JOSÉ LUIZ PADILHA DA SILVA

MÉTODOS DE IMPUTAÇÃO MÚLTIPLA PARA GEE EM ESTUDOS LONGITUDINAIS

Dissertação apresentada ao Programa de Pós-Graduação em Estatística, Setor de Ciências Exatas, Universidade Federal de Minas Gerais.
Orientador: Prof. Dr. Enrico Antonio Colosimo

BELO HORIZONTE

2011

À Juliana e Magda,

razões 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.

Ao professor Enrico Colosimo, pela orientação, paciência e incentivo.

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

A FAPEMIG pela apoio financeiro.

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

SUMÁRIO

LISTA DE TABELAS

RESUMO

Em estudos longitudinais, dados ausentes constituem um grande desafio para análise.

A presente dissertação mostra como dados ausentes podem apresentar grande impacto na estimação de quantidades de interesse quando se opta pelo modelo GEE como método de análise. Esse método – flexível por não requerer a especificação da distribuição da variável resposta do indivíduo – apresenta estimativas válidas dos coeficientes de regressão apenas na situação MCAR, isto é, quando a perda ocorre completamente ao acaso. Como essa suposição é raramente encontrada na prática, exploramos outro mecanismo de perda de dados.

A fim de corrigir o vício nas estimativas dos coeficientes de regressão, focamos na imputação múltipla, técnica proposta por Little & Rubin (1987) e que tem recebido grande destaque na literatura. Consiste em predizer os valores ausentes de forma a obter conjuntos de dados completos que podem ser analisados por meio de métodos padrão de análise.

Abordamos cinco métodos de imputação de dados: três dos quais consideram um modelo de regressão e dois utilizam alguma forma de pareamento.

Além dos resultados de simulação, em que comparamos os desempenhos desses diferentes métodos de imputação, entre eles um proposto, apresentamos também uma aplicação com dados reais.

Os resultados indicam que a imputação de dados é uma ferramenta adequada para remover o vício das estimativas no modelo GEE, sendo o maior ganho obtido com métodos baseados em regressão.

Palavras-chave: estudos longitudinais, dados ausentes, imputação múltipla, GEE, regressão, pareamento.

ABSTRACT

Missing data is a major challenge for longitudinal data analysis.

This dissertation shows how missing data may have a great impact on the estimation of quantities of interest when one chooses to use the GEE model. This approach - flexible in the sense that the joint distribution of a subject's response vector does not need to be specified - yield valid estimates of the regression coefficients only with data missing completely at random (MCAR). Because this assumption is rarely true in practice, we explored another missing data mechanism.

In order to correct the bias in regression coefficient estimates, we focus on multiple imputation, a technique proposed by Little & Rubin (1987) that has received great attention in the literature. It consists of predicting missing values in order to obtain complete data sets that can be analyzed using standard methods.

We discuss five methods for imputing missing data, three of which consider a regression model and two use some form of matching.

Besides the simulation results, in which we compared the performance of these imputation methods, among them one proposed, we present an application with real data.

The results show that multiple imputation is an appropriate tool to remove the bias of the estimates in the GEE model, the largest gain obtained with regression-based models.

Keywords: longitudinal studies, missing data, multiple imputation, GEE, regression, matching.

CAPÍTULO 1

INTRODUÇÃO

A presente dissertação tem por objetivo o estudo do impacto de observações ausentes (*missing data*) em estudos longitudinais. Dados ausentes ou incompletos são muito comuns em variadas situações estatísticas. Dependendo do mecanismo que gera a não resposta podemos nos deparar com resultados enganosos decorrentes de estimativas viesadas para os parâmetros no modelo de regressão. Tal fato ocorre principalmente se os dados completos constituem uma amostra não representativa daquela população para a qual queremos generalizar os resultados. Assim, conhecer o mecanismo de geração da não-resposta será fundamental para fazermos inferências corretas.

Em estudos longitudinais, dados ausentes constituem um grande desafio para análise. Embora a maioria dos estudos longitudinais seja delineada para coletar os dados em todos os indivíduos na amostra em cada instante de tempo, muitos deles apresentam com frequência observações faltantes. Na área da saúde, por exemplo, os dados ausentes são uma regra, não uma exceção (Fitzmaurice *et al.*, 2004).

Entre as possíveis causas para os dados ausentes podemos citar: óbito, perda de contato com o indivíduo, doenças não relacionadas ao medicamento em estudo, erros de digitação, ineficiência do tratamento, questões mal formuladas em questionários, ou mesmo falta de colaboração dos indivíduos integrantes do estudo, etc.

Dados ausentes têm três implicações gerais para a análise:

i *O conjunto de dados torna-se desbalanceado no tempo, o que acarreta complicações para os métodos de análise que requerem dados balanceados;*

ii *Há perda de informação com redução na eficiência ou um decréscimo na precisão com que mudanças na resposta média podem ser estimadas. Análises restritas aos indivíduos com dados completos geralmente terão menos eficiência que os métodos*

que usam todos os dados disponíveis;

iii Sob certas circunstâncias os dados ausentes podem introduzir vícios e levar a inferências enganosas sobre as mudanças na resposta média.

Como forma de tratamento de dados ausentes em estudos longitudinais focaremos, neste trabalho, a imputação múltipla. A imputação múltipla foi proposta por Little & Rubin (1987) e tem recebido grande destaque na literatura. Consiste em substituir cada valor ausente por um conjunto de m valores plausíveis gerando, assim, m conjuntos de dados completados. As estimativas das análises desses conjuntos de dados completados são combinadas, levando em conta a incerteza associada com o processo de imputação.

No Capítulo 2 são apresentados o Modelo Linear Misto e as Equações de Estimação Generalizadas, métodos que têm sido muito utilizados para a análise de dados longitudinais. No Capítulo 3 é abordado o problema de dados ausentes em estudos longitudinais e suas implicações para análise. Os mecanismos que geram a não resposta são também apresentados e discutidos neste capítulo. O Capítulo 4 mostra como a imputação pode ser aplicada a fim se obter resultados mais coerentes na presença de dados incompletos. São mostradas formas de imputação simples e múltipla. Para a imputação múltipla destaca-se como os resultados podem ser combinados com a finalidade de se fazer inferências.

O Capítulo 5 discute os métodos de imputação. São apresentadas várias formas de imputar os dados, seja por meio da geração de um valor através de um modelo de regressão ou por pareamento, o qual busca nos próprios dados disponíveis o valor mais adequado para ser usado como imputação.

No Capítulo 6 são apresentados resultados de simulação que exemplificam numericamente como os dados ausentes podem afetar as estimativas das quantidades de interesse. Nesse capítulo são aplicados os diferentes métodos de imputação abordados no Capítulo 5 e seus desempenhos são comparados.

Por fim, o Capítulo 7 apresenta uma aplicação dos métodos estudados em um banco de dados real.

CAPÍTULO 2

MODELOS PARA DADOS LONGITUDINAIS

2.1 O Modelo Linear Misto

Um enfoque usual para análise de dados longitudinais é utilizar os Modelos Lineares de Efeitos Mistos (Laird & Ware, 1982). Nesses modelos, a premissa subjacente é que algum subconjunto dos modelos de regressão pode variar de indivíduo para indivíduo, levando em conta, portanto, fontes naturais de heterogeneidade na população. Assim, indivíduos na população têm sua própria trajetória média específica sobre o tempo e um subconjunto dos coeficientes da regressão são considerados aleatórios. A característica distintiva dos modelos lineares de efeitos mistos é que a resposta média é modelada como uma combinação das características da população que são compartilhadas por todos os indivíduos, e efeitos específicos, únicos de cada indivíduo. Os primeiros efeitos são chamados de *fixos* e os últimos de *aleatórios*. O termo *misto* é usado nesse contexto para denominar o modelo que contém tanto efeitos fixos como aleatórios.

