specified, providing more flexibility for the researcher.

This work was motivated by the Rheumatic Mitral Stenosis study in which a cohort of 164 patients with rheumatic mitral stenosis (a narrowing of the mitral valve in the heart) were referred for treatment at Hospital das ClÃ**?**nicas of the Federal University of Minas Gerais, Brazil. The response of interest was the functional classification (NYHA), a major determinant of quality of life and survival of the individual. The main objective of the study was to evaluate the improvement of the functional classification over time. This study was characterized by an arbitrary pattern of missing data. The response and a particular covariate (atrial compliance) were missing for about one-third of patients and the MAR mechanism seems to be a reasonable assumption for this data.

We consider a doubly robust approach for the analysis of longitudinal ordinal data with intermittently missing response along with a key covariate that is MAR. Cumulative logit models are used for the marginal means. We extend the doubly robust estimator to accommodate two parametrizations of the association structures: one based on the correlation coefficient (Lipsitz *et al.*, 1994) and the other on local odds ratio (Touloumis *et al.*, 2013). Efficiency and accuracy of the proposed estimator are compared under these association structures.

The paper is organized as follows. In Section 3.2 is defined the notation for GEE with fully observed data and are discussed the two parametrizations of the association structure. Section 3.3 outlines the WGEE and MIGEE approaches. The proposed methodology is established in Section 3.4. A simulation study is presented in Section 3.5, in which the finite-sample biases and mainly standard errors are compared for the standard GEE, MIGEE, WGEE and doubly robust versions, under both the correlation and local odds parametrizations. Data arising from the Rheumatic Mitral Stenosis study are analyzed

in Section 3.6. The paper ends with a discussion and future research directions in Section 3.7.

3.2 Notation and GEE for Complete Data

In this section, it is described the generalized estimating equations approach for the analysis of fully observed ordinal data. Subsection 3.2.1 establishes the model and notation for longitudinal ordinal data. Subsection 3.2.2 presents two competing ways of modeling the association structure in GEE.

3.2.1 GEE for Longitudinal Ordinal Response

Let $O_{it} \in \{1, 2, \ldots, J\}$ be the ordinal response for i-th subject ($i = 1, \ldots, n$) at time t ($t = 1, \ldots, T_i,\ T_i \leq T$). As the response has J levels it can be defined as $Y_{itj} = I(O_{it} = j)$ for $j = 1, \ldots, J$, where $I(A)$ denotes the indicator function. Y_{itj} is converted into the equivalent $(J-1)$-variate vector $\boldsymbol{Y}_{it} = (Y_{it1}, \ldots, Y_{it(J-1)})^T$ and let $\boldsymbol{Y}_i = (Y_{i1}^T, \ldots, Y_{iT_i}^T)^T$ be the stacked response vector. When $J = 2$ the response is binary and $\boldsymbol{Y}_{it}$ is a scalar. Let X_i denote the time-stationary covariate for the i-th subject, and $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}^T, \ldots, \boldsymbol{Z}_{iT_i}^T)^T$ denote the covariate vector that is always observed, where $\boldsymbol{Z}_{it}$ is the covariate vector for subject i at time t.

The marginal distribution of $\boldsymbol{Y}_{it}$ is assumed to be multinomial ($\sum_{j=1}^{J} Y_{itj} = 1$), that is

$$f(\boldsymbol{Y}_{it}|X_i, \boldsymbol{Z}_{it}; \boldsymbol{\beta}) = \prod_{j=1}^{J} \mu_{itj}^{y_{itj}}, \tag{3.1}$$

where $\mu_{itj} = \mu_{itj}(\boldsymbol{\beta}) = E(Y_{itj}|X_i, \boldsymbol{Z}_{it}; \boldsymbol{\beta}) = Pr(O_{it} = j|X_i, \boldsymbol{Z}_{it}; \boldsymbol{\beta})$, is the probability of response j at time t and $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters. In this work, a cumulative logit link is used for modeling μ_{itj}, that is,

$$\text{logit}\left[Pr(O_{it} \leq j|X_i, \boldsymbol{Z}_{it}; \boldsymbol{\beta})\right] = \beta_{0j} + X_i\beta_x + \boldsymbol{Z}_{it}^T\boldsymbol{\beta}_z, \quad j = 1, \ldots, J-1. \tag{3.2}$$

Formulation in (3.2) implies a proportional odds model (McCullagh, 1980). In such

model the interpretation of $\boldsymbol{\beta}$ is the same regardless of the number of categories (i.e., it is invariant to combination of categories). A desired feature is that the exponential of the parameters is interpreted as an odds ratio (Agresti, 2013).

The main interest is to make inferences related on the regression parameters $\boldsymbol{\beta} = (\beta_{01}, \ldots, \beta_{0,J-1}, \beta_x, \boldsymbol{\beta}_z^T)^T$ associated to the $(J-1) \times 1$ marginal probability vectors

$$E(Y_{it}|X_i, \boldsymbol{Z}_{it}; \boldsymbol{\beta}) = \boldsymbol{\mu}_{it}(\boldsymbol{\beta}) = (\mu_{it1}, \ldots, \mu_{it(J-1)})^T.$$

$\boldsymbol{\mu}_{it}$ is grouped to form a vector $E(\boldsymbol{Y}_i|X_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) = \boldsymbol{\mu}_i = (\boldsymbol{\mu}_{i1}^T, \ldots, \boldsymbol{\mu}_{iT_i}^T)^T$ with the same dimension of $\boldsymbol{Y}_i$.

In order to estimate $\boldsymbol{\beta}$, generalized estimation equations are used (Liang & Zeger (1986); Lipsitz *et al.* (1994), Touloumis *et al.* (2013)), which takes the form

$$\sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}, \tag{3.3}$$

where $\boldsymbol{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T}$ and $\boldsymbol{V}_i = \boldsymbol{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is a $T_i(J-1) \times T_i(J-1)$ covariance matrix for $\boldsymbol{Y}_i$. The vector parameter $\boldsymbol{\alpha}$ expresses a 'working' assumption about the correlation/association structure.

Under mild regularity conditions, correct specification of the marginal mean model in (3.2), and provided that a $\sqrt{n}$-consistent of $\boldsymbol{\alpha}$ is available, Liang and Zeger (1986) proved that the estimator $\hat{\boldsymbol{\beta}}$, obtained by solving (3.3), is consistent and $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a multivariate normal distribution with mean vector $\boldsymbol{0}$ and covariance matrix

$$\mathrm{V}_{\boldsymbol{\beta}} = \lim_{n \to \infty} n \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_0^{-1}, \tag{3.4}$$

where $\boldsymbol{\Sigma}_0 = \sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} \boldsymbol{D}_i^T$, and $\boldsymbol{\Sigma}_1 = \sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} \mathrm{Cov}(\boldsymbol{Y}_i) \boldsymbol{V}_i^{-1} \boldsymbol{D}_i^T$.

In practice, the "sandwich" covariance matrix $\mathrm{V}_{\boldsymbol{\beta}}$ in (3.4) is calculated by ignoring the limit and replacing $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and $\mathrm{Cov}(\boldsymbol{Y}_i)$ by $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ and $(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)^T$, respectively (Touloumis *et al.*, 2013).

3.2.2 Estimation of the nuisance parameter vector and covariance matrix

The 'working' assumption is 'independence' when no correlation is assumed between pairs of the response of each individual, that is, $\boldsymbol{\alpha} = \mathbf{0}$. Independent Estimating Equations (IEE) are proved to be efficient only when covariates are constant over time or if the independence structure is actually true (Lipsitz *et al.*, 1994). In this case, score equations of the maximum likelihood method for the regression vector $\boldsymbol{\beta}$ are identical to the IEE if all observations are treated as independent. On the other hand, when there exists within-individual association or time-varying covariates, efficiency can be gained by modeling the correlation structure. Hence, a number of proposals have been formulated to model $\boldsymbol{\alpha}$. These alternatives differ in the efficiency of estimating the covariance matrix and computational simplicity.

Lipsitz *et al.* (1994) defined $\boldsymbol{\alpha}$ as the correlation coefficient, and suggested the use of the method of moments to estimate a number of correlation structures. Parsons *et al.* (2006) proposed an approach that estimate the correlation vector $\boldsymbol{\alpha}$ by minimizing an objective function $Q(\boldsymbol{\alpha}|\boldsymbol{\beta}, \boldsymbol{Y})$. Lumley (1996) proposed to use a common global odds. Heagerty & Zeger (1996) extended the alternating logistic regressions, method proposed by Carey *et al.* (1993), to ordinal responses by using a second set of estimating equations for the global odds ratio. Finally, Touloumis *et al.* (2013) identifies $\boldsymbol{\alpha}$ as a 'nuisance' parameter vector that contains the marginalized local odds ratios structure. They employed a family of association models in order to develop meaningful structures for the ordinal response.

3.2.2.1 Correlation Coefficient

Lipsitz *et al.* (1994) suggested a method that constrains the correlations at different times between two categories of the response. In their approach, the weight matrix $\boldsymbol{V}_i$ is decomposed into the form $\boldsymbol{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{F}_i^{1/2}(\boldsymbol{\beta})\boldsymbol{C}_i(\boldsymbol{\alpha})\boldsymbol{F}_i^{1/2}(\boldsymbol{\beta})$, where $\boldsymbol{F}_i$ is a matrix

containing marginal variances, $\boldsymbol{F}_{it}$, given by

$$\boldsymbol{F}_{it} = \text{diag}\left[\mu_{it1}(1-\mu_{it1}), \ldots, \mu_{it,J-1}(1-\mu_{it,J-1})\right],$$

and $\boldsymbol{C}_i$ is equal to the marginal correlation matrix. The $(J-1)\times(J-1)$ diagonal blocks of $\boldsymbol{C}_i$ are $\boldsymbol{F}_{it}^{-1/2}\boldsymbol{V}_{it}\boldsymbol{F}_{it}^{-1/2}$, with $\boldsymbol{V}_{it} = \text{diag}(\boldsymbol{\mu}_{it}) - \boldsymbol{\mu}_{it}\boldsymbol{\mu}_{it}^T$; and the $(J-1)\times(J-1)$ off-diagonal blocks of $\boldsymbol{C}_i(\boldsymbol{\alpha})$ are $\boldsymbol{\rho}_{itt'} = \boldsymbol{\rho}_{itt'}(\boldsymbol{\alpha}) = Corr(\boldsymbol{Y}_{it}, \boldsymbol{Y}_{it'})$, which represents the correlation between $\boldsymbol{Y}_{it}$ and $\boldsymbol{Y}_{it'}$, $t \neq t'$. The vector $\boldsymbol{\alpha}$ is a parameter vector associated with the model for $\boldsymbol{\rho}_{itt'}$. Define the Pearson residual vector, $\boldsymbol{e}_{it}$ as

$$\boldsymbol{e}_{it} = \boldsymbol{F}_{it}^{-1/2}\left(\boldsymbol{Y}_{it} - \boldsymbol{\mu}_{it}\right).$$

Then, it follows that

$$\boldsymbol{C}_{itt'}(\boldsymbol{\alpha}) = Corr(\boldsymbol{Y}_{it}, \boldsymbol{Y}_{it'}) = E(\boldsymbol{e}_{it}\boldsymbol{e}_{it'}^T).$$

In order to reduce the dimension of the correlation vector, Lipsitz *et al.* (1994) assumed an uniform correlation structure over the individuals. The use of the method of moments was suggested for a variety of correlation matrices such as

- *exchangeable*: $\boldsymbol{\rho}_{itt'} = \boldsymbol{\rho}$, for all $t < t'$;
- *1-dependent*: $\boldsymbol{\rho}_{it,t+1} = \boldsymbol{\rho}_t$, for $t = 1, \ldots, T-1$, and $\boldsymbol{\rho}_{itt'} = \boldsymbol{0}$ otherwise;
- *banded*: $\boldsymbol{\rho}_{itt'} = \boldsymbol{\rho}_\tau$, when $|t'-t| = \tau$, for $\tau = 1, \ldots, T-1$;
- *unstructured*: $\boldsymbol{\rho}_{itt'} = \boldsymbol{\rho}_{itt'}$, that is, no restriction are imposed on the correlations.

The estimate $\hat{\boldsymbol{\alpha}}$ is plugged into (3.3) and a solution is found for $\boldsymbol{\beta}$. The solution might be obtained by a Fisher-scoring algorithm.

The correlation coefficient parametrization ignores the scale of the response variable and may result in loss of information regarding the correlation between the variables (Lumley, 1996). Moreover, Lipsitz *et al.* (1994) observed that the 'working' correlation

matrix is not always positive definite, which may result in a breakdown of the Fisher scoring method. This is specially true for unstructured correlations matrices and small sample sizes. When the given model for $\boldsymbol{\rho}_{itt'}$ contains too many parameters, the resulting estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ may be highly variable.

3.2.2.2 Local Odds Ratio

Instead of modeling the correlation coefficients, some authors (see, for example, Lumley (1996), Touloumis *et al.* (2013)) modeled the off-diagonal matrix $\boldsymbol{V}_{itt'} = Cov(\boldsymbol{Y}_{it}, \boldsymbol{Y}_{it'})$ through the joint probability of the responses $\boldsymbol{Y}_{it}$ and $\boldsymbol{Y}_{it'}$, for $t \neq t'$. The covariance between Y_{itj} and $Y_{itj'}$ can be written as

$$\begin{aligned} Cov(Y_{itj}, Y_{it'j'}|X_i, \boldsymbol{Z}_{it}, \boldsymbol{\beta}) &= E(Y_{itj}, Y_{it'j'}|X_i, \boldsymbol{Z}_{it}, \boldsymbol{\beta}) - E(Y_{itj}|X_i, \boldsymbol{Z}_{it}, \boldsymbol{\beta})E(Y_{it'j'}|X_i, \boldsymbol{Z}_{it}, \boldsymbol{\beta}) \\ &= \mu_{itjt'j'} - \mu_{itj}\mu_{it'j'}. \end{aligned}$$

The product of the first moments can be calculated through the specification of the marginal model (3.2). The joint probabilities $\mu_{itjt'j'} = P(Y_{itj} = 1, Y_{it'j'} = 1|X_i, \boldsymbol{Z}_{it}, \boldsymbol{\beta})$ can be modeled by an association vector that describes the association structure $\forall i = 1, \ldots, n$, $t \neq t' = 1, \ldots, T_i$ and $j, j' = 1, \ldots, J$ (Touloumis *et al.*, 2013). Estimation of these joint probabilities can be achieved through global odds ratios (see, for example, Lumley (1996) and Heagerty & Zeger (1996)) or a local odds ratio (Touloumis *et al.*, 2013). Touloumis *et al.* (2013) argue that the local odds ratio provides the best parametrization in the sense that $\boldsymbol{\alpha}$ is variation independent to $\boldsymbol{\beta}$ and, in addition, it produces valid and unique positive joint probabilities.