Um aspecto muito atrativo desse modelo é a sua flexibilidade em acomodar qualquer grau de desbalanceamento nos dados longitudinais, além de sua habilidade para levar em conta a covariância entre as medidas repetidas de forma relativamente parcimoniosa. Assim, o modelo de efeitos mistos não requer o mesmo número de observações em cada indivíduo nem que as medidas sejam tomadas no mesmo conjunto de ocasiões. Como resultado, esses modelos são particularmente adequados para analisar dados longitudinais inerentemente desbalanceados.

2.1.1 O Modelo

O Modelo Linear Misto é dado por:

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n; \tag{2.1}$$

em que: $\mathbf{Y}_i$ é o vetor de respostas do i-ésimo indivíduo, de dimensão $n_i \times 1$; $\mathbf{X}_i$ é uma matriz conhecida, de dimensão $n_i \times p$, que faz a ligação entre $\boldsymbol{\beta}$ e $\mathbf{y}_i$; $\boldsymbol{\beta}$ é o vetor de efeitos fixos, de dimensão $p \times 1$; $\mathbf{Z}_i$ é uma matriz de covariáveis conhecida, de dimensão $n_i \times q$, que faz a ligação entre $\mathbf{b}_i$ e $\mathbf{y}_i$, sendo $\mathbf{Z}_i$ um subconjunto de $\mathbf{X}_i$; $\mathbf{b}_i$ é o vetor de efeitos aleatórios, de dimensão $q \times 1$; $\boldsymbol{\varepsilon}_i$ é o vetor de erros aleatórios, de dimensão $n_i \times 1$; n_i é o número de observações realizadas no i-ésimo indivíduo; n é o número de indivíduos na amostra; p é o número de parâmetros e q é o número de efeitos aleatórios.

As suposições usuais do modelo (2.1) são:

$$\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \mathbf{E}_i), \quad \mathbf{b}_i \sim N(\mathbf{0}, \mathbf{B})$$

$$Cov(\mathbf{b}_i, \mathbf{b}_{i'}) = 0, \quad Cov(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_{i'}) = 0 \text{ para } i \neq i',$$

e $\mathbf{b}_i$ e $\boldsymbol{\varepsilon}_i$ são assumidos independentes. Como consequência do modelo linear misto (2.1), tem-se que:

- $E(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta} = \boldsymbol{\mu}_i$
- $Var(\mathbf{Y}_i) = \mathbf{Z_i}\mathbf{B}\mathbf{Z_i}' + \mathbf{E}_i = \mathbf{V}_i$
- $Cov(\mathbf{Y}_i, \mathbf{Y}_{i'}) = 0$ para $i \neq i'$.

Como $\mathbf{Y}_i$ é uma combinação linear de $\mathbf{b}_i$ e $\boldsymbol{\varepsilon}_i$ tem-se que $\mathbf{Y}_i \sim N(\boldsymbol{\mu}_i, \mathbf{V}_i)$. O modelo (2.1) pode ser simplificado quando $\mathbf{E}_i = \sigma^2\mathbf{I}_i$, em que $\mathbf{I}_i$ é uma matriz identidade de dimensão $n_i \times n_i$. Quando isso ocorre, o modelo é denominado de *modelo de independência condicional*, pois, ao condicionar as n_i respostas do i-ésimo indivíduo aos vetores $\mathbf{b}_i$ e $\boldsymbol{\beta}$, elas se tornam independentes. Quando a quantidade de observações medidas é a mesma para todos os indivíduos e nas mesmas ocasiões ($n_i = m, \forall i = 1, 2, \ldots, n$) considera-se que o modelo tem dados balanceados com relação ao número de observações, caso contrário, o modelo possui dados desbalanceados.

Como as observações em diferentes indivíduos são consideradas independentes, a matriz de variância-covariância de todas as observações é bloco diagonal, de dimensão $\sum_{i=1}^{n} n_i \times$

$\sum_{i=1}^{n} n_i$, no caso geral (que permite acomodar dados desbalanceados) e bloco diagonal de dimensão $mn \times mn$ no caso de dados desbalanceados, sendo cada bloco igual a $\mathbf{V}_i$

2.1.2 Estimação

Seja $\boldsymbol{\zeta}$ o vetor de componentes de variância em $\boldsymbol{V}_i$, e seja $\boldsymbol{\theta} = (\boldsymbol{\beta'}, \boldsymbol{\zeta'})'$ o vetor que inclui todos os parâmetros do modelo para $\boldsymbol{Y}_i$. A inferência é baseada na seguinte função de log-verossimilhança:

$$l(\boldsymbol{\theta}) \propto -\frac{1}{2}\sum_{i=1}^{n} log|\boldsymbol{V}_i| - \frac{1}{2}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}). \tag{2.2}$$

Fixados valores de $\boldsymbol{\zeta}$, $\boldsymbol{\beta}$ pode ser estimado por máxima verossimilhança perfilada, produzindo

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{X'}_i\boldsymbol{V}^{-1}\boldsymbol{X}_i\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{X'}_i\boldsymbol{V}^{-1}\boldsymbol{Y}_i\right). \tag{2.3}$$

Os estimadores dos componentes de variância podem ser obtidos usando estimação por máxima verossimilhança ou máxima verossimilhança restrita (Patterson & Thompson, 1971). O termo $\boldsymbol{V}_i(\boldsymbol{\zeta})$ em (2.3) é então substituído por $\boldsymbol{V}_i(\hat{\boldsymbol{\zeta}})$. Os efeitos aleatórios podem ser preditos pelo estimador BLUP (veja, por exemplo, Verbeke & Molenberghs (2000)).

2.2 Equações de Estimação Generalizadas

Se o interesse primário está na relação entre a média populacional da resposta e as covariáveis, pode-se usar um enfoque robusto, o método das Equações de Estimação Generalizadas (GEE). O método GEE foi proposto por Liang & Zeger (1986) e pode ser pensado como uma extensão de Modelos Lineares Generalizados (MLG's) para dados correlacionados. Nesse método, ao invés de especificar a distribuição completa da variável resposta, é preciso especificar apenas a média. Nesse sentido o enfoque é semiparamétrico. Para produzir estimativas consistentes, o enfoque GEE requer apenas a especificação correta da estrutura de média das variáveis respostas, sem fazer qualquer suposição distribucional.

Isso torna fácil sua aplicação e extensão a variáveis de vários tipos.

Num enfoque GEE, correlações entre as respostas são tratadas como parâmetros de perturbação, mas a especificação correta da estrutura de variância-covariância melhora a precisão das estimavas. Esses modelos são chamados *marginais* porque modelam uma regressão de $\boldsymbol{Y}$ em $\boldsymbol{X}$ e a dependência intra-indivíduos (isto é, os parâmetros de associação) são tratados separadamente. Assim, o modelo para a resposta média depende apenas das covariáveis de interesse, não de quaisquer efeitos aleatórios ou respostas anteriores.

Para aplicação dos modelos GEE assumimos que haja um número n fixo de avaliações nos quais os indivíduos são medidos. Cada indivíduo não precisa ser medido em todas as n avaliações, mas é a matriz de correlação completa $n \times n$ dos dados longitudinais que é considerada num modelo GEE como parâmetros de perturbação.

2.2.1 O Modelo

Os modelos GEE não requerem quaisquer suposições distribucionais sobre a variável resposta. Esses modelos produzem estimadores consistentes e assintoticamente normais para os coeficientes de regressão $\boldsymbol{\beta}$, mesmo com má especificação da estrutura de covariância para os dados longitudinais (Liang & Zeger, 1986).