Here, we introduce the local odds ratio of Touloumis *et al.* (2013). Denote the $L = T(T-1)/2$ time pairs $(1,2), (1,3), \ldots, (T-1, T)$, where $T = max\{T_1, \ldots, T_n\}$. Let $F_{tjt'j'} = \sum_{i=1}^{n} Y_{itj}Y_{it'j'}$ the observed frequency of the cutpoint (j, j') at the (t, t') time pair of the marginalized table. Define $\theta_{tjt'j'}$ as the local odds, that is,

$$\theta_{tjt'j'} = \frac{F_{tjt'j'}F_{t,j+1,t',j'+1}}{F_{t,j+1,t'j'}F_{tj,t',j'+1}},$$

for $j, j' = 1, \ldots, J-1$, and let $\boldsymbol{\alpha}$ be the $L \times (J-1)^2$ vector consisting of the local odds ratio

$$\boldsymbol{\alpha} = (\theta_{1121}, \ldots, \theta_{112(J-1)}, \ldots, \theta_{(T-1)1T1}, \ldots, \theta_{(T-1)(J-1)T(J-1)})^T.$$

The association vector $\boldsymbol{\alpha}$ can be estimated by fitting a loglinear model for the counts $\{F_{tjt'j'}\}$ simultaneously to all possible L marginalized contingency tables and then calculating the implied local odds ratio (Touloumis *et al.*, 2013). For notational reasons, let A and B be the row and column variable, respectively, and let G be the group variable with levels being the L ordered time pairs. Assuming a Poisson sampling scheme to the L sets of $J \times J$ contingency tables, fit the RC type model (Becker & Clogg, 1989)

$$\log(f_{tjt'j'}) = \lambda + \lambda_j^A + \lambda_{j'}^B + \lambda_{(tt')}^G + \lambda_{(tt')}^{AG} + \lambda_{(tt')}^{BG} + \varphi^{(t,t')}\nu_j^{(t,t')}\nu_{j'}^{(t,t')}, \tag{3.5}$$

where $\left\{\nu_j^{(t,t')} : j = 1, \ldots, J\right\}$ are the score parameters for the J response categories at time pair (t, t'), and $\{f_{tjt'j'} : j, j' = 1, \ldots, J\}$ are the expected frequencies. The maximum likelihood estimate of $\boldsymbol{\alpha}$ are obtained by treating the L marginalized contingency tables as independent. By imposing identifiability constraints on the regression parameters in (3.5), the resulting unrestricted local odds ratio are determined by $\log(\theta_{tj't'j'}) = \varphi^{(t,t')}(\nu_j^{(t,t')} - \nu_{j+1}^{(t,t')})(\nu_{j'}^{(t,t')} - \nu_{j'+1}^{(t,t')})$, where the intrinsic parameter $\varphi^{(t,t')}$ measures the average association of the marginalized contingency table. To increase parsimony, common unit-spaced score parameters ($\nu_j^{(t,t')} = j$) are usually assumed. The main options for the marginalized local odds ratio structures include

- *uniform*: $\log(\theta_{tj't'j'}) = \varphi$, estimates a single parameter;
- *time exchangeability*: $\log(\theta_{tj't'j'}) = \varphi^{(t,t')}$; estimates L parameters;
- *category exchangeability*: $\log(\theta_{tj't'j'}) = \varphi(\nu_j - \nu_{j+1})(\nu_{j'} - \nu_{j'+1})$; estimates $J-1$ parameters, and
- *unstructured*: $\log(\theta_{tj't'j'}) = \varphi^{(t,t')}(\nu_j^{(t,t')} - \nu_{j+1}^{(t,t')})(\nu_{j'}^{(t,t')} - \nu_{j'+1}^{(t,t')})$, that requires $L(J-1)$ parameters.

Conditionally on $\hat{\boldsymbol{\alpha}}$, and the marginal specification (3.2), the joint probabilities $\mu_{itjt'j'}$ are estimated based on the adopted local odds ratio structure using the IPFP (Iterative Proportional Fitting Procedure). This algorithm, proposed by Deming & Stephan (1940), is used to obtain $\mu_{itjt'j'}$ through the marginals μ_{itj} and $\mu_{it'j'}$. Touloumis *et al.* (2013) proved that the IPFP solution preserves local odds ratios of the initial values as long as they are positive. Hence, it is straightforward to calculate the weight matrix $\boldsymbol{V}_i$ and the estimating equations in (3.3) can be solved with respect to $\boldsymbol{\beta}$.

An advantage of the local odds ratio parametrization over the correlation is that the local odds ratio and the marginal regression vector are variation independent. This means that $\boldsymbol{\beta}$ estimates are less sensitive to a possibly wrong specification of $\boldsymbol{\alpha}$. As opposed to the correlation parametrization, the estimation of the association parameters does not depend on covariates and, as long as it is based on maximum likelihood models, the $\boldsymbol{\alpha}$ estimates are consistent under MAR. Thus, no adjustment is needed on $\boldsymbol{\alpha}$ obtained with the available data.

3.3 Available Approaches for Missing Data

Subsection 3.3.1 presents a series of assumptions related to the mechanism causing data to be missing and necessary to be considered in order to build consistent estimators. Multiple imputation and weighted generalized estimation equations are two commonly methods available for missing data under MAR mechanism. These methods are presented in Subsections 3.3.2 and 3.3.3, respectively. They serve as the basis for the construction of the doubly robust estimator, presented in Section 3.4.

3.3.1 Missing Data Framework

In this work, it will be assumed that the time-stationary covariate X_i may be missing for some subjects, whereas the explanatory variables $\boldsymbol{Z}_i$ are fully observed.

For each occasion t, it can be defined $R_{it} = 0$ if O_{it} and X_i are missing, $R_{it} = 1$ if O_{it} is missing and X_i is observed, $R_{it} = 2$ if O_{it} is observed and X_i is missing, and $R_{it} = 3$ if

O_{it} and X_i are both observed. Let $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{iT_i})^T$, and $\bar{\boldsymbol{R}}_{it} = (R_{i1}, \ldots, R_{i,t-1})$.

The marginal probability $Pr(\boldsymbol{R}_i = \boldsymbol{r}_i|\boldsymbol{O}_i, \boldsymbol{Z}_i)$ can be obtained through conditional models of the form $Pr(R_{it} = r_{it}|\bar{\boldsymbol{R}}_{it}, \boldsymbol{O}_i, \boldsymbol{Z}_i)$. This general formulation encompasses the MCAR, MAR and MNAR mechanisms. In particular, the MAR mechanism requires

$$Pr(\boldsymbol{R}_i = \boldsymbol{r}_i|\boldsymbol{O}_i, \boldsymbol{Z}_i) = Pr(\boldsymbol{R}_i = \boldsymbol{r}_i|\boldsymbol{O}_i^o, \boldsymbol{Z}_i), \tag{3.6}$$

where $\boldsymbol{O}_i^o$ denotes the observed components of $\boldsymbol{O}_i$. The following natural and additional assumption is considered

$$Pr(R_{it} = r_{it}|\bar{\boldsymbol{R}}_{it}, \boldsymbol{O}_i, \boldsymbol{Z}_i) = Pr(R_{it} = r_{it}|\bar{\boldsymbol{R}}_{it}, \bar{\boldsymbol{O}}_{it}^o, \boldsymbol{Z}_i), \tag{3.7}$$

for each time t, where $\bar{\boldsymbol{O}}_{it}^o$ is the history of the observed responses up to time $t-1$.

Let $\pi_{it} = Pr(R_{it} = 3|\boldsymbol{O}_i, \boldsymbol{Z}_i)$ be the marginal probability of observing both $\boldsymbol{O}_i$ and X_i at time t, given the entire vectors of responses and covariates. Then, π_{it} is expressed by

$$\pi_{it} = \sum_{r_{i1},\ldots,r_{i,t-1}} Pr(R_{it} = 3, R_{i,t-1} = r_{i,t-1}, \ldots, R_{i1} = r_{i1}|\boldsymbol{O}_i, \boldsymbol{Z}_i).$$

This marginal probability can be expressed in terms of the conditional probabilities $Pr(R_{it} = k|\bar{\boldsymbol{R}}_{it}, \boldsymbol{O}_i, \boldsymbol{Z}_i)$, for $k = 0, 1, 2, 3$. Throughout this paper, it will be required the so-called *positivity assumption*, that is, π_{it} must be bounded away from zero. This condition is needed in order to guarantee the existence of $\sqrt{n}$-consistent estimators of $\boldsymbol{\beta}$ (Robins *et al.*, 1995).

3.3.2 Multiple Imputation Generalized Estimating Equations

An imputation model commonly used to handle intermittently missing response and covariate is the imputation using chained equations (van Buuren *et al.* (1999), van Buuren (2007)), which is more commonly referred to as full conditional specification (FCS). This approach specifies conditional distributions for each incomplete variable, conditional on all others variables in the imputation model. Starting from an initial imputation, FCS

draws imputations by iterating over the conditional densities.

Denote by $\tilde{\boldsymbol{\beta}}_m$ and $\tilde{\boldsymbol{U}}_m$, respectively, the estimate of $\boldsymbol{\beta}$ and its covariance matrix from the GEE analysis of the m-th completed data set, $(m = 1, \ldots, M)$. Following Rubin (1987), the combined point estimate for the parameter of interest $\boldsymbol{\beta}$ based on MI is simply the average of the M complete-data point estimates

$$\hat{\boldsymbol{\beta}}_{MI} = \frac{1}{M} \sum_{m=1}^{M} \tilde{\boldsymbol{\beta}}_m,$$

and an estimate of the covariance matrix of $\hat{\boldsymbol{\beta}}_{MI}$ is given by

$$\widehat{\boldsymbol{U}}_{MI} = \tilde{\boldsymbol{W}} + \left(\frac{M+1}{M}\right) \tilde{\boldsymbol{B}},$$

where

$$\tilde{\boldsymbol{W}} = \frac{1}{M} \sum_{m=1}^{M} \tilde{\boldsymbol{U}}_m \quad \text{and} \quad \tilde{\boldsymbol{B}} = \frac{1}{M-1} \sum_{m=1}^{M} (\tilde{\boldsymbol{\beta}}_m - \hat{\boldsymbol{\beta}}_{\boldsymbol{MI}})(\tilde{\boldsymbol{\beta}}_m - \hat{\boldsymbol{\beta}}_{\boldsymbol{MI}})^T.$$

3.3.3 Weighted Generalized Estimating Equations

Robins *et al.* (1995) proposed a class of weighted estimating equations to allow for MAR mechanism. In binary longitudinal data, Chen & Zhou (2011) extended the method to accommodate arbitrary patterns of both missing response and covariate. Their method was adapted here for longitudinal ordinal responses.

Define a weight matrix $\boldsymbol{\Delta}_i = [\delta_{itt'}]_{T_i(J_i-1)\times T_i(J_i-1)}$, $t = 1, \ldots, T_i, t' = 1, \ldots, T_i$, where $\delta_{itt'} = \{I(R_{it} = 1, R_{it'} = 3) + I(R_{it} = 3, R_{it'} = 3)\} / \pi_{itt'}$ for $t \neq t'$, $\delta_{itt} = I(R_{it} = 3)/\pi_{it}$, and $\pi_{itt'} = Pr(R_{it} = 1, R_{it'} = 3|\boldsymbol{O}_i, \boldsymbol{Z}_i) + Pr(R_{it} = 3, R_{it'} = 3|\boldsymbol{O}_i, \boldsymbol{Z}_i)$. In order to construct the weight matrix $\boldsymbol{\Delta}_i$ the conditional probabilities $Pr(R_{it} = k|\bar{\boldsymbol{R}}_{it}, \boldsymbol{O}_i, \boldsymbol{Z}_i)$ are decomposed as the product of two separated logistic models. The first one models the probability of observing the potentially missing covariate X_i, whereas, the latter models the probability of observing Y_{it} conditional on observed responses and covariates up to time $t - 1$.

The main idea of the weighted generalized estimating equations (WGEE) lies on weighting the individual contribution to the estimating equation by introducing the weight matrix $\boldsymbol{\Delta}_i$ into the covariance matrix $\boldsymbol{V}_i$. This task can be accomplished by different ways depending on the parametrization of the association structure. The general WGEE for $\boldsymbol{\beta}$ are given by

$$\sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\psi}) = \sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{M}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}. \tag{3.8}$$

When the correlation coefficient is adopted, the $\boldsymbol{M}_i$ matrix is decomposed as $\boldsymbol{M}_i = \boldsymbol{F}_i^{-1/2}(\boldsymbol{C}_i^{-1} \cdot \boldsymbol{\Delta}_i)\boldsymbol{F}_i^{-1/2}$, where $\boldsymbol{A} \cdot \boldsymbol{B} = [a_{it} \cdot b_{it}]$ denotes the Hadamard product of matrices $\boldsymbol{A} = [a_{it}]$ and $\boldsymbol{B} = [b_{it}]$. Under the local odds approach, the matrix $\boldsymbol{M}_i$ is given by $\boldsymbol{M}_i = \boldsymbol{V}_i^{-1} \cdot \boldsymbol{\Delta}_i$.

Conditionally on a consistent estimate of the correlation/association structure $\boldsymbol{\alpha}$, a consistent estimate for $\boldsymbol{\beta}$ can be obtained by solving (3.8), under the correct specification of the missing data model.

In the presence of missing data, a consistent estimate for the correlation parameters can be obtained by defining a weighted observed pair of Pearson residual vector as $e_{it}^* = e_{it}(I(R_{it} = 3)/\pi_{it})$. Then, a moment-based estimator can be constructed using the weighted pair contribution. For instance, an exchangeable correlation estimate can be obtained through

$$\hat{\rho}_{itt'} = \frac{1}{\sum_{i=1}^{n} \frac{1}{2} T_i(T_i - 1) - p} \sum_{i=1}^{n} \sum_{t'>t} e_{it}^* e_{it'}^{*T} \frac{\pi_{it}\pi_{it'}}{\pi_{itt'}}.$$

It is easy to show that $\hat{\rho}_{itt'}$ is an unbiased estimator.

3.4 Doubly Robust GEE for Longitudinal Ordinal Data

Some authors (e.g., Scharfstein *et al.* (1999), Tsiatis (2006)) noted that adding a term of expectation zero, say $\phi(\cdot)$, to the inverse probability weighted estimators would still result in consistent estimates under the MAR mechanism. The solutions of these augmented estimating equations give rise to the so-called *doubly robust* estimators.

Following Chen & Zhou (2011), the optimal ϕ_{opt} for missing response and covariate is given by $\phi_{opt} = E_{(\boldsymbol{Y}_i^m, X_i^m | \boldsymbol{Y}_i^o, X_i^o, \boldsymbol{Z}_i, \boldsymbol{R}_i)} \{\boldsymbol{D}_i \boldsymbol{N}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)\}$, where $\boldsymbol{Y}_i^m$ and X_i^m denote the missing components of $\boldsymbol{Y}_i$ and X_i, respectively. When the correlation coefficient parametrization is adopted, $\boldsymbol{N}_i$ is defined as $\boldsymbol{N}_i = \boldsymbol{F}_i^{-1/2} \left\{\boldsymbol{C}_i^{-1} \cdot (\boldsymbol{1}\boldsymbol{1}^T - \boldsymbol{\Delta}_i)\right\} \boldsymbol{F}_i^{-1/2}$, where $\boldsymbol{1}$ is a vector of 1's of length $T_i(J-1)$. With local odds, $\boldsymbol{N}_i$ can be defined as $\boldsymbol{N}_i = \boldsymbol{V}_i^{-1} \cdot (\boldsymbol{1}\boldsymbol{1}^T - \boldsymbol{\Delta}_i)$.

Conditionally on a consistent estimate of the correlation/association structure $\boldsymbol{\alpha}$, an improved estimate for $\boldsymbol{\beta}$ can then be obtained by solving the estimating equations

$$\sum_{i=1}^{n} \left[\boldsymbol{D}_i \boldsymbol{M}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) + E_{(\boldsymbol{Y}_i^m, X_i^m | \boldsymbol{Y}_i^o, X_i^o, \boldsymbol{Z}_i, \boldsymbol{R}_i)} \{\boldsymbol{D}_i \boldsymbol{N}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)\}\right] = \boldsymbol{0}. \tag{3.9}$$

The estimator for $\boldsymbol{\beta}$ in (3.9) is doubly-robust in the sense that it is consistent if *at least one* of the missing data model or the covariate model is correctly specified (Chen & Zhou, 2011).