Inicialmente é necessário especificarmos as respostas médias como função das covariáveis, $E(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta} = \boldsymbol{\mu}_i$, depois especificamos uma matriz de correlação "de trabalho" das medidas repetidas, $\boldsymbol{R}$, que depende de um vetor de parâmetros de associação denotados por $\boldsymbol{\alpha}$. Esses parâmetros são os mesmos para todos os indivíduos e representam a dependência média entre as observações repetidas dos indivíduos. Embora seja recomendado que a escolha de $\boldsymbol{R}$ seja consistente com as correlações observadas, o método GEE produz estimativas consistentes dos coeficientes de regressão e de seus erros padrões, mesmo com a má especificação da estrutura de correlação. Essa é uma propriedade bastante atrativa dos modelos GEE. A eficiência (isto é, poder estatístico) é reduzida se se opta por uma escolha incorreta de $\boldsymbol{R}$; contudo, a perda de eficiência é reduzida à medida que o número de indivíduos aumenta.

2.2.2 Formas de Correlação de Trabalho

A forma mais simples é a de *independência*, $\boldsymbol{R}_i(\boldsymbol{\alpha}) = \sigma^2 \boldsymbol{I}_i$, equivalente a assumir que os dados longitudinais não são correlacionados.

A próxima estrutura mais simples é assumir que todas as correlações em $\boldsymbol{R}$ são as mesmas, ou "permutáveis". Essa estrutura permutável, referida aqui como *simetria composta*, especifica que $\boldsymbol{R}_i(\boldsymbol{\alpha}) = \rho 11'$. Isso é equivalente a correlação imposta por um modelo linear misto com apenas o intercepto aleatório.

Uma outra estrutura útil é a *AR-1*, para a qual $\boldsymbol{R}_i(\boldsymbol{\alpha}) = \rho^{|j-j'|}$, em que j é o índice do tempo. Essa estrutura é somente válidas para medidas igualmente espaçadas no tempo. Aqui a correlação intra-indivíduo ao longo do tempo é uma função exponencial do distanciamento entre as observações. Essa forma é bastante parcimoniosa para dados longitudinais pois depende de apenas um termo, embora permita que as correlações declinem com o afastamento temporal das observações.

Finalmente, a forma não especificada ou *não estruturada* estima todas as $n(n-1)/2$ correlações de $\boldsymbol{R}$. Essa forma é a mais eficiente, mas é mais útil quando há relativamente poucos períodos de tempo. Quando há muitas ocasiões de medição, a estimação de $n(n-1)/2$ correlações não é parcimoniosa.

Dados ausentes complicam a estimação de $\boldsymbol{R}$ (Hedeker & Gibbons, 2006).

2.2.3 Estimação

O estimador GEE de $\boldsymbol{\beta}$, específico para o modelo linear $E(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta} = \boldsymbol{\mu}_i$, é a solução de

$$\sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{R}_i(\boldsymbol{\alpha})^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0, \tag{2.4}$$

que produz, resolvendo para $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{R}_i(\hat{\boldsymbol{\alpha}})^{-1} \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{R}_i(\hat{\boldsymbol{\alpha}})^{-1} \boldsymbol{y}_i \right), \tag{2.5}$$

em que $\hat{\boldsymbol{\alpha}}$ é o estimador de $\boldsymbol{\alpha}$, que pode ser obtido conforme algoritmo descrito no final desta seção. A forma geral desse estimador para o Modelo Linear Generalizado é encontrada em Liang & Zeger (1986). Esse estimador é o de mínimos quadrados ponderados, sendo a matriz de peso $\boldsymbol{R}_i(\hat{\boldsymbol{\alpha}})$. Aqui a matriz de peso depende de parâmetros a serem estimados (isto é, de $\boldsymbol{\alpha}$). Neste caso, a solução pode ser obtida usando o algoritmo iterativo de mínimos quadrados reponderados (IRLS) no qual estimativas iterativas de $\boldsymbol{\alpha}$ são usadas para produzir novas estimativas de $\boldsymbol{\beta}$, com o procedimento continuando até a convergência. Porque a equação (2.4) depende apenas da média e da variância de $\boldsymbol{Y}$, essas são estimativas de quase-verossimilhança (McCullagh & Nelder, 1989).

Os erros padrões associados com os coeficientes estimados de regressão são obtidos a partir de duas versões.

1. *Naive* ou "baseada no modelo"

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{R}_i(\hat{\boldsymbol{\alpha}})^{-1} \boldsymbol{X}_i \right)^{-1}. \tag{2.6}$$

2. *Robusta* ou "empírica"

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = \boldsymbol{M}_0^{-1} \boldsymbol{M}_1 \boldsymbol{M}_0^{-1}, \tag{2.7}$$

em que

$$\begin{aligned} \boldsymbol{M}_0 &= \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{R}_i(\hat{\boldsymbol{\alpha}})^{-1} \boldsymbol{X}_i, \\ \boldsymbol{M}_1 &= \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{R}_i(\hat{\boldsymbol{\alpha}})^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)' \boldsymbol{R}_i(\hat{\boldsymbol{\alpha}})^{-1} \boldsymbol{X}_i. \end{aligned}$$

O estimador robusto ou "sanduíche" fornece um estimador consistente para $\widehat{Var}(\hat{\boldsymbol{\beta}})$ mesmo quando a estrutura de correlação $\boldsymbol{R}_i(\boldsymbol{\alpha})$ não é a corretamente especificada.

Desta forma, a obtenção de $\hat{\boldsymbol{\beta}}$ e $\widehat{Var}(\hat{\boldsymbol{\beta}})$ para o modelo GEE envolve os seguintes passos:

1. Especificar a forma de $\boldsymbol{R}_i(\boldsymbol{\alpha})$ e tomar

$$\hat{\boldsymbol{\beta}}^* = \left(\sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{y}_i \right).$$

2. Encontrar os resíduos $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}^*$. Através dos resíduos é possível estimar $\boldsymbol{\alpha}$.

3. Encontrar $\hat{\boldsymbol{\beta}}^*$ a partir de (2.5) e $\widehat{Var}(\hat{\boldsymbol{\beta}})$ a partir de (2.7).

CAPÍTULO 3

DADOS AUSENTES EM ESTUDOS LONGITUDINAIS

3.1 Introdução

Um dos maiores desafios para a análise de dados longitudinais é o problema de dados ausentes (*missing data*). Embora a maioria dos estudos longitudinais seja delineada para coletar os dados em todos os indivíduos na amostra em cada instante de tempo, muitos deles têm algumas observações faltantes. Na área da saúde, por exemplo, os dados ausentes são uma regra, não uma exceção (Fitzmaurice *et al.*, 2004).

Em estudos longitudinais o problema de dados ausentes é muito mais grave que nos estudos tranversais, pois a não-resposta pode ocorrer em qualquer ocasião. O termo *missing data* (dado ausente) é usado para indicar que uma medida do indivíduo foi programada, porém não pode ser coletada. Uma distinção geralmente feita é se os dados ausentes são *intermitentes* ou *dropout*. Para compreender a distinção, consideramos a seguinte situação: é tomada uma sequência de medições $Y_1, Y_2, \ldots, Y_m$ na i-ésima unidade experimental. Ocorre *dropout* quando ao ser observado um Y_j faltoso também será observado um Y_k faltoso para todo $k \geq j$; caso contrário os dados são chamados *intermitentes*. Com dados *intermitentes* há uma ou mais perdas pontuais. Em estudos longitudinais que sofrem da condição de *dropout* há perda completa da informação a partir de um certo instante de tempo. Alguns autores preferem denominar as situações descritas como padrão de ausência monótono e não-monótono. Nesse trabalho focaremos nos dados ausentes como padrão *monótono* ou *dropout*, por ser a situação mais encontrada na prática (Verbeke & Molenberghs, 2000).

Dados ausentem têm três implicações gerais para a análise:

- O conjunto de dados torna-se desbalanceado no tempo, já que nem todos os indivíduos têm o mesmo número de medidas repetidas no conjunto comum de ocasiões.