The referred expectation in the second part of (3.9) is over the conditional distribution of $(\boldsymbol{Y}_i^m, X_i^m | \boldsymbol{Y}_i^o, X_i^o, \boldsymbol{Z}_i, \boldsymbol{R}_i)$, which can be written as

$$\begin{aligned} P(\boldsymbol{Y}_i^m = \boldsymbol{y}_i^m, X_i^m = x_i^m | \boldsymbol{Y}_i^o, X_i^o, \boldsymbol{Z}_i, \boldsymbol{R}_i; \boldsymbol{\beta}^*, \boldsymbol{\gamma}) &= P(\boldsymbol{Y}_i^m = \boldsymbol{y}_i^m, X_i^m = x_i^m | \boldsymbol{Y}_i^o, X_i^o, \boldsymbol{Z}_i; \boldsymbol{\beta}^*, \boldsymbol{\gamma}) \\ &= P(\boldsymbol{Y}_i^m = \boldsymbol{y}_i^m | \boldsymbol{Y}_i^o, X_i = x_i, \boldsymbol{Z}_i; \boldsymbol{\beta}^*) \\ &\quad \times P(X_i^m = x_i^m | \boldsymbol{Y}_i^o, X_i^o, \boldsymbol{Z}_i; \boldsymbol{\gamma}). \end{aligned}$$

The multivariate distribution $P(\boldsymbol{Y}_i^m = \boldsymbol{y}_i^m | \boldsymbol{Y}_i^o, X_i = x_i, \boldsymbol{Z}_i; \boldsymbol{\beta}^*)$ is expressed by a product of univariate ordinal models. When X is discrete, the second term in (3.9) can be written as

$$E_{(\boldsymbol{Y}_i^m, X_i^m | \boldsymbol{Y}_i^o, X_i^o, \boldsymbol{Z}_i, \boldsymbol{R}_i)} \{\boldsymbol{D}_i \boldsymbol{N}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)\} = \sum_{(\boldsymbol{y}_i^m, x_i^m)} w_{ixy} \{\boldsymbol{D}_i \boldsymbol{N}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)\},$$

where the weight w_{ixy} is given by

$$w_{ixy} = P(\boldsymbol{Y}_i^m = \boldsymbol{y}_i^m | \boldsymbol{Y}_i^o, X_i = x_i, \boldsymbol{Z}_i; \boldsymbol{\beta}^*) \times P(X_i^m = x_i^m | \boldsymbol{Y}_i^o, X_i^o, \boldsymbol{Z}_i; \hat{\boldsymbol{\gamma}}).$$

In the case of X continuous, the second term in (3.9) takes the form

$$E_{(\boldsymbol{Y}_i^m, X_i^m | \boldsymbol{Y}_i^o, X_i^o, \boldsymbol{Z}_i, \boldsymbol{R}_i)} \{\boldsymbol{D}_i \boldsymbol{N}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)\} = \int_{(\boldsymbol{Y}_i^m, X_i^m)} w_{ixy} . \{\boldsymbol{D}_i \boldsymbol{N}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)\} \, d\boldsymbol{Y}_i^m X_i^m,$$

This expectation can be cumbersome, depending on the missing data pattern. In such case, instead of using numerical integration, a Monte Carlo method can be applied to approximate the corresponding integral.

Inspired by doubly robust ideas, we constructed the following estimator for the correlation structure

$$\hat{\rho}_{itt'} = \frac{\omega}{n^\dagger} \sum_{i=1}^{n} \sum_{t'>t} e_{it}^* e_{it'}^{*T} \frac{\pi_{it} \pi_{it'}}{\pi_{itt'}} + \frac{(1-\omega)}{n^\dagger} \sum_{i=1}^{n} \sum_{t'>t} \left[\sum_{(y_i^m, x_i^m)} w_{ixy} \hat{e}_{it} \hat{e}_{it'}^T \right],$$

where $n^\dagger = \sum_{i=1}^{n} \frac{1}{2} T_i (T_i - 1) - p$ and $0 \leq \omega \leq 1$.

A sandwich estimator for the standard error of $\hat{\boldsymbol{\beta}}$ is given in Appendix A.

3.5 Simulation Study

In this section, a simulation study is presented in order to investigate the performance of the proposed method under the two parametrizations of the association vector as well as its robustness to misspecification of the predictive models. It is considered a study with $T_i = T = 3$ repeated ordinal measures (with three categories) and two covariates (one quantitative and other qualitative). The true marginal model is

$$logit \; Pr(O_{it} \leq j | X_i, Z_{it}) = \beta_{0j} + \beta_1 X_i + \beta_2 Z_{it}, \quad j = 1, 2. \tag{3.10}$$

where $Z_{it} \sim N(0, 1/2)$ for $t = 1, 2, 3$.

The binary covariate X_i may be missing for some subjects and is generated according to

$$logit \; Pr(X_i = 1 | Z_{i1}) = \gamma_0 + \gamma_1 Z_{i1}. \tag{3.11}$$

It is assumed that $\beta_{01} = -0.4$, $\beta_{02} = 1.2$, $\beta_1 = -0.35$, $\beta_2 = 0.35$, $\gamma_0 = log(1)$, $\gamma_1 = 2$.

The correlated ordinal responses were generated using the NORTA method (Touloumis, 2015) with constant correlation between the latent vectors as $\rho = 0.7$.

In order to model R_{it}, two new indicators were defined. Let R_i^x the indicator of observing X_i and R_{it}^y the indicator of observing O_{it}. The response variable in the first time occasion was allowed to be fully observed. The model for R_i^x was defined as

$$log\left(\frac{Pr(R_i^x = 1)}{Pr(R_i^x = 0)}\right) = \psi_0^x + \psi_1^x O_{i1} + \psi_2^x Z_{i1}, \tag{3.12}$$

and the model for R_{it}^y was taken as

$$log\left(\frac{Pr(R_{it}^y = 1)}{Pr(R_{it}^y = 0)}\right) = \psi_0^y + \psi_1^y O_{i,t-1}^* + \psi_2^y I(R_{i,t-1}^y = 1) + \psi_3^y Z_{it}, \quad t = 2, 3, \tag{3.13}$$

where $O_{i,t-1}^* = O_{i,t-1}$, if $O_{i,t-1}$ is observed and 0 otherwise. The true values are: $\psi_0^x = 1.2$, $\psi_1^x = -1.5$, $\psi_2^x = -1.5$, $\psi_0^y = 0.6$, $\psi_1^y = -1.5$, $\psi_2^y = 2.5$, and $\psi_3^y = -1.3$. It was observed about 30% of missing observations under this setup.

For comparison purposes, it was considered ordinary GEE for the complete and available data, respectively, weighted GEE (WGEE), multiple imputation (MIGEE) by chained equations with $M = 10$ multiple imputations, and the doubly robust versions (DRGEE). The primary goal of this simulation was to compare the performance of the above mentioned methods under the correlation and local odds ratio parametrizations. Three correlation structures (independent – ind, exchangeable – exch, and unstructured – unst) and four local odds ratio structures (uniform – unif, category exchangeability – cat.exch, time exchangeability – time.exch, and unstructured – RC) are compared. Under independence, the estimates from the two parametrization are identical.

In order to investigate robustness of these methods, the predicted models were also misspecified by omitting the covariate Z_1 from the covariate model (3.11) or the missing data model (3.12). Correctly specified models are indicated with a *plus* sign ('x^+' and 'r^+', for the covariate and missing data model, respectively) and incorrectly specified models are indicated with a *minus* sign ('x^-' and 'r^-', for the covariate and missing data model, respectively).

Let $S = 1000$ be the total number of Monte Carlo replications. Whenever, in a given iteration, a working association structure (C) failed to converge, a new sample data were generated. Denote by $\hat{\beta}_r^C$ the corresponding GEE estimator at the r-th Monte Carlo replication and let $\hat{\beta}^C$ be the arithmetic mean, $\hat{\beta}^C = 1/S \sum_{r=1}^{S} \hat{\beta}_r^C$. To evaluate the consistency of the competing methods the relative bias, defined as $100 \times (\hat{\beta}^C - \beta)/\beta$, was calculated for each parameter. Interest is in quantifying the gain in efficiency by the working association structures over the independence structure. The Monte Carlo relative efficiency was defined as $\sum_{r=1}^{S} \hat{EP}(\hat{\beta}_r^I) / \sum_{r=1}^{S} \hat{EP}(\hat{\beta}_r^C)$, where $\hat{EP}(\hat{\beta}_r^C)$ is the standard error of $\hat{\beta}_r^C$ based on the estimated robust covariance matrix under the (C) working association structure. The estimated coverage probability for a nominal 95% level based on the asymptotic normality of the GEE estimators is also reported. All methods were implemented in the software R (R Core Team, 2015). For the generation of the multiple imputed values, the package *mice* (van Buuren & Groothuis-Oudshoorn, 2011) was used. Simulation was conducted for the sample sizes $n = 50, 150, 300$ and 600. Here, results are presented only for sample size $n = 300$ subjects. For this sample size, no convergence issues was observed for the independent structure, as well as the local odds ratio structures. The convergence rate for the exchangeable and unstructured correlation matrices was 96% and 89%, respectively.

Table 3.1 presents the simulation results associated with ordinary GEE for available data and incorrectly specified methods, those in which the covariate Z_1 was omitted from their predictive models. For each method, the first three lines refer to correlation structures and the last four refer to local odds ratio structures. These seven structures differ in terms of the number of parameters being estimated as well as the restrictions placed on associations/correlations between the response indicators at different time pairs.

Regarding the GEE for available data, the missing data impact on bias can be noticed for all parameters: higher biases being observed in the first intercept, followed by the parameter associated with the covariate Z. The impact of the bias on parameter estimates was also clearly noticed by the low coverage rates. Still considering the available data, it is worth noting that the bias for the parameter associated with covariate Z was reduced by

Table 3.1: Evaluation criteria for misspecified models. Results for $n = 300$ and $S = 1000$ simulations.

	Relative Bias				Relative Efficiency				Empirical Coverage			
Structure	β_{01}	β_{02}	X	Z	β_{01}	β_{02}	X	Z	β_{01}	β_{02}	X	Z
						Available						
ind	121.1	-31.5	-23.9	60.2	1.00	1.00	1.00	1.00	0.15	0.36	0.93	0.73
exch	94.0	-25.2	-26.4	32.3	1.02	1.01	1.01	1.20	0.35	0.55	0.94	0.86
unst	94.3	-25.9	-31.4	26.7	1.03	1.01	1.02	1.19	0.31	0.48	0.90	0.90
unif	93.7	-23.9	-27.8	26.4	1.02	0.99	1.01	1.23	0.33	0.57	0.92	0.88
cat.exch	93.8	-23.9	-28.0	25.5	1.02	0.99	1.01	1.24	0.33	0.57	0.92	0.89
time.exch	93.7	-24.0	-27.9	26.3	1.02	0.99	1.01	1.24	0.36	0.56	0.93	0.89
RC	93.1	-24.2	-30.8	22.5	1.02	1.00	1.01	1.24	0.34	0.56	0.92	0.91
						WGEE(r^-)						
ind	19.9	-5.4	-28.9	32.8	1.00	1.00	1.00	1.00	0.94	0.93	0.92	0.89
exch	17.9	-5.1	-29.9	22.2	1.00	0.99	1.00	1.17	0.94	0.94	0.94	0.91
unst	21.0	-6.5	-35.0	15.2	1.01	0.99	1.01	1.16	0.93	0.92	0.91	0.94
unif	18.8	-5.1	-28.6	22.7	1.00	0.99	1.00	1.16	0.95	0.94	0.92	0.92
cat.exch	19.0	-5.2	-28.8	22.0	1.00	0.99	1.00	1.17	0.95	0.94	0.92	0.92
time.exch	18.3	-5.0	-27.7	22.8	1.00	0.99	1.00	1.17	0.95	0.93	0.93	0.92
RC	18.7	-5.6	-31.4	19.8	1.01	0.99	1.00	1.17	0.94	0.93	0.92	0.91
						MIGEE(x^-)						
ind	20.9	-7.1	-48.0	-8.3	1.00	1.00	1.00	1.00	0.93	0.93	0.92	0.94
exch	19.9	-6.8	-47.9	-1.6	0.99	0.99	1.00	1.22	0.94	0.94	0.93	0.95
unst	21.5	-7.6	-51.9	-5.9	0.99	1.00	1.00	1.22	0.94	0.93	0.92	0.94
unif	19.7	-6.7	-46.8	-2.2	1.00	1.00	1.00	1.22	0.95	0.94	0.93	0.94
cat.exch	19.9	-6.8	-47.0	-3.1	1.00	1.00	1.00	1.23	0.95	0.94	0.93	0.94
time.exch	20.0	-6.8	-46.6	-2.8	1.00	1.00	1.00	1.23	0.94	0.94	0.94	0.94
RC	20.1	-7.3	-49.7	-5.8	1.00	1.01	1.00	1.23	0.94	0.93	0.92	0.93
						DRGEE(x^-, r^-)						
ind	17.0	-5.5	-41.4	-8.9	1.00	1.00	1.00	1.00	0.93	0.92	0.89	0.94
exch	15.7	-5.1	-40.7	-2.4	0.99	0.99	1.00	1.14	0.94	0.94	0.91	0.94
unst	18.5	-6.3	-45.5	-8.5	1.00	1.00	1.01	1.14	0.93	0.92	0.89	0.95
unif	15.5	-5.0	-39.1	-3.1	1.00	0.99	1.00	1.14	0.94	0.94	0.91	0.94
cat.exch	15.6	-5.1	-39.3	-3.5	1.00	0.99	1.00	1.15	0.95	0.94	0.91	0.94
time.exch	15.0	-4.8	-38.1	-2.5	1.00	0.99	1.00	1.14	0.94	0.92	0.91	0.94
RC	15.5	-5.5	-41.1	-5.2	1.00	0.99	1.00	1.14	0.94	0.93	0.91	0.93

"+" indicates correctly specified model and "−" indicates misspecified model omitting the Z_1 predictor

more than a half when the association structure is modeled. All methods being compared exhibited bias when their predictive models are incorrectly specified, although the bias for the two intercepts was considerably reduced. For the parameters associated with the covariates, the bias of the WGEE method was of the same magnitude as those provided the by ordinary GEE, and the performance of DRGEE was slightly superior to MIGEE.

Independent estimating equations are efficient for the intercept parameters and regression coefficient associated with the baseline covariate X. As expected, the gain in efficiency by modeling the association structure occurs only for the time-varying covariate Z. Comparing to the independence structure, the gain in efficiency ranged from 14% on average for DRGEE, 23% for multiple imputation and about 17% for WGEE.

All misspecified methods presented empirical coverage rates below nominal level although they are somewhat close to the expected value in some cases, particularly for MIGEE.

Table 3.2 presents the simulation results for complete data in addition to correctly specified methods. In terms of bias, there is no clear distinction between the two approaches, except when an unstructured matrix is chosen, which causes an increase in bias, especially for WGEE and DRGEE when only the weight model is correctly specified. For the scenario under consideration, the results suggest that the performances of both WGEE and DRGEE are slightly superior to multiple imputation, especially for estimates associated with covariate X.