Isso acarreta complicações para os métodos de análise que requerem dados balanceados;

- Há perda de informação. Dados ausentes causam redução na eficiência ou diminuição na precisão com que mudanças na resposta média podem ser estimadas. Análises restritas aos indivíduos com dados completos geralmente terão menos eficiência que os métodos que usam todos os dados disponíveis;

- Sob certas circunstâncias os dados ausentes podem introduzir vícios e levar a inferências enganosas sobre as mudanças na resposta média. Assim, as razões da perda de dados (mecanismo de dado ausente) devem ser cuidadosamente consideradas.

3.2 Hierarquia de Mecanismos de Dados Ausentes

O mecanismo de dados ausentes pode ser pensado como um modelo probabilístico para a distribuição de um conjunto de variáveis resposta indicadoras.

Um indivíduo tem um vetor de respostas denotado por $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{im})'$, com distribuição governada pelos parâmetros $\boldsymbol{\theta}$. Seja $\boldsymbol{D}_i$ um vetor $m \times 1$ de indicadoras da resposta $\boldsymbol{D}_i = (D_{i1}, D_{i2}, \ldots, D_{im})'$, com $D_{ij} = 1$ se Y_{ij} é observado e $D_{ij} = 0$ se Y_{ij} é dado ausente. Em geral, a distribuição de $\boldsymbol{D}$ pode estar relacionada com $\boldsymbol{Y}$, assim, admitimos um modelo de probabilidade para $\boldsymbol{D}$, $P(\boldsymbol{D}|\boldsymbol{Y}, \boldsymbol{\xi})$, que depende de $\boldsymbol{Y}$ assim como de parâmetros desconhecidos $\boldsymbol{\xi}$. Dado $\boldsymbol{D}_i$, o conjunto de respostas, $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{im})'$, pode ser particionado dentro de duas componentes, $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i,obs}, \boldsymbol{Y}_{i,mis})$, correspondendo às respostas observadas e aos dados ausentes, respectivamente.

Uma hierarquia de três diferentes tipos de mecanismos de dados ausentes pode ser distinguida avaliando como $\boldsymbol{D}_i$ está relacionado com $\boldsymbol{Y}_i$ (Little & Rubin, 1987):

1. *Missing Completely at Random* (MCAR): quando a probabilidade de não resposta é independente de dados observados ou não observados, isto é:

$$P(\boldsymbol{D}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{X}, \boldsymbol{\xi}) = P(\boldsymbol{D}|\boldsymbol{\xi}).$$

Nesse caso não são necessários cuidados adicionais na análise. Exemplo de MCAR: erros administrativos que ocorrem ao acaso, tais como acidentes em laboratório, perda de formulário, etc.

2. *Missing at Random* (MAR): quando a probabilidade de não resposta é independente de $\boldsymbol{Y}_{mis}$:

$$P(\boldsymbol{D}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{X}, \boldsymbol{\xi}) = P(\boldsymbol{D}|\boldsymbol{Y}_{obs}, \boldsymbol{X}, \boldsymbol{\xi}).$$

   Essa suposição é atendida se a não resposta está relacionada apenas a valores medidos nos dados mas não a valores não medidos. Exemplos de MAR incluem valores ausentes em indivíduos mais velhos, indivíduos de certa região, ou tempo de calendário. Num mecanismo MAR, os indivíduos com as covariáveis e resposta completos não são mais representativos da população para a qual queremos generalizar os resultados.

3. *Not Missing at Random* (NMAR): quando a probabilidade de não resposta depende de dados não observados $\boldsymbol{Y}_{mis}$:

$$P(\boldsymbol{D}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{X}, \boldsymbol{\xi}) = P(\boldsymbol{D}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{X}, \boldsymbol{\xi}).$$

   O caso NMAR também é referido como não resposta *informativa* ou *não ignorável*. Em contraste com a situação MAR, em que a não resposta pode ser explicada pelas outras variáveis no estudo, NMAR surge devido a um padrão de não resposta explicado pelas variáveis que estão ausentes. Essa é uma situação crítica, já que a não-resposta depende agora dos valores ausentes ou então de valores de variáveis que não foram observadas. Exemplos incluem a não-resposta em certas questões (orientação sexual, renda, etc...), ou condição clínica (não-resposta se uma condição está presente, a qual não pôde ser avaliada de forma precisa).

O primeiro mecanismo possui uma suposição muito forte que raramente é satisfeita na prática. Quando ele ocorre, os dados não observados constituem uma sub-amostra aleatória, sem acarretar problemas de viés à análise. O segundo é também denominado

ignorável. O termo *ignorável* é usado para indicar que não é necessário especificar um modelo para a não resposta. O terceiro mecanismo de ausência é *não ignorável* devido à falta de aleatoriedade da não resposta; ignorar esse mecanismo pode levar a sérios vícios nas estimativas.

Como vimos, compreender qual mecanismo de não-resposta é mais coerente com o fenômeno em estudo é fundamental para fazer inferências corretas. Nesta dissertação estaremos interessados na situação MAR, em que a não resposta depende apenas de quantidades observada.

3.3 *Dropout*

O *dropout* é caracterizado pela perda completa a partir de um instante específico. O termo *dropout* refere-se a um caso especial em que se Y_{ik} é dado ausente, então $Y_{ik+1}, \ldots, Y_{im}$ são dados ausentes também. Quando há *dropout* em um estudo longitudinal, a questão chave é determinar se há alguma diferença relevante entre os indivíduos que tiveram *dropout* e os que permaneceram no estudo longitudinal até o fim. Se eles diferem as análises com os dados completos são potencialmente viesadas.

A mesma taxonomia usada na seção anterior pode ser aplicada ao *dropout* (Fitzmaurice *et al.*, 2004). Isto é, o *dropout* pode *ser completely at random*, *at random* ou *not at random*.

Quando o *dropout* é *completely at random* a probabilidade de *dropout* independe de todo passado, resultados atuais e futuros (dadas as covariáveis). Com o *dropout completely at random*, um indivíduo deixa o estudo em um processo que não está relacionado com os valores observados (respostas) do indivíduo.

Em contraste, quando o *dropout* é *at random*, a probabilidade de *dropout* em cada instante pode depender dos resultados anteriores observados até, mas não incluindo, o instante atual.

Finalmente, quando o *dropout* é *not at random*, a probabilidade de *dropout* em cada instante pode depender dos resultados atuais e futuros não-observados. Nesse caso depende de valores da variável resposta que seriam observados se o indivíduo tivesse permanecido no estudo.

No contexto de *dropout* num estudo longitudinal o termo *informativo* frequentemente é usado para se referir ao *dropout* que é NMAR (similarmente o termo *não informativo* se refere aos casos MAR e MCAR). Aqui o fato do *dropout* é informativo sobre a distribuição das observações futuras. Por exemplo, considere dois indivíduos com o mesmo histórico de respostas (e covariáveis) até o tempo t. Um abandona e o outro não abandona o estudo. Com MAR, a distribuição das observações futuras é a mesma. Em contraste *dropout* NMAR nos informa que a distribuição das observações futuras vai diferir. No caso mais geral, nada nos dados pode ser usado para determinar a distribuição das observações futuras dos *dropouts*, assim a análise depende fortemente da especificação da probabilidade de não-resposta.

3.4 Métodos para Tratar Dados Ausentes em Estudos Longitudinais

Três métodos comumente usados para lidar com dados ausentes em estudos longitudinais são (Fitzmaurice *et al.*, 2009): (i) métodos de imputação; (ii) métodos baseados em verossimilhança; e (iii) métodos de ponderação. De forma geral, esses enfoques estão interrelacionados no sentido que eles "imputam"certos valores para os dados ausentes; a diferença é que em (i) a imputação é explícita e em (ii) e (iii), contudo, a imputação é implícita. Teremos como foco a primeira situação.

- **Métodos de Imputação**

A ideia básica por trás da imputação (Little & Rubin, 1987) é relativamente simples: substituir os valores que não foram registrados por valores imputados. Um dos grandes atrativos dos métodos de imputação é que, uma vez que o conjunto de dados foi construído, métodos padrões para dados completos podem ser aplicados. Na imputação múltipla (IM), os valores ausentes são substituídos por um conjunto de k valores plausíveis, permitindo assim que a incerteza sobre os valores imputados seja levada em conta. Os k conjuntos de dados preenchidos produzem k diferentes conjuntos de estimativas de parâmetros e erros padrões. Estas estimativas são então combinadas para fornecer uma única estimativa dos

parâmetros de interesse, junto com erros padrões que refletem a incerteza inerente na imputação das respostas não observadas.

3.4.1 Justificativa da Imputação

Segundo Little & Rubin (1987), pode ser mostrado que, sob ignorabilidade, não precisamos considerar um modelo para $\boldsymbol{D}$ envolvendo os parâmetros de perturbação $\boldsymbol{\xi}$ ao se fazer inferência sobre $\boldsymbol{\theta}$.

Por que os dados observados consistem não apenas de $\boldsymbol{Y}_{obs}$ mas também de $\boldsymbol{D}$, a distribuição de probabilidade dos dados observados é dada por:

$$\begin{aligned} P(\boldsymbol{D}, \boldsymbol{Y}_{obs}|\boldsymbol{\theta}, \boldsymbol{\xi}) &= \int P(\boldsymbol{D}, \boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\xi}) d\boldsymbol{Y}_{mis} \\ &= \int P(\boldsymbol{D}|\boldsymbol{Y}, \boldsymbol{\xi}) \boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{\theta}) d\boldsymbol{Y}_{mis} \end{aligned} \tag{3.1}$$

Sob MAR (3.1) se torna:

$$\begin{aligned} P(\boldsymbol{D}, \boldsymbol{Y}_{obs}|\boldsymbol{\theta}, \boldsymbol{\xi}) &= P(\boldsymbol{D}|\boldsymbol{Y}_{obs}, \boldsymbol{\xi}) \int P(\boldsymbol{Y}|\boldsymbol{\theta}) d\boldsymbol{Y}_{mis} \\ &= P(\boldsymbol{D}|\boldsymbol{Y}_{obs}, \boldsymbol{\xi}) P(\boldsymbol{Y}_{obs}|\boldsymbol{\theta}). \end{aligned} \tag{3.2}$$

Quando os dois parâmetros $\boldsymbol{\xi}$ e $\boldsymbol{\theta}$ são distintos, inferências de máxima verossimilhança sobre $\boldsymbol{\theta}$ não serão afetadas por $\boldsymbol{\xi}$ ou $P(\boldsymbol{D}|\boldsymbol{Y}_{obs}, \boldsymbol{\xi})$. Assim, na situação MAR (e MCAR como caso particular) estimação por máxima verossimilhança de $\boldsymbol{\theta}$, testes de razão de verossimilhanças sobre $\boldsymbol{\theta}$, etc, podem ser realizados sem levar em conta o mecanismo de geração da não resposta; isto é, o mecanismo de não resposta pode ser seguramente ignorado (Schafer, 1997a).

A função de verossimilhança, ignorando o mecanismo de geração da não resposta, é dada por:

$$L(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}) \propto P(\boldsymbol{Y}_{obs}|\boldsymbol{\theta}). \tag{3.3}$$

Métodos não baseados em verossimilhança, tais como GEE (Liang & Zeger, 1986)

geralmente requerem a forte suposição MCAR para produzirem estimativas consistentes (Fitzmaurice *et al.*, 2009). O modelo GEE padrão requer que tenhamos um modelo para o valor esperado das respostas dado as covariáveis. Quando os dados são MAR, esse modelo para a resposta média geralmente não se manterá para os dados observados e, consequentemente, a validade dos métodos GEE aplicados aos dados disponíveis fica comprometida (Fitzmaurice *et al.*, 2009).

Quando os dados são NMAR, praticamente todos os métodos padrão de análise de dados longitudinais são inválidos (Fitzmaurice *et al.*, 2009). Por exemplo, métodos baseados em verossimilhança que ignoram o mecanismo de geração da não resposta produzirão estimativas viesadas das tendências médias da resposta. Para se obter estimadores válidos é preciso a modelagem conjunta do vetor de resposta média e do mecanismo gerador dos dados ausentes.

Imputação múltipla de valores ausentes de $\boldsymbol{Y}$ sob um modelo paramétrico explícito, está intimamente relacionada à inferência Bayesiana baseada na distribuição *a posteriori* para aquele modelo.

Uma vantagem da IM é que o modelo para gerar as imputações pode ser diferente do modelo usado na análise (Schafer, 1997a); por exemplo, num cenário de ensaio clínico, o modelo de imputação pode condicionar em informações sobre efeitos colaterais que não são parte dos modelos de substantivo interesse, os quais focam nos desfechos clínicos de interesse primário.

CAPÍTULO 4

IMPUTAÇÃO DE DADOS

Podemos classificar a imputação de dados em dois grupos: os métodos de imputação simples e aqueles de imputação múltipla. Nos primeiros, para cada valor ausente é gerado apenas um valor para ser imputado; não é levada em conta a incerteza decorrente do não conhecimento do valor real que foi perdido. Nos últimos, cada valor ausente é substituído por um conjunto de k valores plausíveis, levando assim em consideração a incerteza associada com a imputação.

Neste capítulo apresentaremos brevemente na Seção 4.1 os métodos mais utilizados de imputação simples, e, na Seção 4.2, os de nosso interesse, imputação múltipla. No Capítulo 5, os métodos de imputação múltipla são apresentados com mais detalhes.

4.1 Imputação Simples (IS)

No contexto de dados longitudinais, os valores observados podem ser usados de forma simples para imputar os valores dos dados ausentes. Dentre as abordagens comuns de imputação simples destacamos: imputação pela última observação, pela média e através da regressão.

4.1.1 Imputação Através da Última Observação (IUO)

Os dados ausentes em cada indivíduo são substituídos pelo último valor observado no mesmo indivíduo. Essa técnica, geralmente usada em situações em que há *dropout*, requer suposições muito fortes e frequentemente irrealistas para ter validade. Nota-se que com esse procedimento a tendência longitudinal de aumento/decréscimo do valor da variável resposta em consideração é perdida.

4.1.2 Imputação Através da Média

Em um estudo longitudinal duas maneiras distintas de imputação através da média podem ser consideradas:

1. **Média dos Tempos (IMT)**: média dos valores observados em tempos distintos para a i-ésima unidade experimental.

2. **Média dos Indivíduo (IMI)** média dos valores observados nas diferentes unidades experimentais do t-ésimo tempo.

No primeiro caso, a ideia é calcular a média das observações presentes para a i-ésima unidade experimental nos diferentes tempos, utilizando assim esse valor como imputação para essa unidade experimental nos tempos ausentes. Nota-se que esse método pode não fazer sentido em experimentos onde existe tendência natural de aumento do valor da resposta como função do tempo, por exemplo em medidas de crescimento. No segundo caso, o valor imputado é calculado como a média das observações presentes das unidades experimentais em um tempo fixado. Assim, cada valor ausente em um determinado tempo é substituído por esse valor. Tal procedimento pode fazer mais sentido quando existem poucos grupos sob estudo e os indivíduos apresentam trajetórias semelhantes ao longo do tempo.

4.1.3 Imputação Através da Regressão

Uma forma mais promissora de imputação é através de modelos de regressão. Com esse método os valores ausentes são substituídos por valores preditos através de um modelo de regressão ajustado com os dados observados. Essa técnica pode ser combinada com algum tipo de pareamento, situação na qual não é diretamente imputado o valor predito pelo modelo, mas escolhido dentre os observados aquele mais próximo do predito. Isso restringe o intervalo de valores que podem ser imputados, impedindo discrepâncias entre estes valores imputados e aqueles realmente observados para os indivíduos com dados completos.