Compared to independence structure, all other association structures presented smaller standard errors for the time-dependent covariate Z. Although they are very similar within each method, the largest gain in efficiency, around 23%, is obtained for the MIGEE method, followed by WGEE with 19%. For DRGEE estimators, it ranged from 14% to 18%.

All correctly specified methods showed coverage rates close to the nominal levels for all association structures. In terms of empirical bias, relative efficiency and empirical coverage, there was no clear distinction between the correlation and local odds parametrizations for $n = 300$ subjects. For the scenario under consideration, an exchangeable correlation

Table 3.2: Evaluation criteria for correctly specified models. Results for $n = 300$ and $S = 1000$ simulations.

	Relative Bias				Relative Efficiency				Empirical Coverage			
Structure	β_{01}	β_{02}	X	Z	β_{01}	β_{02}	X	Z	β_{01}	β_{02}	X	Z
						Complete						
ind	2.1	0.1	-1.1	-0.5	1.00	1.00	1.00	1.00	0.96	0.94	0.95	0.95
exch	0.3	0.6	-0.2	1.3	0.99	0.99	0.99	1.25	0.96	0.95	0.97	0.96
unst	0.9	0.0	-3.4	-1.7	1.00	1.00	1.01	1.25	0.95	0.94	0.95	0.96
unif	0.1	0.7	0.7	0.3	1.00	1.00	1.00	1.27	0.96	0.95	0.94	0.94
cat.exch	0.2	0.7	0.6	-0.2	1.00	1.00	1.00	1.27	0.96	0.95	0.95	0.94
time.exch	0.2	0.3	0.1	0.3	1.00	1.00	1.00	1.27	0.95	0.95	0.96	0.94
RC	0.2	0.3	-0.8	-1.9	1.00	1.00	1.00	1.27	0.95	0.94	0.95	0.94
						WGEE(r^+)						
ind	3.6	-0.4	-1.8	-0.7	1.00	1.00	1.00	1.00	0.96	0.95	0.94	0.96
exch	1.3	0.3	0.2	0.6	0.99	0.98	1.00	1.19	0.96	0.96	0.96	0.96
unst	6.2	-1.6	-6.8	-4.1	1.00	1.00	1.01	1.19	0.95	0.94	0.93	0.96
unif	2.4	-0.1	-0.3	-0.4	0.99	0.99	1.00	1.18	0.97	0.96	0.96	0.95
cat.exch	2.6	-0.2	-0.5	-1.0	0.99	0.99	1.00	1.18	0.97	0.95	0.96	0.95
time.exch	1.4	0.3	1.8	1.1	0.99	0.99	1.00	1.19	0.97	0.97	0.96	0.96
RC	2.5	-0.5	-2.5	-1.6	0.99	0.99	1.00	1.19	0.96	0.96	0.95	0.94
						MIGEE(x^+)						
ind	5.3	-1.3	-7.1	-0.4	1.00	1.00	1.00	1.00	0.96	0.94	0.94	0.94
exch	3.5	-0.8	-5.6	0.7	0.99	0.99	1.00	1.22	0.95	0.95	0.96	0.95
unst	4.6	-1.5	-9.4	-3.7	0.99	0.99	1.00	1.21	0.95	0.95	0.95	0.94
unif	3.9	-0.8	-5.4	0.2	1.00	0.99	1.00	1.22	0.96	0.95	0.95	0.94
cat.exch	4.2	-0.9	-5.7	-0.7	1.00	0.99	1.00	1.23	0.96	0.95	0.96	0.94
time.exch	4.2	-1.0	-5.2	-0.2	1.00	0.99	1.00	1.22	0.95	0.96	0.96	0.94
RC	3.7	-1.3	-8.0	-3.1	1.00	1.00	1.00	1.23	0.94	0.95	0.95	0.92
						DRGEE(x^+, r^+)						
ind	1.0	0.4	-1.7	-1.4	1.00	1.00	1.00	1.00	0.95	0.95	0.93	0.95
exch	-0.8	0.9	0.3	1.4	0.99	0.99	1.00	1.16	0.95	0.95	0.96	0.95
unst	2.4	-0.5	-5.7	-3.6	1.00	0.99	1.01	1.16	0.94	0.94	0.94	0.94
unif	0.0	0.7	-0.6	-0.1	1.00	0.99	1.00	1.16	0.95	0.95	0.95	0.94
cat.exch	0.1	0.6	-0.8	-0.6	1.00	0.99	1.00	1.17	0.95	0.95	0.95	0.94
time.exch	-0.9	1.2	2.0	1.4	0.99	0.99	1.00	1.16	0.95	0.95	0.95	0.94
RC	-0.2	0.3	-2.3	-1.8	1.00	1.00	1.00	1.16	0.95	0.94	0.94	0.93
						DRGEE(x^-, r^+)						
ind	2.2	0.0	-4.7	-1.5	1.00	1.00	1.00	1.00	0.95	0.95	0.95	0.94
exch	0.2	0.6	-2.5	1.3	0.98	0.99	1.00	1.17	0.96	0.96	0.96	0.95
unst	4.7	-1.3	-11.1	-3.7	1.00	1.01	1.03	1.18	0.94	0.94	0.94	0.94
unif	1.2	0.3	-3.3	-0.1	1.00	0.99	1.00	1.17	0.96	0.96	0.97	0.94
cat.exch	1.3	0.3	-3.5	-0.5	1.00	0.99	1.00	1.18	0.96	0.96	0.97	0.94
time.exch	0.3	0.7	-1.2	1.3	1.00	1.00	1.00	1.18	0.96	0.96	0.97	0.94
RC	1.2	-0.1	-5.5	-1.6	1.00	1.00	1.00	1.17	0.95	0.96	0.96	0.93
						DRGEE(x^+, r^-)						
ind	1.0	0.2	-2.5	-1.8	1.00	1.00	1.00	1.00	0.95	0.94	0.93	0.94
exch	-0.4	0.6	-1.4	0.5	0.99	0.99	1.00	1.15	0.94	0.95	0.96	0.95
unst	2.1	-0.5	-5.7	-5.6	1.00	0.99	1.01	1.14	0.93	0.94	0.94	0.95
unif	-0.2	0.7	-0.6	0.0	1.00	0.99	1.00	1.15	0.95	0.95	0.95	0.94
cat.exch	-0.1	0.6	-0.8	-0.5	1.00	0.99	1.00	1.15	0.94	0.95	0.95	0.94
time.exch	-0.4	0.8	0.1	0.5	1.00	0.99	1.00	1.14	0.94	0.94	0.95	0.94
RC	0.0	0.1	-3.3	-2.2	1.00	1.00	1.00	1.14	0.94	0.94	0.94	0.93

"+" indicates correctly specified model and "−" indicates misspecified model omitting the Z_1 predictor

structure or an uniform local odds structure resulted in good marginal mean estimates.

The simulation results for $n = 50, 150$ and 600 are presented in the Appendix B. With sample sizes $n = 50$ or $n = 150$ subjects, convergence issues were observed for unstructured correlation matrices (convergence rate of 8% and 59% for $n = 50$ and $n = 150$, respectively) and exchangeable correlation matrices (convergence rate of 59% and 89% for n = 50 and n = 150, respectively). The local parameter odds presented low convergence rate only for the unstructured matrix and sample size $n = 50$, where 67% of the samples converged. In general, for small sample sizes, simulation results suggest that the local odds parametrization outperforms the correlation coefficient parametrization in terms of bias and convergence issues.

3.6 Data Analysis: Functional Classification in Rheumatic Mitral Stenosis

A cohort of 164 patients with rheumatic mitral stenosis who were referred for treatment at Hospital das ClÃ?nicas of the Federal University of Minas Gerais, Brazil, was selected for a mitral valve invervention. Patients were included before intervention and then followed up in the outpatient clinic every 4 months according to their clinical status. The first three measurements were available for analysis.

Mitral stenosis is a narrowing of the mitral valve in the heart caused by rheumatic disease, which restricts the flow of blood through the valve. The main clinical manifestation of this disease is shortness of breath, classified in four categories based on how much the patients are limited during physical activity. The response of interest is The New York Heart Association (NYHA) Functional Classification, that provides a simple way of classifying the extent of shortness of breath. Patients with no symptoms and no limitation in ordinary physical activity were classified in class I; slight limitation of physical activity in class II; marked limitation of physical activity in class III; and patients with severe limitations resulting in inability to carry on any physical activity without discomfort in class IV. Only one patient were classified into class IV in the followup evaluation and

hence class IV was combined with class III in the analysis. Thus, the ordinal response was defined as (1: if class I, 2: if class II, 3 if class III or class IV).

Percutaneous mitral valvuloplasty (PMV) is an effective treatment for stretching the stenosed mitral valve. This procedure is carried out by inserting a catheter with a balloon at its tip to open the narrowed mitral valve. This procedure causes improvement of the functional class in the majority of the patients. A number of patient characteristics were measured at baseline, such as atrial compliance (Cn: defined as 1, if ≤ 4 mL/mmHg, and 0 otherwise), cardiac rhythm (1: normal; 0: atrial fibrillation), morphological features of the mitral valve expressed as an echocardiographic score (varying from 4 to 12), mitral valve area (in cm^2), pressure transmitral gradients (in mmHg), and pulmonary artery pressure (in mmHg). Some variables, measured after the procedure, are: the success of the procedure to open the mitral valve without complications (1: success; 0: otherwise), long-term event-free survival (1: event-free; 0: otherwise), mitral valve area (in cm^2), pressure transmitral gradients (in mmHg), and pulmonary artery pressure at the follow-up appointment (in mmHg). Figure 3.1 shows the observed longitudinal profile of the observed functional class. The proportion of patients classified in class I is clearly higher after the intervention.

This study was characterized by an arbitrary pattern of missing data. The response were fully observed for 125 patients, 29 of them had only the first two measurements, 1 patient had only the first data collected, and for 9 patients the second occasion was missing. There was no missing data in response at baseline. Some collected variables, such as success of the procedure and long term events were responsible for the missingness at followup. A baseline covariate of particular interest, atrial compliance, were missing for 54 (32.9%) of the patients. Among the reasons for not observing such predictor it can be included morphological characteristics of the mitral valve and valve calcification. Therefore a MAR mechanism seems to be a reasonable assumption for this data set.

The models used for analysis are described below. For the ordinal response, the

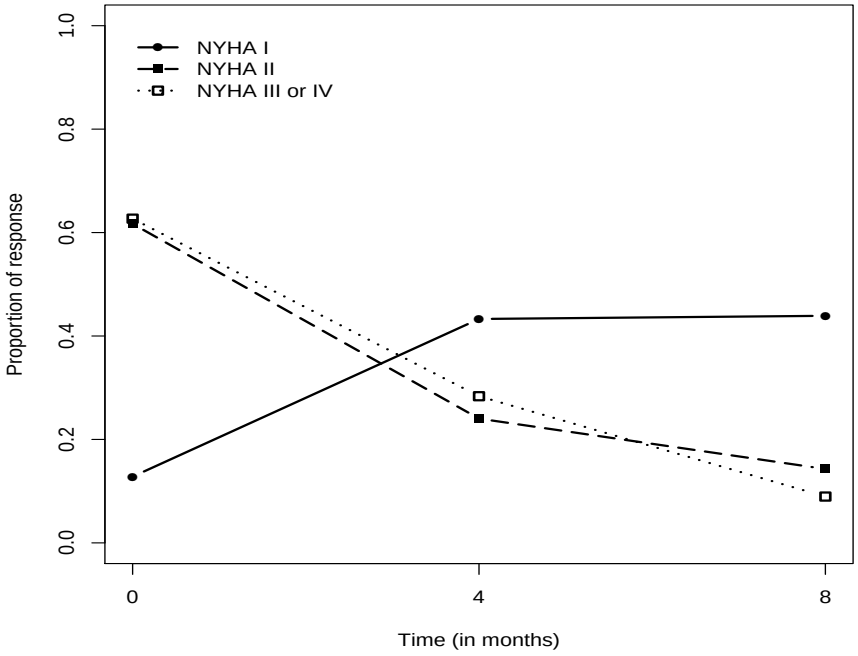

Figure 3.1: Observed longitudinal profile of functional class.

following proportional odds model was used

$$logit\ Pr(NYHA_{it} \leq j|\boldsymbol{u_{it}}) = \beta_{0j} + \boldsymbol{u}_{it}^T\boldsymbol{\beta}, \quad j = 1, 2, \quad t = 1, 2, 3, \tag{3.14}$$

where $\boldsymbol{u}_{it}$ is the covariate vector at time t, and it is formed by time, atrial compliance, echocardiographic score, and success of the procedure.

When using WGEE or DRGGE, it is necessary to correctly model the missingness mechanism in order to obtain consistent estimates of $\boldsymbol{\beta}$. For the missing data process, R_{it}^y was defined as the indicator of observing the response $NYHA_{it}$ and R_i^x was defined as the indicator of observing the baseline atrial compliance. The probability of observing the response was modeled as

$$logit\ Pr(R_{it}^y = 1) = \psi_0^y + \boldsymbol{w}_{it}^{yT}\boldsymbol{\psi}^y, \quad t = 2, 3, \tag{3.15}$$

where the vector $\boldsymbol{w}_{it}^y$ includes: success of the procedure, long term events, and the previous

indicator of missing data. The model for R_i^x was specified as

$$logit\ Pr(R_i^x = 1) = \psi_0^x + \boldsymbol{w}_i^{xT}\boldsymbol{\psi}^x, \tag{3.16}$$

where the vector $\boldsymbol{w}_i^x$ includes: the baseline response, echocardiographic score, valve calcification and ECG rhythm.

For the covariate, the following model was built

$$logit\ Pr(Cn = 1) = \boldsymbol{v}_i^T\boldsymbol{\gamma}, \tag{3.17}$$

where $\boldsymbol{v}_i^T$ included the baseline right ventricular systolic pressure, mean gradient and mitral valve area. An imputation model for NYHA was also specified including the covariates in (3.14) in addition to the baseline right ventricular systolic pressure and the response history.

Here, four methods were used to analyze the data. Results are shown in Table 3.3. The first method is the usual GEE method using the available data; the second is the weighted method (WGEE) using models (3.15) and (3.16) for the weights; the third is the multiple imputation by chained equation (MIGEE); and the fourth, labeled DRGEE, is the doubly robust method using (3.15) and (3.16) for the missing data process and (3.17) for the covariate models. In order to account for the dependence structure of the repeated measures the correlation coefficient and the local odds ratio parametrization were both initially applied with the same association structures presented in the simulation. Convergence issues were observed for unstructured correlation matrix. Motivated by the simulation results suggesting better performance of the local odds ratio structures, results are shown only for the independence, uniform and category exchangeability local odds ratio structures. The independence structure estimates no parameters, while the uniform represents the dependence structure by a single parameter. The category exchangeability assumes constant odds ratio among the levels of the response, but different odds ratio between time pairs. Thus, the association is explained by three parameters in this structure. Similar results were obtained for the other dependence structures; they are shown

in Appendix C.

Table 3.3: Results for Rheumatic Mitral Stenosis study under independence, uniform and category exchangeability association structures

	Independence			Uniform			Cat. Exchangeability		
Parameter	Est.	SE	p	Est.	SE	p	Est.	SE	p
				Available					
β_{01}	-0.800	0.765	0.295	-0.838	0.735	0.254	-0.768	0.751	0.307
β_{02}	0.997	0.745	0.181	0.954	0.726	0.189	1,037	0.746	0.164
Success	0.733	0.441	0.096	0.662	0.391	0.090	0.692	0.400	0.083
Total Score	-0.147	0.104	0.158	-0.142	0.101	0.161	-0.156	0.103	0.130
Cn	0.533	0.328	0.104	0.498	0.315	0.113	0.499	0.318	0.116
time=2	1,003	0.413	0.015	1,084	0.389	0.005	1,126	0.390	0.004
time=3	1,216	0.420	0.004	1,214	0.390	0.002	1,198	0.394	0.002
				WGEE					
β_{01}	-0.852	0.786	0.278	-0.883	0.763	0.247	-0.792	0.776	0.307
β_{02}	0.895	0.768	0.244	0.882	0.748	0.238	0.974	0.767	0.204
Success	0.750	0.449	0.094	0.732	0.405	0.071	0.749	0.416	0.072
Total Score	-0.139	0.107	0.192	-0.136	0.104	0.190	-0.153	0.106	0.148
Cn	0.524	0.336	0.118	0.464	0.325	0.153	0.472	0.327	0.148
time=2	1,030	0.413	0.013	1,039	0.401	0.010	1,085	0.406	0.007
time=3	1,269	0.419	0.002	1,352	0.404	0.001	1,331	0.410	0.001
				MIGEE					
β_{01}	-0.856	0.582	0.141	-0.805	0.580	0.165	-0.788	0.572	0.169
β_{02}	0.988	0.582	0.090	1,019	0.574	0.076	1,035	0.572	0.070
Success	0.938	0.367	0.010	0.922	0.333	0.006	0.980	0.338	0.004
Total Score	-0.141	0.080	0.077	-0.140	0.079	0.075	-0.147	0.078	0.060
Cn	0.425	0.305	0.163	0.307	0.261	0.239	0.343	0.267	0.199
time=2	0.995	0.334	0.003	1,106	0.308	0.000	1,127	0.330	0.001
time=3	0.969	0.338	0.004	1,068	0.304	0.000	1,079	0.333	0.001
				DRGEE					
β_{01}	-0.793	0.730	0.277	-0.830	0.711	0.243	-0.726	0.711	0.308
β_{02}	0.952	0.722	0.187	0.934	0.707	0.186	1.039	0.712	0.145
Sucess	0.825	0.422	0.050	0.806	0.382	0.035	0.841	0.388	0.030
Total Score	-0.150	0.099	0.131	-0.147	0.097	0.132	-0.165	0.098	0.092
Cn	0.571	0.341	0.094	0.520	0.329	0.113	0.523	0.333	0.116
time=2	1.013	0.499	0.042	1.005	0.512	0.050	1.020	0.525	0.052
time=3	1.135	0.476	0.017	1.185	0.473	0.012	1.145	0.491	0.020

The significance of the time effect indicates the effectiveness of the intervention and improvement of functional class over time. Similar coefficients for the two followup occasions suggests that the major change in functional class occurs right after the valvuloplasty intervention. The total score was not significant for all the four methods, although a marginal significance ($p = 0.060$) is noted for MIGEE when the category exchangeabil-

ity odds ratio structure is adopted, as an indication that higher scores may be related to cardiac insufficiency. All methods provide the same conclusion for effects of the missing covariate Cn. It can be noticed that estimates for success of the procedure effect goes from non significant in the standard GEE to significant for all methods, except WGEE, regardless of the adopted association structure. It is interesting to note that the estimated effect are increased for all methods, especially for the multiple imputation procedure. So, for example, for the uniform structure, the estimated odds of the response in class I for subjects with success in the procedure compared to patients with suboptimal results were $e^{0.806} = 2.24$ (95% CI: $1.528 - 3.281$) for DRGEE and $e^{0.922} = 2.51$ (95% CI: $1.802 - 3.508$) for MIGEE. Regarding the association structures, it can be noticed that the uniform and category exchangeability choices presented some gain in efficiency, specially for the covariate success of procedure effect. Similar conclusions were reached for the other association structures.

3.7 Discussion

In this paper, it was considered a doubly robust estimator for the analysis of longitudinal data when missingness can occur in a baseline covariate or intermittently in the ordinal response. The main objective was to compare the performance of the proposed method in terms of bias and efficiency for two different approaches to model the covariance matrix of the longitudinal outcome, namely, the correlation coefficient proposed by Lipsitz *et al.* (1994) and local odds ratio proposed by Touloumis *et al.* (2013). Although a complete data comparison between these two distinct approaches had already appeared elsewhere (Nooraee *et al.*, 2014) such a comparison with MAR missing data was still in need of investigation.

The covariate design plays an important role in the efficiency of GEE estimators. The working independence structure is expected to be efficient for time-stationary covariates. This is no longer true for time-varying covariates and/or missing data (Lipsitz *et al.*, 1994). The simulation results agreed with the literature, that is, the gain in efficiency by adopting more complicated association structures was noticed for the time-varying

covariate and it has varied through the different methods. For the complete data, the standard error for the independence was, on average, about 27% larger compared to an uniform structure that estimates a single association parameter. The gain in efficiency reached 23% for MIGEE, 19% for WGEE and 18% for DRGEE.

As Liang & Zeger (1986) pointed out, when the assumed correlation is the true one, the missing completely at random assumption can be unnecessary. However, this is not true when missingness occurs in a covariate that is MAR given the response. In this case, even likelihood methods are biased (Carpenter & Kenward, 2013) and the missingness mechanism must be modeled in order to obtain consistent estimates for the regression parameters.

The dependence structures compared here differ in terms of the number of parameters and restrictions imposed on the correlation/association between levels of ordinal response at different time pairs. Although the same definition can be used for an independence, exchangeable and unstructured association matrix, this does not imply that identical associations are fit (Nooraee *et al.*, 2014). Under the independence working assumption, all off-diagonal blocks of the covariance matrix are constant and equal to zero. Exchangeability over time indicates that the association between Y_{itj} and $Y_{it'j'}$ is independent of time, but it depends on the levels j and j'. For unstructured associations there are no restrictions implied. With moderate to large number of subjects (at least 300 subjects) the simulation results did not suggest a very clear distinction in terms of bias, relative efficiency and empirical coverage between the correlation or local odds ratio parametrizations, although some bias was observed for unstructured matrices. On the other side, for small sample sizes ($n = 50$ or 150) simulation results did suggest that the local odds ratio outperforms the correlation coefficient parametrization in terms of bias and convergence issues. It seems that local odds structures works fine for small samples sizes and the correlation coefficient needs relatively more subjects to achieve the same reduction in bias.

Two important differences are noted between the local odds ratio and the correlation coefficient parametrizations. First, the local odds ratios does not depend on the marginal

specification, that is, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are variation independent (Touloumis *et al.*, 2013). Unlike correlations, the local odds estimates for the association structure does not depend on the values of the observed covariates and thus these estimates do not need to be obtained at each step of the modified Fisher Scoring algorithm. As a consequence, the iterative procedure converges faster. Next, the correlation estimates are obtained through a moment based estimator, thus some work was necessary to ensure their validity in the presence of data that is MAR. The estimates of local odds are based on maximum likelihood methods, assuming independent Poisson sampling for the observed counts in the marginalized contingency tables. As it considers only the observed responses the resulting estimates are consistent under a MAR mechanism.

In some situations (small samples, complex patterns of missing data), it is possible that the algorithm will not converge because the estimated correlation matrix is not guaranteed to be positive definite (Lipsitz *et al.*, 1994). For small sample sizes there may not be enough data to estimate both the regression parameters and a correlation matrix that is highly unstructured. Simulation results for unstructured correlation matrices and small sample sizes presented very low convergence rates (in addition to increased bias) and thus an exchangeable correlation matrix is preferred. There were no serious convergence issues with the local odds ratio parametrization, except for an unstructured matrix applied to $n = 50$.

Although the latent vectors have been generated from an exchangeable correlation, the correlation coefficient between binary variables Y_{itj} and $Y_{it'j'}$ also depends on the linear predictor at times t and t' (Nooraee *et al.*, 2014). That is, the correlation between the latent vectors was exchangeable, but the correlation between the binary variables was not exchangeable due to the dependent mean in the marginal model. Nevertheless, it is expected that the GEE estimator remains consistent even when assuming a misspecified correlation matrix provided the marginal mean is correctly specified (Molenberghs & Kenward, 2010).

Overall, the doubly robust method performed well for all working association structures under comparison except when applied to a very small number of subjects (i.e.,

$n = 50$) and when an unstructured matrix is adopted. The coverage probabilities were relatively close to the nominal level of 95%.

The doubly robust estimator considered here is restricted to missingness in the response and a baseline covariate. A natural extension is to consider an intermittently missing time-varying covariate and/or allow multiple missing covariates. In the proposed approach, the marginal means were modeled by cumulative logits. This implies a proportional odds model that in some cases may not be valid. Another possible extension of the proposed model is, therefore, to allow non-proportional odds for a subset of the explanatory variables (Peterson & Harrell Jr, 1990).

BIBLIOGRAPHY

Agresti, Alan. 2013. *Categorical Data Analysis*. 3 edn. John Wiley & Sons.

Bang, Heejung, & Robins, James M. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61**(4), 962–973.

Becker, Mark P, & Clogg, Clifford C. 1989. Analysis of sets of two-way contingency tables using association models. *Journal of the American Statistical Association*, **84**(405), 142–151.

Beunckens, Caroline, Sotto, Cristina, & Molenberghs, Geert. 2008. A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational Statistics & Data Analysis*, **52**(3), 1533–1548.

Birhanu, Teshome, Molenberghs, Geert, Sotto, Cristina, & Kenward, Michael G. 2011. Doubly robust and multiple-imputation-based generalized estimating equations. *Journal of Biopharmaceutical Statistics*, **21**(2), 202–225.

Carey, Vincent, Zeger, Scott L, & Diggle, Peter. 1993. Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, **80**(3), 517–526.

Carpenter, James R., & Kenward, Michael G. 2013. *Multiple Imputation and its Application*. Wiley & Sons.

Carpenter, James R, Kenward, Michael G, & Vansteelandt, Stijn. 2006. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **169**(3), 571–584.

Chen, Baojiang, & Zhou, Xiao-Hua. 2011. Doubly robust estimates for binary longitudinal data analysis with missing response and missing covariates. *Biometrics*, **67**(3), 830–842.

Deming, W Edwards, & Stephan, Frederick F. 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, **11**(4), 427–444.

Donneau, Anne-Françoise, Mauer, Murielle, Molenberghs, Geert, & Albert, Adelin. 2015a. A simulation study comparing multiple imputation methods for incomplete longitudinal ordinal data. *Communications in Statistics-Simulation and Computation*, **44**(5), 1311–1338.

Donneau, Anne-Françoise, Mauer, Murielle, Lambert, Philippe, Molenberghs, Geert, & Albert, Adelin. 2015b. Simulation-based study comparing multiple imputation methods for non-monotone missing ordinal data in longitudinal settings. *Journal of Biopharmaceutical Statistics*, **25**(3), 570–601.

Fitzmaurice, G., Davidian, M., Molenberghs, G., & Verbeke, G. 2009. *Longitudinal Data Analysis*. Handbooks of Modern Statistical Methods. Chapman & Hall/CRC.

Fitzmaurice, Garrett M., Laird, M., & Ware, James H. 2004. *Applied Longitudinal Analysis*. Wiley-Interscience.

Goodman, Leo A. 1985. The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *The Annals of Statistics*, **13**(1), 10–69.

Heagerty, Patrick J, & Zeger, Scott L. 1996. Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association*, **91**(435), 1024–1036.

Li, Lingling, Shen, Changyu, Li, Xiaochun, & Robins, James M. 2013. On weighting approaches for missing data. *Statistical Methods in Medical Research*, **22**(1), 14–30.

Liang, Kung-Yee, & Zeger, Scott L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22.

Lipsitz, Stuart R, Laird, Nan M, & Harrington, David P. 1991. Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, **78**(1), 153–160.

Lipsitz, Stuart R, Kim, Kyungmann, & Zhao, Lueping. 1994. Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine*, **13**(11), 1149–1163.

Little, Roderick JA, & Rubin, Donald B. 1987. *Statistical Analysis with Missing Data*. 1 edn. Wiley New York.

Little, Roderick JA, & Rubin, Donald B. 2002. *Statistical Analysis with Missing Data*. 2 edn. Wiley New York.

Lumley, Thomas. 1996. Generalized estimating equations for ordinal data: a note on working correlation structures. *Biometrics*, **52**(1), 354–361.

McCullagh, Peter. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, **42**(2), 109–142.

Molenberghs, Geert, & Kenward, Michael G. 2010. Semi-parametric marginal models for hierarchical data and their corresponding full models. *Computational Statistics & Data Analysis*, **54**(2), 585–597.

Molenberghs, Geert, & Verbeke, Geert. 2005. *Models for discrete longitudinal data*. 1 edn. Springer Series in Statistics. Springer-Verlag New York.

Nooraee, Nazanin, Molenberghs, Geert, & van den Heuvel, Edwin R. 2014. GEE for longitudinal ordinal data: Comparing R-geepack, R-multgee, R-repolr, SAS-GENMOD, SPSS-GENLIN. *Computational Statistics & Data Analysis*, **77**, 70–83.

Parsons, Nicholas R, Edmondson, RN, & Gilmour, SG. 2006. A generalized estimating equation method for fitting autocorrelated ordinal score data with an application in horticultural research. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **55**(4), 507–524.

Parsons, Nick R, Costa, Matthew L, Achten, Juul, & Stallard, Nigel. 2009. Repeated measures proportional odds logistic regression analysis of ordinal score data in the statistical software package R. *Computational Statistics & Data Analysis*, **53**(3), 632–641.

Peterson, Bercedis, & Harrell Jr, Frank E. 1990. Partial proportional odds models for ordinal response variables. *Applied Statistics*, **39**(2), 205–217.

Pierce, Donald A. 1982. The asymptotic effect of substituting estimators for parameters in certain types of statistics. *The Annals of Statistics*, **10**(2), 475–478.

Poleto, Frederico Z, Singer, Julio M, & Paulino, Carlos Daniel. 2014. A product-multinomial framework for categorical data analysis with missing responses. *Brazilian Journal of Probability and Statistics*, **28**(1), 109–139.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robins, James M, Rotnitzky, Andrea, & Zhao, Lue Ping. 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**(429), 106–121.

Rubin, Donald B. 1976. Inference and missing data. *Biometrika*, **63**(3), 581–592.

Rubin, Donald B. 1978. Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. *Pages 20–34 of: Proceedings of the survey research methods section of the American Statistical Association*, vol. 1. American Statistical Association.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. John Wiley New York.

Schafer, Joseph L. 1997. *Analysis of incomplete multivariate data*. Monographs on Statistics & Applied Probability. Chapman and Hall/CRC.

Schafer, Joseph L. 1999. Multiple imputation: a primer. *Statistical Methods in Medical Research*, **8**(1), 3–15.

Scharfstein, Daniel O, Rotnitzky, Andrea, & Robins, James M. 1999. Adjusting for non-ignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, **94**(448), 1096–1120.

Seaman, Shaun, & Copas, Andrew. 2009. Doubly robust generalized estimating equations for longitudinal data. *Statistics in Medicine*, **28**(6), 937–955.

Toledano, Alicia Y, & Gatsonis, Constantine. 1999. Generalized estimating equations for ordinal categorical data: arbitrary patterns of missing responses and missingness in a key covariate. *Biometrics*, **55**(2), 488–496.