4.2 Imputação Múltipla (IM)

A imputação múltipla tem sido aplicada a muitos cenários com dados ausentes (Horton & Lipsitz, 2001). O processo de imputação múltipla consiste basicamente de três passos:

1. **Imputação:** Para cada valor ausente são gerados $k(k \geq 2)$ valores;

2. **Análise:** Os k valores são organizados de forma que o primeiro valor imputado para cada dado ausente produz o primeiro conjunto de dados completado, o segundo valor imputado para cada dado ausente produz o segundo conjunto de dados completado e assim por diante. Cada conjunto de dados completado é analisado por métodos tradicionais para dados completos;

3. **Combinação:** Finalmente, os resultados das k análises são combinados numa análise final permitindo que a incerteza associada à imputação seja considerada.

Os cálculos para o passo da combinação das análises são descritos por Little & Rubin (1987):

Sejam $\hat{\beta}_i$ e $\hat{U}_i$ as estimativas pontuais e de variância para o i-ésimo conjunto de dados imputado $i = 1, 2, \ldots, k$. Então a estimativa pontual para β das múltiplas imputações é a média das k estimativas dos dados completos:

$$\bar{\beta} = \frac{1}{k} \sum_{i=1}^{k} \hat{\beta}_i.$$

Seja $\bar{U}$ a variância entre imputações, que é a média das k estimativas de dados completos:

$$\bar{U} = \frac{1}{k} \sum_{i=1}^{k} \hat{U}_i,$$

e B a variância intra imputações:

$$B = \frac{1}{k-1} \sum_{i=1}^{k} (\hat{\beta}_i - \bar{\beta})^2.$$

Então, a variância estimada associada com $\bar{\beta}$ é a variância total:

$$T = \bar{U} + \left(1 + \frac{1}{k}\right) B.$$

A estatística $(\beta - \bar{\beta})T^{-1/2}$ é aproximadamente distribuída com distribuição t com v_k graus de liberdade, em que

$$v_k = (k-1)\left\{1 + \frac{\bar{U}}{(1+k^{-1})B}\right\}^2. \tag{4.1}$$

Tem sido notado que um número relativamente pequeno de imputações produz estimavas de erros padrões que são quase totalmente eficientes (Schafer, 1997a). Na prática não mais de 10 imputações são geralmente usadas.

4.2.1 Notas Sobre Imputação

Segundo Schafer (1997a) inferência por imputação múltipla pode ser robusta a desvios do modelo de imputação se a quantidade de informação ausente não é grande, porque o modelo de imputação é aplicado não ao conjunto completo dos dados mas apenas às porções ausentes deste.

Deve ser notado que uma vantagem da imputação múltipla como enfoque analítico é que esta permite a incorporação de informações adicionais no modelo de imputação. Essa informação auxiliar pode não ser de interesse no modelo de regressão, mas pode tornar a suposição MAR mais plausível (Horton & Lipsitz, 2001). A inclusão dessa informação no modelo de imputação é direta.

Algumas sugestões práticas de quais variáveis incluir no modelo de imputação foram dadas por van Buuren *et al.* (1999). Eles recomendam que esse conjunto de variáveis inclua aquelas que estão no modelo de dados completos, fatores que conhecidamente estejam associados com a probabilidade de não resposta e fatores que expliquem uma considerável porção de variância na variável resposta.

CAPÍTULO 5

MÉTODOS DE IMPUTAÇÃO MÚLTIPLA

Nesse capítulo são apresentados os métodos de imputação de interesse desta dissertação. Os dois primeiros utilizam uma abordagem bayesiana para imputação, sendo que o primeiro considera o modelo normal multivariado e o segundo o modelo linear misto de regressão. Ambos estão implementados no *software R*, nos pacotes *norm* e *pan*, respectivamente. Nestes métodos os valores imputados são gerados explicitamente de um modelo de regressão.

Os dois próximos métodos consideram uma forma de imputação implícita, ou seja, na qual os valores imputados não são preditos por um modelo de regressão, mas obtidos a partir de valores observados de indivíduos com trajetórias longitudinais semelhantes. Esses dois métodos baseiam-se em alguma forma de pareamento para construção das imputações.

Por fim, o último método considerado aqui adota também o modelo misto de regressão para imputar os dados perdidos. De forma semelhante aos dois primeiros procedimentos, os coeficientes usados no modelo não são aqueles obtidos por máxima verossimilhanca, mas sim simulados da distribuição dos seus estimadores.

5.1 IMPUTAÇÃO PELO MODELO NORMAL – BAYESIANO

O modelo de probabilidade mais comum para dados multivariados contínuos é a distribuição normal multivariada. Embora algumas variáveis incompletamente observadas sejam claramente não normais, ainda assim pode ser razoável usar o modelo normal como uma ferramenta conveniente para criar imputações múltiplas.

5.1.1 Características Relevantes do Modelo de Dados Completos

Seja $\boldsymbol{Y}$ uma matriz com n linhas e m colunas, com as linhas correspondendo às unidades observacionais e colunas correspondendo às variáveis. Seja $\boldsymbol{y}_{ij}$ um elemento individual de $\boldsymbol{Y}$, $i = 1, \ldots, n$ e $j = 1, \ldots, m$. A i-ésima linha de $\boldsymbol{Y}$, é

$$\boldsymbol{y}_i = (\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2}, \ldots, \boldsymbol{y}_{im})'.$$

Assumimos que $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ são realizações independentes de um vetor aleatório, denotado por $(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m)'$, que tem uma distribuição normal multivariada com vetor de média $\boldsymbol{\mu}$ e covariância $\boldsymbol{V}$, isto é:

$$\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n | \boldsymbol{\theta} \overset{iid}{\sim} N_m(\boldsymbol{\mu}, \boldsymbol{V}),$$

em que $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{V})$ são os parâmetros desconhecidos. A única restrição em $\boldsymbol{\theta}$ é que $\boldsymbol{V}$ seja positiva definida.

A densidade de uma única linha é:

$$P(\boldsymbol{y}_i|\boldsymbol{\theta}) = |2\pi\boldsymbol{V}|^{-1/2} exp\left\{-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu})'\boldsymbol{V}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})\right\},$$

e a função de verossimilhança para os dados completos é, descartando a constante de proporcionalidade,

$$L(\boldsymbol{\theta}|\boldsymbol{Y}) = |\boldsymbol{V}|^{-n/2} exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})'\boldsymbol{V}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})\right\}. \tag{5.1}$$

5.1.2 Inferência sob uma *Priori* Conjugada

A forma mais simples de se conduzir inferência Bayesiana no caso completo é aplicar uma família paramétrica ou classe de distribuições *a priori* que é *conjugada* da função de verossimilhança (5.1). Uma classe conjugada tem a propriedade que qualquer *priori*

$\pi(\boldsymbol{\theta})$ na classe leva a uma *posteriori* que também está na classe. Quando ambos $\boldsymbol{\mu}$ e $\boldsymbol{V}$ são desconhecidos, a classe conjugada mais natural para o modelo multivariado normal é a família normal Wishart invertida. Ou seja, $\pi(\boldsymbol{\mu}, \boldsymbol{V}) = \pi(\boldsymbol{\mu}|\boldsymbol{V})\pi(\boldsymbol{V})$ a distribuição *a priori* para $\boldsymbol{\theta}$, em que

$$\boldsymbol{\mu}|\boldsymbol{V} \sim N_m(\boldsymbol{\mu}_0, \tau^{-1}\boldsymbol{V}), \tag{5.2}$$

$$\boldsymbol{V} \sim W^{-1}(m, \boldsymbol{\Lambda}). \tag{5.3}$$

Neste caso, $\boldsymbol{V}$ segue uma distribuição Wishart invertida com parâmetros m e Λ, em que

$$P(\boldsymbol{V}|m, \boldsymbol{\Lambda}) \propto |\boldsymbol{V}|^{-(\frac{m+p+1}{2})} exp\left\{-\frac{1}{2} tr\boldsymbol{\Lambda}^{-1}\boldsymbol{V}^{-1}\right\}. \tag{5.4}$$

Detalhes da distribuição Wishart podem ser encontrados em Muirhead (1982).