Touloumis, Anestis. 2015. *SimCorMultRes: Simulates Correlated Multinomial Responses*. R package version 1.3.0.

Touloumis, Anestis, Agresti, Alan, & Kateri, Maria. 2013. GEE for multinomial responses using a local odds ratios parameterization. *Biometrics*, **69**(3), 633–640.

Tsiatis, Anastasios A. 2006. *Semiparametric theory and missing data*. Springer Series in Statistics. Springer New York.

van Buuren, S., Boshuizen, H. C., & Knook, D. L. 1999. Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. *Statistics in Medicine*, **18**(6), 681–694.

van Buuren, Stef. 2007. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, **16**(3), 219–242.

van Buuren, Stef, & Groothuis-Oudshoorn, Karin. 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, **45**(3), 1–67.

Vansteelandt, Stijn, Carpenter, James, & Kenward, Michael G. 2010. Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, **6**(1), 37–48.

CAPÍTULO 4

CONCLUSÕES GERAIS

A presente tese teve por objetivo o tratamento de dados ausentes em estudos longitudinais com respostas ordinais. Foi abordado o caso geral em que a perda na resposta ocorre de forma intermitente. Tal padrão arbitrário de perda engloba a situação comumente estudada de perda do tipo abandono. Um aspecto bastante relevante, e muitas vezes desconsiderado, é quando a não resposta ocorre em alguma covariável de interesse. Neste caso, valiosa informação pode ser recuperada na análise.

Quando ocorrem dados ausentes, é fundamental distinguir os mecanismos que geram a não resposta a fim de se obterem estimadores consistentes dos parâmetros de regressão. Quando o mecanismo gerador das perdas é do tipo MAR, propomos um estimador GEE duplamente robusto para análise de dados ordinais correlacionados com perda na resposta e na covariável. Os resultados de simulação sugeriram que tal método é uma boa alternativa à imputação múltipla e ao método GEE ponderado, mas com a vantagem de requerer a correta especificação de apenas um de seus modelos preditos a fim de retornar estimativas consistentes dos parâmetros de interesse. De forma bastante geral, os resultados de simulação sugerem que o estimador é consistente e apresenta viés de pequenas amostras comparável (e em vários casos menor) aos concorrentes citados, embora perca em eficiência para a imputação múltipla. A eficiência do estimador ponderado é afetada pela variabilidade dos pesos estimados enquanto a eficiência da imputação múltipla depende da incerteza do valor imputado. Como o estimador duplamente robusto combina esses dois métodos, é de se esperar que a variabilidade estimada dos coeficientes fique num meio termo. É natural que a redução do viés introduzido pelos dados ausentes dependa da capacidade preditiva das variáveis nos modelos auxiliares, seja para explicar a probabilidade de não resposta ou para predizer o real valor não observado. Em algumas situações, um ou outro modelo pode ser mais facilmente especificado e, por isso, o estimador proposto

ganha em flexibilidade de seus concorrentes.

Há certo debate na literatura sobre se vale a pena a modelagem da estrutura de dependência entre as medidas repetidas, já que o estimador GEE é consistente para o caso de dados completos, independentemente de como a matriz de covariâncias seja especificada. Como já apontado na literatura, a resposta a tal questão depende das características das covariáveis preditoras no modelo marginal. A modelagem da associação por meio do coeficiente de correlação e também das razões de chances locais indicaram, via resultados de simulação, que na presença de covariáveis que mudam de valor no tempo, pode-se ganhar consideravelmente eficiência mesmo assumindo uma estrutura de associação uniforme (a qual descreve toda a associação entre as medidas repetidas por meio de um único parâmetro). Dados ausentes levam naturalmente à perda de eficiência, porque as inferências são baseadas em uma amostra menor do que aquela planejada, e por isso parece que parametrizar a matriz de covariância traz vantagens. Um ponto bastante relevante é que a modelagem da associação pode reduzir sobremaneira o viés das estimativas, mesmo que a estrutura de correlação adotada não corresponda de fato à verdadeira correlação entre as medidas repetidas.

Entre as duas abordagens consideradas, notou-se diferença em termos de eficiência apenas para amostras pequenas, com vantagem para a parametrização de razões de chances locais. Tal abordagem apresentou ainda menos erros de convergência comparado com o coeficiente de correlação, além de ter a vantagem de ser estimada por métodos de verossimilhança usando apenas as respostas observadas, o que garante sua consistência no caso de perda MAR. Além disso, como as estimativas das razões de chances locais não são atualizadas a cada iteração, o processo de estimação se torna mais rápido.

O modelo adotado para a resposta marginal foi o chamado modelo de razões de chances proporcionais, o qual supõe que o efeito do preditor nas razões de chances não dependa do nível da resposta. Uma proposta de trabalho futuro é desenvolver um teste para verificar tal suposição e/ou adotar modelos mais flexíveis para a resposta ordinal. Outra possibilidade é permitir a incorporação de múltiplas covariáveis ausentes.

CAPÍTULO 5

APÊNDICES

APÊNDICE A

VARIÂNCIA ASSINTÓTICA

To state the asymptotic properties of $\hat{\boldsymbol{\beta}}$, let

$\boldsymbol{S}_{1i}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma})$ be the individual's contribution to the estimating equations for $\boldsymbol{\beta}$,

$\boldsymbol{S}_{2i}(\boldsymbol{\psi})$ be the individual's contribution to the estimating equations for $\boldsymbol{\psi}$, and

$\boldsymbol{S}_{3i}(\boldsymbol{\gamma})$ be the individual's contribution to the estimating equations for $\boldsymbol{\gamma}$.

Define $\boldsymbol{\Gamma}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma}) = E\left\{\partial \boldsymbol{S}_{1i}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma})/\partial \boldsymbol{\beta}^T\right\}$, $\boldsymbol{I}_{12}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma}) = E\left\{\partial \boldsymbol{S}_{1i}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma})/\partial \boldsymbol{\psi}^T\right\}$, $\boldsymbol{I}_{13}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma}) = E\left\{\partial \boldsymbol{S}_{1i}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma})/\partial \boldsymbol{\gamma}^T\right\}$, $\boldsymbol{I}_2(\boldsymbol{\psi}) = E\left\{\partial \boldsymbol{S}_{2i}(\boldsymbol{\psi})/\partial \boldsymbol{\psi}^T\right\}$, $\boldsymbol{I}_3(\boldsymbol{\gamma}) = E\left\{\partial \boldsymbol{S}_{3i}(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}^T\right\}$, and $\boldsymbol{Q}_i(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma}) = \boldsymbol{S}_{1i}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma}) - \boldsymbol{I}_{12}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma})\boldsymbol{I}_2^{-1}(\boldsymbol{\psi})\boldsymbol{S}_{2i}(\boldsymbol{\psi}) - \boldsymbol{I}_{13}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma})\boldsymbol{I}_3^{-1}(\boldsymbol{\gamma})\boldsymbol{S}_{3i}(\boldsymbol{\gamma})$.

Theorem 1 *If either the missing data model or the covariate model is correctly specified, then*

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \longrightarrow N(\mathbf{0}, \boldsymbol{\Gamma}^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\psi}_0, \boldsymbol{\gamma}_0)\boldsymbol{\Sigma}\left\{\boldsymbol{\Gamma}^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\psi}_0, \boldsymbol{\gamma}_0)\right\}^T), \tag{A.1}$$

where $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$, $\boldsymbol{\psi}_0$ and $\boldsymbol{\gamma}_0$ are the probability limits of $\hat{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\gamma}}$, and $\boldsymbol{\Sigma} = E\left\{\boldsymbol{Q}_i(\boldsymbol{\beta}_0, \boldsymbol{\psi}_0, \boldsymbol{\gamma}_0)\boldsymbol{Q}_i^T(\boldsymbol{\beta}_0, \boldsymbol{\psi}_0, \boldsymbol{\gamma}_0)\right\}$.

Inferences for $\boldsymbol{\beta}$ follows by replacing the unknown quantities in (A.1) by its consistent estimators. We make use of "generalized information equality" (Pierce, 1982) that

$E\left\{\partial \boldsymbol{S}_{1i}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma})/\partial \boldsymbol{\psi}^T\right\} = -E\left\{\boldsymbol{S}_{1i}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma})\boldsymbol{S}_{2i}^T(\boldsymbol{\psi})\right\}$, and

$E\left\{\partial \boldsymbol{S}_{1i}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma})/\partial \boldsymbol{\gamma}^T\right\} = -E\left\{\boldsymbol{S}_{1i}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma})\boldsymbol{S}_{3i}^T(\boldsymbol{\gamma})\right\}$. Similarly (Robins *et al.*, 1995),

$E\left\{\partial \boldsymbol{S}_{2i}(\boldsymbol{\psi})/\partial \boldsymbol{\psi}^T\right\} = -Var\left\{\boldsymbol{S}_{2i}(\boldsymbol{\psi})\right\}$, and $E\left\{\partial \boldsymbol{S}_{3i}(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}^T\right\} = -Var\left\{\boldsymbol{S}_{3i}(\boldsymbol{\gamma})\right\}$.

The matrix $\boldsymbol{\Gamma}$ is replaced by $\hat{\boldsymbol{\Gamma}} = n^{-1}\sum_{i=1}^n \left\{\partial \boldsymbol{S}_{1i}(\hat{\boldsymbol{\theta}})/\partial \boldsymbol{\beta}^T\right\}$, and $\boldsymbol{\Sigma}$ by $\hat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^n \left\{\hat{\boldsymbol{Q}}_i\hat{\boldsymbol{Q}}_i^T\right\}$, $\hat{\boldsymbol{Q}}_i = \boldsymbol{S}_{1i}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{I}}_{12}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{I}}_2^{-1}(\hat{\boldsymbol{\psi}})\boldsymbol{S}_{2i}(\hat{\boldsymbol{\psi}}) - \hat{\boldsymbol{I}}_{13}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{I}}_3^{-1}(\hat{\boldsymbol{\gamma}})\boldsymbol{S}_{3i}(\hat{\boldsymbol{\gamma}})$, $\hat{\boldsymbol{I}}_{12}(\hat{\boldsymbol{\theta}}) = n^{-1}\sum_{i=1}^n \left\{\partial \boldsymbol{S}_{1i}(\hat{\boldsymbol{\theta}})/\partial \boldsymbol{\psi}^T\right\}$, $\hat{\boldsymbol{I}}_{13}(\hat{\boldsymbol{\theta}}) = n^{-1}\sum_{i=1}^n \left\{\partial \boldsymbol{S}_{1i}(\hat{\boldsymbol{\theta}})/\partial \boldsymbol{\gamma}^T\right\}$, $\hat{\boldsymbol{I}}_2(\hat{\boldsymbol{\psi}}) = n^{-1}\sum_{i=1}^n \left\{\partial \boldsymbol{S}_{2i}(\hat{\boldsymbol{\psi}})/\partial \boldsymbol{\psi}^T\right\}$, $\hat{\boldsymbol{I}}_3(\hat{\boldsymbol{\gamma}}) = n^{-1}\sum_{i=1}^n \left\{\partial \boldsymbol{S}_{3i}(\hat{\boldsymbol{\gamma}})/\partial \boldsymbol{\gamma}^T\right\}$.

The proof is similar to Chen & Zhou (2011) and is omitted here.

APÊNDICE B

RESULTADOS ADICIONAIS PARA AS SIMULAÇÕES NO ARTIGO 2

Tabela B.1: Evaluation criteria for misspecified models. Results for $n = 50$ and $S = 1000$ simulations.

	Relative Bias				Relative Efficiency				Empirical Coverage			
Structure	β_{01}	β_{02}	X	Z	β_{01}	β_{02}	X	Z	β_{01}	β_{02}	X	Z
						Available						
ind	101.9	-25.2	-10.0	48.4	1.00	1.00	1.00	1.00	0.85	0.85	1.00	0.95
exch	96.2	-22.6	-36.0	19.5	1.02	1.03	1.03	1.15	0.87	0.88	1.00	0.95
unst	91.0	-21.8	-44.1	-1.3	1.08	1.10	1.10	1.17	0.86	0.90	1.00	0.96
unif	80.2	-18.6	-14.8	26.5	1.01	0.99	1.00	1.19	0.91	0.90	1.00	0.94
cat.exch	80.9	-19.0	-16.8	19.5	1.01	1.00	1.01	1.23	0.90	0.89	1.00	0.94
time.exch	82.8	-18.2	-16.9	25.1	1.00	0.99	1.01	1.20	0.88	0.90	1.00	0.93
RC	83.6	-20.0	-19.2	9.1	1.01	1.00	1.01	1.26	0.95	0.93	1.00	0.93
						WGEE(r^-)						
ind	24.5	-0.7	-13.5	30.3	1.00	1.00	1.00	1.00	0.94	0.93	1.00	0.96
exch	36.3	-5.2	-29.7	14.2	1.01	1.02	1.03	1.14	0.94	0.94	1.00	0.95
unst	43.3	-7.0	-44.7	-3.0	1.07	1.11	1.09	1.17	0.93	0.93	1.00	0.96
unif	17.9	0.6	-12.6	24.8	1.00	0.99	0.99	1.14	0.96	0.94	1.00	0.96
cat.exch	19.2	0.0	-11.6	19.0	1.00	0.99	0.99	1.15	0.95	0.93	1.00	0.95
time.exch	18.1	0.6	-12.6	23.6	0.98	0.98	0.99	1.15	0.96	0.95	1.00	0.94
RC	22.3	-1.4	-12.7	9.9	0.99	0.99	0.99	1.17	0.97	0.95	1.00	0.95
						MIGEE(x^-)						
ind	25.6	-5.0	-34.7	-2.2	1.00	1.00	1.00	1.00	0.96	0.95	1.00	0.96
exch	35.4	-7.8	-53.5	-8.3	1.00	1.01	1.00	1.18	0.95	0.97	1.00	0.94
unst	35.2	-6.7	-59.9	-28.8	1.03	1.06	1.05	1.16	0.96	0.97	1.00	0.95
unif	23.0	-5.1	-36.4	2.4	1.00	0.99	1.00	1.19	0.97	0.96	1.00	0.94
cat.exch	24.8	-5.8	-35.3	-3.5	0.99	0.99	1.00	1.24	0.96	0.96	1.00	0.95
time.exch	22.9	-4.0	-36.0	-1.2	0.98	0.99	0.99	1.21	0.97	0.96	1.00	0.92
RC	25.5	-5.5	-35.8	-16.3	1.00	0.99	1.00	1.28	0.95	0.94	1.00	0.93
						DRGEE(x^-, r^-)						
ind	21.1	-1.4	-19.0	-4.0	1.00	1.00	1.00	1.00	0.94	0.93	1.00	0.95
exch	33.8	-6.0	-35.9	-14.4	1.02	1.03	1.04	1.18	0.93	0.94	1.00	0.92
unst	38.2	-7.8	-55.3	-34.2	1.08	1.13	1.13	1.20	0.93	0.93	1.00	0.93
unif	15.4	-0.1	-18.2	-1.0	1.00	0.98	0.99	1.15	0.96	0.94	1.00	0.94
cat.exch	16.3	-0.7	-18.9	-4.8	1.00	0.98	0.99	1.17	0.95	0.93	1.00	0.94
time.exch	15.5	-0.1	-17.1	-2.5	0.99	0.98	1.00	1.17	0.96	0.94	1.00	0.92
RC	20.1	-1.7	-20.1	-15.2	0.99	0.98	0.99	1.21	0.95	0.94	1.00	0.91

"$^+$"indicates correctly specified model and "$^-$"indicates misspecified model omitting the Z_1 predictor

Tabela B.2: Evaluation criteria for correctly specified models. Results for $n = 50$ and $S = 1000$ simulations.