A densidade *a priori* para $\boldsymbol{\theta}$ é então

$$\pi(\boldsymbol{\theta}) \propto |\boldsymbol{V}|^{-(\frac{m+p+1}{2})} exp\left\{-\frac{1}{2} tr\boldsymbol{\Lambda}^{-1}\boldsymbol{V}^{-1}\right\} \times exp\left\{-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\boldsymbol{V}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right\} \tag{5.5}$$

Multiplicando (5.1) por (5.5), pode-se mostrar que $P(\boldsymbol{\theta}|\boldsymbol{Y})$ tem a mesma forma de (5.5) mas com novos valores para $(\tau, m, \boldsymbol{\mu}_0, \boldsymbol{\Lambda})$; isto é, a *posteriori* de dados completos é normal Wishart invertida,

$$\boldsymbol{V}|\boldsymbol{Y} \sim W^{-1}(m', \boldsymbol{\Lambda}'), \tag{5.6}$$

$$\boldsymbol{\mu}|\boldsymbol{V}, \boldsymbol{Y} \sim N(\boldsymbol{\mu}_0', (\tau')^{-1}\boldsymbol{V}), \tag{5.7}$$

em que os hiperparâmetros atualizados são

$$\begin{aligned} \tau' &= \tau + n \\ m' &= m + n \\ \boldsymbol{\mu}_o' &= \left(\frac{n}{\tau + n}\right)\bar{\boldsymbol{y}} + \left(\frac{\tau}{\tau + n}\right)\boldsymbol{\mu}_0 \end{aligned}$$

e

$$\boldsymbol{\Lambda}' = \left[\boldsymbol{\Lambda}^{-1} + n\boldsymbol{S} + \left(\frac{\tau n}{\tau + n}\right)(\bar{\boldsymbol{y}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{y}} - \boldsymbol{\mu}_0)'\right]^{-1},$$

sendo $\boldsymbol{S} = \sum(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})'$.

Usualmente, utiliza-se a distribuição *a priori* imprópria

$$\pi(\boldsymbol{\theta}) \propto |\boldsymbol{V}|^{-\left(\frac{p+1}{2}\right)} \tag{5.8}$$

que pode ser vista como um caso limite da distribuição normal Wishart invertida (5.2)-(5.3) à medida que $\tau \to 0, m \to -1$ e $\boldsymbol{\Lambda}^{-1} \to 0$. Como $\boldsymbol{\mu}$ não aparece no lado direito de (5.8), a *'distribuição' a priori* para $\boldsymbol{\mu}$ é considerada uniforme sobre o espaço real p-dimensional.

Assumindo essa distribuição *a priori* não informativa, a distribuição *a posteriori* é dada por (Schafer, 1997a):

$$\boldsymbol{V}|\boldsymbol{Y} \sim W^{-1}\left(n-1, (n\boldsymbol{S})^{-1}\right), \tag{5.9}$$

$$\boldsymbol{\mu}|(\boldsymbol{V}, \boldsymbol{Y}) \sim N\left(\bar{\boldsymbol{y}}, n^{-1}\boldsymbol{V}\right). \tag{5.10}$$

A justificativa não Bayesiana para o uso dessa *priori* é que a distribuição *a posteriori* da quantidade pivotal

$$\boldsymbol{T}^2 = (n-1)(\bar{\boldsymbol{y}} - \boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{y}} - \boldsymbol{\mu})$$

é $(n-1)p(n-1)^{-1}F_{n,n-p}$, a mesma de sua distribuição amostral condicionada em $\boldsymbol{\theta}$ (DeGroot, 1970). A região elipsoide HPD $(1-\alpha)100\%$ para $\boldsymbol{\mu}$ sob essa priori é idêntica à região de confiança clássica $(1-\alpha)100\%$ para $\boldsymbol{\mu}$ da teoria amostral, e as inferências bayesianas e frequentistas sobre $\boldsymbol{\mu}$ coincidem. A *priori* imprópria (5.8) também surge aplicando o princípio da invariância de Jeffreys a $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$.

Adotando a regra de que uma estimativa razoável para $\boldsymbol{V} \sim W^{-1}(m, \boldsymbol{\Lambda})$ é $m^{-1}\boldsymbol{\Lambda}^{-1}$, então (5.9) leva à estimativa pontual $(n-1)^{-1}n\boldsymbol{S}$. Essa é a estimativa de $\boldsymbol{V}$ mais usada na prática, porque é não viciada para $\boldsymbol{\theta}$ sobre repetições do procedimento amostral. Por

essas razões aceitaremos (5.8) como uma distribuição a *priori* razoável para $\boldsymbol{\theta}$.

5.1.3 Imputação Utilizando o Modelo Normal

Em muitos problemas de dados incompletos, a distribuição *a posteriori* de dados observados, $P(\boldsymbol{\theta}|\boldsymbol{Y})$, é intratável e não pode ser facilmente simulada. Contudo, quando $\boldsymbol{Y}_{obs}$ é aumentado por um valor estimado/simulado do dado ausente $\boldsymbol{Y}_{mis}$, a distribuição *a posteriori* de dados completos $P(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$ fica mais fácil de ser simulada.

Aumentação de Dados (Tanner & Wong, 1987) para dados incompletos é um procedimento bastante similar ao algoritmo EM. Os passo E e M determinísticos são substituídos pelos passos I e P estocásticos, a serem descritos a seguir:

1. **Passo I – Imputação:** Dadas as estimativas do vetor de médias e da matriz de covariância, $\boldsymbol{\theta}^{(t)}$ no Passo t, o Passo I imputa os valores ausentes para cada observação independentemente. Isto é, denotando as variáveis com valores ausentes para a observação i por $\boldsymbol{Y}_{i(mis)}$ e as variáveis com valores observados por $\boldsymbol{Y}_{i(obs)}$, então o Passo I retira valores para $\boldsymbol{Y}_{i(mis)}$ da distribuição condicional de $\boldsymbol{Y}_{i(mis)}$ dado $\boldsymbol{Y}_{i(obs)}$.

$$\boldsymbol{Y}_{mis}^{(t+1)} \sim P(\boldsymbol{Y}_{mis}|\boldsymbol{y}_{obs}, \boldsymbol{\theta}^{(t)})$$

   Porque as linhas $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ de Y são condicionalmente independentes dado $\boldsymbol{\theta}$, o Passo I pode ser conduzido retirando

$$\boldsymbol{y}_{i(mis)}^{(t+1)} \sim P(\boldsymbol{y}_{i(mis)}|\boldsymbol{y}_{i(obs)}, \boldsymbol{\theta}^{(t)})$$

   independentemente para $i = 1, 2, \ldots, n$.

   Cálculo da distribuição condicional $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \boldsymbol{\theta})$

   Seja $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \boldsymbol{\mu}_2')'$ em que $\boldsymbol{\mu}_1$ é o vetor de médias para as variáveis $\boldsymbol{Y}_{obs}$, $\boldsymbol{\mu}_2$ é o

vetor de médias para as variáveis $\boldsymbol{Y}_{mis}$; e

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}'_{12} & \boldsymbol{V}_{22} \end{bmatrix}$$

a matriz de covariância para essas variáveis, em que $\boldsymbol{V}_{11}$ é a matriz de covariância para $\boldsymbol{Y}_{obs}$, $\boldsymbol{V}_{22}$ é a matriz de covariância para Y_{mis} e $\boldsymbol{V}_{12}$ é a matriz de covariância entre as variáveis $\boldsymbol{Y}_{obs}$ e $\boldsymbol{Y}_{mis}$.