	Relative Bias				Relative Efficiency				Empirical Coverage			
Structure	β_{01}	β_{02}	X	Z	β_{01}	β_{02}	X	Z	β_{01}	β_{02}	X	Z
						Complete						
ind	3.4	4.4	3.3	3.5	1.00	1.00	1.00	1.00	0.95	0.94	0.99	0.96
exch	13.6	1.3	-12.4	-8.0	0.99	1.01	1.01	1.23	0.94	0.96	1.00	0.94
unst	17.0	2.3	-25.5	-18.9	1.03	1.06	1.05	1.21	0.95	0.96	0.98	0.94
unif	0.1	4.6	4.0	2.8	1.00	1.00	1.00	1.25	0.95	0.93	1.00	0.95
cat.exch	0.9	4.2	3.6	-1.8	1.00	1.00	1.00	1.28	0.95	0.93	0.99	0.95
time.exch	3.3	4.8	-2.6	1.6	0.99	0.99	1.00	1.26	0.95	0.95	0.99	0.93
RC	5.1	3.0	-0.6	-8.1	1.00	1.00	1.00	1.31	0.95	0.93	1.00	0.93
						WGEE(r^+)						
ind	14.8	3.2	1.8	9.1	1.00	1.00	1.00	1.00	0.95	0.94	1.00	0.97
exch	27.7	-1.7	-16.3	-1.6	1.01	1.03	1.03	1.16	0.94	0.94	1.00	0.95
unst	37.5	-4.7	-29.3	-13.5	1.08	1.12	1.10	1.18	0.93	0.93	1.00	0.97
unif	8.1	5.1	4.7	10.1	1.00	0.98	0.99	1.15	0.96	0.93	1.00	0.95
cat.exch	10.0	4.4	4.4	4.9	1.00	0.98	1.00	1.16	0.96	0.94	1.00	0.95
time.exch	6.9	5.7	8.6	5.7	0.97	0.97	0.99	1.16	0.96	0.95	1.00	0.95
RC	12.4	2.9	0.8	-5.9	0.98	0.98	0.99	1.19	0.96	0.94	1.00	0.94
						MIGEE(x^+)						
ind	12.1	0.6	-5.1	6.5	1.00	1.00	1.00	1.00	0.96	0.95	1.00	0.96
exch	23.7	-2.9	-20.3	-7.3	0.99	1.01	1.01	1.18	0.95	0.96	1.00	0.94
unst	26.8	-3.3	-32.4	-27.0	1.02	1.06	1.04	1.15	0.95	0.97	0.99	0.95
unif	7.5	0.9	0.0	5.1	1.00	1.00	1.00	1.19	0.96	0.94	1.00	0.94
cat.exch	9.1	0.3	0.5	-0.6	1.00	1.00	1.00	1.23	0.95	0.94	1.00	0.94
time.exch	7.8	1.6	-2.1	2.1	0.99	0.99	1.00	1.21	0.96	0.95	1.00	0.93
RC	13.4	-0.1	-4.9	-14.0	1.00	0.99	1.00	1.28	0.95	0.94	1.00	0.92
						DRGEE(x^+, r^+)						
ind	7.0	4.1	0.1	2.4	1.00	1.00	1.00	1.00	0.94	0.94	1.00	0.96
exch	19.3	-0.5	-13.9	-10.0	1.01	1.03	1.04	1.19	0.94	0.95	1.00	0.93
unst	25.0	-2.5	-28.3	-23.3	1.08	1.13	1.13	1.21	0.92	0.94	1.00	0.94
unif	3.9	5.5	3.7	4.5	0.99	0.98	0.99	1.16	0.95	0.93	1.00	0.94
cat.exch	4.4	4.8	3.9	-0.1	0.99	0.98	0.99	1.18	0.94	0.93	1.00	0.94
time.exch	2.7	5.7	10.5	0.8	0.97	0.97	0.99	1.16	0.96	0.95	1.00	0.93
RC	9.0	3.2	1.0	-10.7	0.97	0.97	0.98	1.20	0.95	0.94	1.00	0.91
						DRGEE(x^-, r^+)						
ind	11.6	3.3	-2.4	2.3	1.00	1.00	1.00	1.00	0.94	0.94	1.00	0.96
exch	25.8	-2.0	-22.2	-9.0	1.02	1.06	1.06	1.22	0.93	0.95	1.00	0.93
unst	31.8	-5.0	-39.5	-26.1	1.09	1.18	1.18	1.26	0.92	0.93	1.00	0.93
unif	5.4	5.1	-1.3	3.2	1.00	0.98	1.01	1.17	0.95	0.94	1.00	0.95
cat.exch	7.3	4.5	-2.7	-1.3	1.00	0.98	1.01	1.20	0.95	0.94	1.00	0.94
time.exch	5.2	5.2	1.6	1.1	0.98	0.97	0.99	1.18	0.96	0.95	1.00	0.94
RC	11.2	3.1	-4.7	-12.5	0.98	0.97	1.00	1.22	0.94	0.94	1.00	0.92
						DRGEE(x^+, r^-)						
ind	8.9	3.9	-2.0	2.0	1.00	1.00	1.00	1.00	0.93	0.93	1.00	0.96
exch	18.3	-0.5	-14.8	-12.9	1.01	1.03	1.03	1.18	0.94	0.95	1.00	0.92
unst	23.7	-2.3	-28.5	-30.3	1.06	1.11	1.11	1.20	0.92	0.94	1.00	0.93
unif	4.7	4.9	3.7	2.6	0.99	0.98	0.99	1.15	0.95	0.93	1.00	0.95
cat.exch	5.7	4.2	2.4	-2.4	0.99	0.98	0.99	1.17	0.95	0.92	1.00	0.94
time.exch	3.3	5.3	4.4	1.3	0.98	0.98	1.00	1.16	0.96	0.94	1.00	0.92
RC	10.5	3.4	-2.2	-11.9	0.99	0.98	0.99	1.19	0.88	0.89	1.00	0.92

"+"indicates correctly specified model and "−"indicates misspecified model omitting the Z_1 predictor

Tabela B.3: Evaluation criteria for misspecified models. Results for $n = 150$ and $S = 1000$ simulations.

	Relative Bias				Relative Efficiency				Empirical Coverage			
Structure	β_{01}	β_{02}	X	Z	β_{01}	β_{02}	X	Z	β_{01}	β_{02}	X	Z
						Available						
ind	117.0	-29.9	-20.9	60.5	1.00	1.00	1.00	1.00	0.46	0.66	0.94	0.84
exch	96.4	-25.5	-32.2	30.2	1.02	1.01	1.02	1.20	0.58	0.7	0.92	0.92
unst	97.4	-26.3	-40.0	17.4	1.04	1.03	1.04	1.21	0.59	0.71	0.94	0.94
unif	92.5	-22.6	-25.9	26.8	1.02	1.00	1.01	1.23	0.63	0.78	0.93	0.91
cat.exch	92.6	-22.7	-26.4	24.8	1.02	1.00	1.01	1.24	0.64	0.77	0.93	0.92
time.exch	93.6	-22.8	-25.6	30.0	1.02	1.00	1.01	1.23	0.61	0.76	0.94	0.91
RC	93.2	-22.4	-25.1	24.1	1.02	1.00	1.01	1.25	0.62	0.78	0.94	0.92
						WGEE(r^-)						
ind	19.0	-3.4	-22.9	37.5	1.00	1.00	1.00	1.00	0.94	0.94	0.94	0.9
exch	21.5	-5.5	-32.9	22.2	1.00	1.00	1.00	1.16	0.94	0.93	0.94	0.94
unst	27.3	-7.4	-44.2	10.0	1.02	1.02	1.02	1.17	0.94	0.94	0.95	0.94
unif	18.7	-3.4	-24.0	24.7	1.00	1.00	1.00	1.16	0.94	0.94	0.94	0.93
cat.exch	19.1	-3.6	-24.1	23.0	1.00	1.00	1.00	1.17	0.94	0.94	0.94	0.93
time.exch	19.0	-4.0	-25.3	26.4	1.00	1.00	1.00	1.16	0.94	0.94	0.94	0.92
RC	20.3	-4.0	-24.8	20.9	1.00	1.00	1.00	1.16	0.94	0.96	0.94	0.94
						MIGEE(x^-)						
ind	21.2	-5.9	-46.4	-7.4	1.00	1.00	1.00	1.00	0.94	0.95	0.94	0.95
exch	22.7	-7.5	-53.4	-2.2	0.98	0.99	0.99	1.21	0.95	0.95	0.95	0.96
unst	24.9	-8.1	-57.2	-10.0	0.99	1.00	1.00	1.21	0.95	0.96	0.96	0.94
unif	20.8	-5.9	-45.9	-1.8	1.00	1.00	1.00	1.22	0.94	0.95	0.94	0.93
cat.exch	21.3	-6.1	-46.4	-3.7	1.00	1.00	1.00	1.24	0.94	0.95	0.94	0.93
time.exch	20.1	-6.2	-44.8	0.4	1.00	1.00	1.00	1.22	0.95	0.96	0.96	0.94
RC	20.8	-6.1	-45.3	-3.8	1.00	1.00	1.00	1.25	0.95	0.96	0.96	0.94
						DRGEE(x^-, r^-)						
ind	16.6	-3.5	-32.9	-5.4	1.00	1.00	1.00	1.00	0.93	0.94	0.94	0.94
exch	19.4	-5.5	-42.8	-4.9	1.00	1.00	1.01	1.13	0.93	0.93	0.93	0.95
unst	24.7	-7.4	-54.6	-16.1	1.01	1.02	1.03	1.16	0.93	0.93	0.93	0.93
unif	16.0	-3.3	-31.9	-1.7	1.00	1.00	0.99	1.14	0.93	0.94	0.95	0.92
cat.exch	16.4	-3.5	-32.2	-2.9	1.00	1.00	0.99	1.16	0.94	0.94	0.94	0.92
time.exch	16.0	-3.8	-34.2	0.6	1.00	1.00	1.00	1.16	0.93	0.94	0.93	0.95
RC	17.0	-4.0	-35.3	-5.7	1.00	1.00	1.00	1.17	0.94	0.95	0.94	0.94

"+"indicates correctly specified model and "−"indicates misspecified model omitting the Z_1 predictor

Tabela B.4: Evaluation criteria for correctly specified models. Results for $n = 150$ and $S = 1000$ simulations.

	Relative Bias				Relative Efficiency				Empirical Coverage			
Structure	β_{01}	β_{02}	X	Z	β_{01}	β_{02}	X	Z	β_{01}	β_{02}	X	Z
						Complete						
ind	0.2	1.9	2.9	1.0	1.00	1.00	1.00	1.00	0.94	0.94	0.95	0.95
exch	0.9	0.8	-3.8	0.7	0.99	1.00	1.00	1.26	0.93	0.93	0.95	0.97
unst	3.4	0.4	-8.5	-6.1	1.01	1.01	1.01	1.26	0.95	0.95	0.97	0.95
unif	-0.1	2.0	3.2	1.6	1.00	1.00	1.00	1.26	0.94	0.94	0.94	0.95
cat.exch	0.0	1.9	2.8	0.3	1.00	1.00	1.00	1.27	0.94	0.94	0.94	0.94
time.exch	0.3	1.6	3.3	2.9	1.00	1.00	1.00	1.27	0.95	0.96	0.95	0.94
RC	0.7	1.8	1.3	-0.1	1.01	1.01	1.00	1.28	0.95	0.95	0.95	0.96
						WGEE(r^+)						
ind	3.3	1.5	3.9	6.0	1.00	1.00	1.00	1.00	0.96	0.95	0.96	0.95
exch	5.7	-0.5	-7.1	1.2	0.99	0.99	1.00	1.16	0.94	0.94	0.95	0.96
unst	13.3	-2.8	-18.2	-7.9	1.02	1.02	1.02	1.18	0.96	0.97	0.97	0.95
unif	2.6	1.6	4.0	3.7	0.99	0.99	0.99	1.17	0.95	0.95	0.96	0.94
cat.exch	3.0	1.4	3.9	2.5	0.99	0.99	0.99	1.18	0.95	0.96	0.96	0.94
time.exch	2.1	1.3	2.1	4.4	0.98	0.98	0.99	1.17	0.95	0.95	0.95	0.94
RC	4.4	0.9	1.5	-0.3	1.00	0.99	0.99	1.17	0.95	0.96	0.96	0.95
						MIGEE(x^+)						
ind	5.4	0.1	-3.3	0.9	1.00	1.00	1.00	1.00	0.94	0.95	0.95	0.95
exch	7.7	-1.8	-14.5	0.3	0.99	0.99	0.99	1.20	0.94	0.94	0.96	0.96
unst	9.6	-2.6	-18.9	-7.5	1.00	1.00	1.00	1.20	0.96	0.96	0.96	0.95
unif	5.0	0.1	-3.0	1.4	1.00	1.00	1.00	1.21	0.94	0.95	0.95	0.93
cat.exch	5.6	-0.1	-3.8	-0.6	1.00	1.00	1.00	1.23	0.94	0.95	0.95	0.93
time.exch	4.7	-0.3	-4.1	3.3	1.00	1.00	1.00	1.22	0.95	0.95	0.95	0.94
RC	5.3	0.0	-3.7	-0.8	1.00	1.00	1.00	1.24	0.95	0.95	0.96	0.94
						DRGEE(x^+, r^+)						
ind	0.3	2.5	5.0	1.9	1.00	1.00	1.00	1.00	0.95	0.94	0.95	0.94
exch	3.0	0.3	-5.7	-0.1	0.98	0.99	1.00	1.14	0.93	0.94	0.95	0.94
unst	8.9	-1.7	-16.5	-9.6	1.00	1.01	1.03	1.16	0.94	0.95	0.97	0.93
unif	-0.2	2.5	5.5	1.9	1.00	1.00	1.00	1.15	0.95	0.94	0.95	0.93
cat.exch	0.2	2.3	5.1	0.7	1.00	1.00	1.00	1.16	0.94	0.94	0.95	0.92
time.exch	-0.6	2.2	3.6	4.3	0.98	0.98	0.99	1.15	0.94	0.95	0.95	0.94
RC	1.7	1.8	2.0	-1.5	0.99	1.00	0.99	1.17	0.95	0.95	0.96	0.94
						DRGEE(x^-, r^+)						
ind	1.8	1.9	0.7	1.9	1.00	1.00	1.00	1.00	0.95	0.95	0.97	0.94
exch	4.6	-0.2	-10.2	-0.7	0.99	1.00	1.01	1.17	0.94	0.94	0.96	0.95
unst	11.4	-2.5	-22.0	-10.3	1.02	1.04	1.05	1.19	0.95	0.96	0.96	0.93
unif	1.3	2.1	0.8	1.8	1.00	0.99	1.00	1.18	0.95	0.96	0.97	0.94
cat.exch	1.7	1.9	0.2	0.6	1.00	0.99	1.01	1.19	0.95	0.96	0.97	0.93
time.exch	1.0	1.8	0.3	4.4	0.99	0.99	0.99	1.18	0.94	0.95	0.97	0.94
RC	3.1	1.4	-1.6	-1.8	1.00	1.00	1.00	1.20	0.95	0.95	0.97	0.94
						DRGEE(x^+, r^-)						
ind	1.2	2.1	3.0	1.7	1.00	1.00	1.00	1.00	0.95	0.94	0.95	0.94
exch	3.0	0.3	-6.6	-1.6	1.00	1.00	1.01	1.13	0.93	0.94	0.94	0.95
unst	8.1	-1.6	-16.0	-13.3	1.00	1.01	1.03	1.16	0.94	0.95	0.97	0.93
unif	0.7	2.2	3.4	1.6	1.00	1.00	1.00	1.14	0.94	0.94	0.95	0.93
cat.exch	1.1	2.0	2.6	0.4	1.00	1.00	1.00	1.15	0.94	0.94	0.94	0.93
time.exch	0.3	1.8	0.6	3.7	1.00	1.00	1.00	1.15	0.93	0.94	0.95	0.95
RC	1.5	1.6	2.0	-2.5	1.00	1.00	1.00	1.17	0.94	0.95	0.96	0.94

"+"indicates correctly specified model and "−"indicates misspecified model omitting the Z_1 predictor

Tabela B.5: Evaluation criteria for misspecified models. Results for $n = 600$ and $S = 1000$ simulations.

	Relative Bias				Relative Efficiency				Empirical Coverage			
Structure	β_{01}	β_{02}	X	Z	β_{01}	β_{02}	X	Z	β_{01}	β_{02}	X	Z
						Available						
ind	119.9	-31.8	-22.5	61.8	1.00	1.00	1.00	1.00	0.01	0.08	0.92	0.50
exch	95.2	-24.9	-25.1	30.8	1.02	1.01	1.01	1.21	0.08	0.27	0.89	0.76
unst	94.3	-26.2	-27.9	29.8	1.03	1.02	1.02	1.21	0.09	0.20	0.88	0.79
unif	93.5	-24.5	-28.5	25.3	1.02	1.00	1.01	1.24	0.08	0.26	0.90	0.82
cat.exch	93.5	-24.5	-28.5	24.7	1.02	1.00	1.01	1.24	0.08	0.26	0.90	0.83
time.exch	94.9	-24.7	-29.4	25.3	1.02	1.00	1.01	1.25	0.07	0.28	0.89	0.83
RC	93.4	-24.0	-26.9	26.0	1.02	1.00	1.01	1.24	0.08	0.31	0.89	0.82
						WGEE(r^-)						
ind	19.1	-5.6	-28.1	34.8	1.00	1.00	1.00	1.00	0.92	0.92	0.91	0.82
exch	18.5	-5.6	-28.5	21.7	1.00	0.98	0.99	1.18	0.92	0.92	0.90	0.89
unst	19.0	-6.0	-31.2	20.3	1.00	0.99	0.99	1.18	0.92	0.93	0.90	0.88
unif	19.2	-6.0	-30.3	20.8	1.00	1.00	0.99	1.18	0.92	0.92	0.91	0.90
cat.exch	19.3	-6.0	-30.3	20.4	1.00	1.00	0.99	1.18	0.92	0.92	0.91	0.90
time.exch	19.8	-6.0	-30.9	20.8	1.00	1.00	0.99	1.18	0.92	0.91	0.91	0.90
RC	18.7	-5.2	-27.9	22.1	1.00	0.99	0.99	1.18	0.92	0.92	0.89	0.88
						MIGEE(x^-)						
ind	19.9	-7.2	-47.9	-8.1	1.00	1.00	1.00	1.00	0.92	0.91	0.86	0.92
exch	20.5	-7.0	-46.7	-2.6	1.00	1.00	1.00	1.21	0.91	0.90	0.87	0.94
unst	20.5	-7.4	-48.8	-3.2	1.00	1.00	0.99	1.22	0.90	0.90	0.84	0.96
unif	19.8	-7.3	-47.8	-3.7	1.00	1.00	0.99	1.22	0.91	0.90	0.86	0.95
cat.exch	19.8	-7.3	-47.8	-4.2	1.00	1.00	0.99	1.22	0.92	0.90	0.86	0.94
time.exch	20.3	-7.4	-48.5	-3.7	1.00	1.01	1.00	1.22	0.91	0.89	0.83	0.94
RC	19.7	-6.9	-46.6	-3.3	1.00	1.00	0.99	1.22	0.90	0.90	0.85	0.95
						DRGEE(x^-, r^-)						
ind	16.1	-5.8	-41.1	-6.7	1.00	1.00	1.00	1.00	0.92	0.91	0.86	0.92
exch	15.8	-5.5	-39.7	-2.9	0.99	0.98	0.99	1.14	0.93	0.90	0.86	0.94
unst	16.4	-5.9	-42.1	-3.9	0.99	0.98	0.99	1.14	0.92	0.92	0.86	0.96
unif	15.7	-5.8	-41.1	-3.4	0.99	0.99	0.99	1.15	0.92	0.91	0.86	0.95
cat.exch	15.8	-5.9	-41.1	-3.8	0.99	0.99	0.99	1.15	0.92	0.91	0.86	0.95
time.exch	16.3	-5.9	-41.5	-3.5	0.99	0.99	0.99	1.15	0.92	0.90	0.87	0.94
RC	15.2	-5.2	-38.6	-2.0	1.00	0.99	0.99	1.14	0.92	0.91	0.86	0.95

"+"indicates correctly specified model and "−"indicates misspecified model omitting the Z_1 predictor

Tabela B.6: Evaluation criteria for correctly specified models. Results for $n = 600$ and $S = 1000$ simulations.

	Relative Bias				Relative Efficiency				Empirical Coverage			
Structure	β_{01}	β_{02}	X	Z	β_{01}	β_{02}	X	Z	β_{01}	β_{02}	X	Z
						Complete						
ind	1.0	-0.2	-1.0	0.3	1.00	1.00	1.00	1.00	0.96	0.96	0.95	0.93
exch	1.2	-0.1	0.1	0.2	1.00	1.00	1.00	1.26	0.97	0.96	0.96	0.95
unst	0.2	-0.2	-1.0	0.3	1.00	1.00	1.00	1.26	0.96	0.95	0.96	0.95
unif	0.9	-0.2	-0.9	0.2	1.00	1.01	1.00	1.28	0.96	0.95	0.95	0.94
cat.exch	0.9	-0.2	-1.0	-0.2	1.00	1.01	1.00	1.28	0.96	0.95	0.95	0.94
time.exch	0.4	0.0	0.5	0.2	1.00	1.00	1.00	1.28	0.96	0.95	0.94	0.95
RC	0.0	0.4	1.4	0.1	1.00	1.00	1.00	1.28	0.95	0.95	0.96	0.96
						WGEE(r^+)						
ind	2.2	-0.6	-0.3	0.6	1.00	1.00	1.00	1.00	0.98	0.97	0.97	0.95
exch	1.7	-0.3	0.9	-0.3	0.99	0.98	0.99	1.20	0.97	0.96	0.95	0.96
unst	2.2	-0.8	-1.5	-0.8	0.99	0.98	0.99	1.20	0.97	0.96	0.97	0.97
unif	2.0	-0.8	-0.7	-1.7	0.99	0.99	0.99	1.20	0.97	0.96	0.96	0.95
cat.exch	2.1	-0.9	-0.7	-2.1	0.99	0.99	0.99	1.20	0.97	0.96	0.96	0.95
time.exch	2.3	-0.7	-0.9	-1.5	0.99	0.99	0.99	1.20	0.96	0.95	0.95	0.95
RC	1.7	-0.1	1.5	0.3	0.99	0.99	0.99	1.20	0.97	0.96	0.96	0.95
						MIGEE(x^+)						
ind	3.6	-1.1	-5.6	-0.2	1.00	1.00	1.00	1.00	0.96	0.95	0.97	0.92
exch	4.3	-1.1	-5.0	-0.3	0.99	0.99	1.00	1.21	0.95	0.95	0.96	0.93
unst	4.2	-1.5	-7.5	-0.9	0.99	0.99	1.00	1.21	0.95	0.95	0.97	0.95
unif	3.6	-1.2	-5.6	-0.8	1.00	1.00	1.01	1.22	0.96	0.96	0.96	0.94
cat.exch	3.7	-1.3	-5.8	-1.4	1.00	1.00	1.01	1.22	0.96	0.96	0.96	0.94
time.exch	3.9	-1.4	-6.0	-1.0	1.00	1.00	1.01	1.22	0.95	0.94	0.95	0.94
RC	3.4	-0.8	-4.0	-0.8	1.00	1.00	1.01	1.22	0.94	0.96	0.96	0.95
						DRGEE(x^+, r^+)						
ind	-0.3	0.2	-0.5	1.0	1.00	1.00	1.00	1.00	0.97	0.96	0.97	0.94
exch	-0.6	0.3	0.2	0.9	1.01	0.98	0.99	1.17	0.96	0.94	0.96	0.93
unst	-0.3	-0.1	-1.8	0.5	1.01	0.98	0.99	1.16	0.95	0.94	0.96	0.95
unif	-0.5	0.0	-0.9	0.1	1.00	1.00	1.00	1.17	0.95	0.95	0.96	0.95
cat.exch	-0.5	0.0	-1.0	-0.3	1.00	1.00	1.00	1.17	0.95	0.95	0.96	0.95
time.exch	0.4	0.0	-1.6	0.4	1.00	1.00	0.99	1.17	0.94	0.95	0.95	0.95
RC	-1.0	0.7	1.4	1.6	1.00	1.00	1.00	1.16	0.95	0.95	0.95	0.95
						DRGEE(x^-, r^+)						
ind	0.9	-0.3	-3.6	0.6	1.00	1.00	1.00	1.00	0.97	0.96	0.97	0.93
exch	0.4	0.0	-2.3	0.9	0.99	0.98	1.00	1.18	0.96	0.95	0.95	0.92
unst	0.9	-0.5	-4.7	0.4	0.99	0.98	1.01	1.18	0.96	0.96	0.96	0.95
unif	0.7	-0.4	-4.0	0.2	1.00	1.00	1.01	1.18	0.96	0.96	0.96	0.95
cat.exch	0.8	-0.4	-4.1	-0.2	1.00	1.00	1.01	1.19	0.96	0.96	0.96	0.95
time.exch	1.2	-0.2	-3.7	0.7	0.99	0.99	1.00	1.18	0.95	0.95	0.96	0.95
RC	0.3	0.3	-1.9	1.8	1.00	0.99	1.01	1.18	0.96	0.96	0.96	0.95
						DRGEE(x^+, r^-)						
ind	0.0	0.0	-1.4	0.7	1.00	1.00	1.00	1.00	0.97	0.96	0.96	0.94
exch	-0.2	0.2	-0.6	0.1	0.99	0.98	0.99	1.15	0.95	0.94	0.95	0.93
unst	-0.1	-0.1	-2.1	-1.0	0.99	0.98	0.99	1.15	0.95	0.94	0.96	0.95
unif	-0.2	-0.1	-1.9	-0.3	1.00	0.99	0.99	1.16	0.95	0.95	0.96	0.95
cat.exch	-0.1	-0.2	-1.9	-0.7	1.00	0.99	0.99	1.16	0.95	0.95	0.96	0.95
time.exch	0.6	-0.2	-2.8	-0.5	0.99	0.99	0.99	1.16	0.94	0.94	0.94	0.95
RC	-0.5	0.5	0.1	1.1	1.00	0.99	0.99	1.15	0.94	0.94	0.96	0.95

"+"indicates correctly specified model and "−"indicates misspecified model omitting the Z_1 predictor

Tabela B.7: Convergence rate for the simulation study

	sample size			
	50	150	300	600
ind	0.97	1.00	1.00	1.00
exch	0.59	0.89	0.96	1.00
unst	0.08	0.59	0.89	0.99
unif	0.97	1.00	1.00	1.00
cat.exch	0.97	1.00	1.00	1.00
time.exch	0.96	1.00	1.00	1.00
RC	0.67	0.99	1.00	1.00

Table B.7 shows the convergence rate (CR) obtained from the simulation results for seven association structures and four sample sizes. The convergence rate for the working association structure (C) was calculated as

$$CR^C = \frac{1000}{\text{total number of simulations}}.$$

The local odds ratio parametrization presents a clear advantage over the correlation coefficient in terms of convergence issues, specially for unstructured association matrices and small sample sizes.

APÊNDICE C

RESULTADOS ADICIONAIS PARA A ANÁLISE DOS DADOS REAIS NO ARTIGO 2

Tabela C.1: Results for Rheumatic Mitral Stenosis study under exchangeability, time exchangeability and RC association structures

	Exchangeable			Time Exchangeability			RC		
Parameter	Est.	SE	p	Est.	SE	p	Est.	SE	p
				Available					
β_{01}	-1.064	0.782	0.174	-0.909	0.730	0.213	-0.881	0.744	0.236
β_{02}	0.721	0.776	0.353	0.890	0.724	0.219	0.924	0.738	0.210
Sucess	0.680	0.397	0.086	0.626	0.387	0.105	0.641	0.391	0.101
Total Score	-0.108	0.107	0.313	-0.130	0.101	0.197	-0.137	0.102	0.179
Cn	0.468	0.339	0.168	0.477	0.314	0.129	0.496	0.317	0.118
time=2	1.083	0.398	0.007	1.022	0.383	0.008	1.081	0.390	0.006
time=3	1.208	0.402	0.003	1.120	0.380	0.003	1.214	0.392	0.002
				WGEE					
β_{01}	-1178	0.807	0.144	-0.934	0.756	0.217	-0.892	0.771	0.248
β_{02}	0.611	0.797	0.443	0.848	0.741	0.252	0.898	0.757	0.236
Sucess	0.744	0.410	0.070	0.623	0.400	0.119	0.645	0.404	0.111
Total Score	-0.101	0.109	0.355	-0.127	0.103	0.216	-0.136	0.105	0.195
Cn	0.490	0.351	0.162	0.433	0.322	0.179	0.449	0.325	0.167
time=2	1.037	0.414	0.012	1.013	0.398	0.011	1.060	0.404	0.009
time=3	1.395	0.418	0.001	1.363	0.405	0.001	1.437	0.415	0.001
				MIGEE					
β_{01}	-1.100	0.618	0.075	-0.865	0.573	0.131	-0.908	0.567	0.109
β_{02}	0.743	0.621	0.231	0.973	0.575	0.091	0.934	0.572	0.102
Sucess	1.079	0.334	0.001	0.834	0.322	0.010	0.887	0.324	0.006
Total Score	-0.098	0.083	0.239	-0.131	0.078	0.094	-0.131	0.078	0.093
Cn	0.237	0.280	0.397	0.281	0.265	0.288	0.350	0.272	0.198
time=2	1.095	0.317	0.001	1.075	0.320	0.001	1.033	0.323	0.001
time=3	1.071	0.316	0.001	1.091	0.308	0.000	1.072	0.313	0.001
				DRGEE					
β_{01}	-1.199	0.781	0.125	-0.896	0.692	0.195	-0.852	0.700	0.224
β_{02}	0.591	0.785	0.451	0.888	0.686	0.195	0.934	0.696	0.180
Sucess	0.808	0.380	0.034	0.727	0.368	0.048	0.748	0.371	0.044
Total Score	-0.097	0.108	0.368	-0.137	0.094	0.148	-0.143	0.096	0.135
Cn	0.475	0.373	0.202	0.490	0.326	0.133	0.495	0.330	0.134
time=2	0.937	0.630	0.137	0.985	0.534	0.065	0.983	0.540	0.069
time=3	1.203	0.601	0.046	1.189	0.498	0.017	1.208	0.507	0.017