Usando o operador *sweep* (Goodnight, 1979) nos pivôs da submatriz $\boldsymbol{V}_{11}$, a matriz se torna

$$\begin{bmatrix} \boldsymbol{V}_{11}^{-1} & \boldsymbol{V}_{11}^{-1}\boldsymbol{V}_{12} \\ -\boldsymbol{V}'_{12}\boldsymbol{V}_{11}^{-1} & \boldsymbol{V}_{22.1} \end{bmatrix}$$

em que $\boldsymbol{V}_{22.1} = \boldsymbol{V}_{22} - \boldsymbol{V}'_{12}\boldsymbol{V}_{11}^{-1}\boldsymbol{V}_{12}$.

A distribuição condicional de $\boldsymbol{Y}_{mis}$ dado $\boldsymbol{Y}_{obs} = \boldsymbol{y}_1$ e $\boldsymbol{\theta}$ é normal multivariada com vetor de médias

$$\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \boldsymbol{V}'_{12}\boldsymbol{V}_{11}^{-1}(\boldsymbol{y}_1 - \boldsymbol{\mu}_1) \tag{5.11}$$

e matriz de covariância condicional

$$\boldsymbol{V}_{22.1} = \boldsymbol{V}_{22} - \boldsymbol{V}'_{12}\boldsymbol{V}_{11}^{-1}\boldsymbol{V}_{12}. \tag{5.12}$$

2. **Passo P – Simulação da distribuição *a posteriori*:** Dado a amostra completa, o Passo P simula a média e o vetor de covariância da população *a posteriori*.

$$\boldsymbol{\theta}^{(t+1)} \sim P(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}^{(t+1)}).$$

Sob as distribuições *a priori* discutidas na Subseção 5.1.2 a *posteriori* de dados completos $P(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$ é uma distribuição normal Wishart invertida. O Passo

P, portanto, é meramente uma simulação da distribuição normal Wishart invertida,

$$\begin{aligned} \boldsymbol{V} &\sim W^{-1}(m, \boldsymbol{\Lambda}) \\ \boldsymbol{\mu}|\boldsymbol{V} &\sim N(\boldsymbol{\mu}_0, \tau^{-1}\boldsymbol{V}) \end{aligned}$$

para algum $(\tau, m, \boldsymbol{\mu}_0, \boldsymbol{\Lambda})$ determinados pela *priori*, os dados observados $\boldsymbol{Y}_{obs}$ e os dados ausentes $\boldsymbol{Y}_{mis}^{(t)}$ imputados no último Passo I. Os valores específicos de $(\tau, m, \boldsymbol{\mu}_0, \boldsymbol{\Lambda})$ são calculados usando as fórmulas de atualização dos hiperparâmetros dadas na Subseção 5.1.2.

Os dois passos são iterados até se obter resultados confiáveis para um conjunto de dados com imputação múltipla (Schafer, 1997a).

Isso cria uma Cadeia de Markov $(\boldsymbol{Y}_{miss}^{(1)}, \boldsymbol{\theta}^{(1)}), (\boldsymbol{Y}_{miss}^{(2)}, \boldsymbol{\theta}^{(2)}), \ldots, (\boldsymbol{Y}_{miss}^{(t)}, \boldsymbol{\theta}^{(t)}), \ldots$, que converge em distribuição para $P(\boldsymbol{Y}_{mis}, \boldsymbol{\theta}|\boldsymbol{Y}_{obs})$. Assumindo que as iterações convergem para uma distribuição estacionária, o objetivo é simular uma retirada aproximadamente independente dos valores ausentes dessa distribuição.

Após um período de aquecimento suficiente, valores sucessivos de $\boldsymbol{\theta}$,

$$\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t+2)}, \ldots,$$

constituem uma amostra dependente de $P(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})$. Iterações de $\boldsymbol{Y}_{mis}$ tomadas suficientemente distantes da sequência, digamos

$$\boldsymbol{Y}_{mis}^{(t)}, \boldsymbol{Y}_{mis}^{(t+c)}, \boldsymbol{Y}_{mis}^{(t+2c)}, \ldots$$

para algum valor grande de c, podem ser tomados como imputações de $\boldsymbol{Y}_{mis}$. Para obtermos k imputações de $\boldsymbol{Y}_{mis}$, devemos gerar k valores de $\boldsymbol{\theta}$ que sejam aproximadamente independentes, digamos

$$\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+c)}, \boldsymbol{\theta}^{(t+kc)}, \ldots,$$

e então retirar um valor de $\boldsymbol{Y}_{mis}$ dado cada um,

$$\begin{aligned}
\boldsymbol{Y}_{mis}^{(1)} &\sim P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \boldsymbol{\theta}^{(t)}) \\
\boldsymbol{Y}_{mis}^{(2)} &\sim P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \boldsymbol{\theta}^{(t+c)}) \\
&\vdots \\
\boldsymbol{Y}_{mis}^{(k)} &\sim P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \boldsymbol{\theta}^{(t+kc)}).
\end{aligned}$$

Para se obter valores independentes de $\boldsymbol{\theta}$ não é preciso necessariamente subamostrar a cada c-ésimo valor de uma única cadeia; podemos também rodar k cadeias independentes de comprimento c de um valor inicial comum, ou, melhor ainda, de k valores iniciais independentes.

5.2 IMPUTAÇÃO PELO MODELO MISTO

Nesta seção apresentamos dois métodos para imputação múltipla considerando o modelo linear misto, como apresentado na Seção 2.1. O primeiro método utiliza um enfoque bayesiano e o segundo o clássico.

5.2.1 Enfoque Bayesiano

A imputação com o modelo linear misto na versão multivariada sob enfoque bayesiano é discutida em detalhes em Schafer & Yucel (2002). A seguir mostramos o algoritmo para o caso univariado, que é de nosso interesse.

5.2.1.1 Algoritmo de Imputação via Amostrador de Gibbs

Considere um algoritmo de simulação iterativo no qual versões atualizadas dos parâmetros desconhecidos $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\beta}^{(t)}, \boldsymbol{\sigma}^{2(t)}, \boldsymbol{B}^{(t)})$ e os dados ausentes $\boldsymbol{Y}_{mis}^{(t)}$ são atualizados em três passos. Inicialmente,

$$\boldsymbol{b}_i^{(t+1)} \sim P(\boldsymbol{b}_i|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}^{(t)}, \boldsymbol{\theta}^{(t)}) \tag{5.13}$$

independentemente para $i = 1, 2, \ldots, n$. A seguir fazemos

$$\boldsymbol{\theta}^{(t+1)} \sim P(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}^{(t)}, \boldsymbol{b}^{(t+1)}), \tag{5.14}$$

e finalmente,

$$\boldsymbol{y}_{i(mis)}^{(t+1)} \sim P(\boldsymbol{y}_{i(mis)}|\boldsymbol{Y}_{obs}, \boldsymbol{b}^{(t+1)}, \boldsymbol{\theta}^{(t+1)}), \tag{5.15}$$

para $i = 1, \ldots, n$. Dados valores iniciais $\boldsymbol{\theta}^{(0)}$ e $\boldsymbol{Y}_{mis}^{(0)}$, esses passos definem um ciclo de um procedimento MCMC chamado Amostrador de Gibbs. Executar o ciclo repetidamente cria sequências $\left\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots\right\}$ e $\left\{\boldsymbol{Y}_{mis}^{(1)}, \boldsymbol{Y}_{mis}^{(2)}, \ldots\right\}$ cujas distribuições limites são $P(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})$ e $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs})$, respectivamente.

Como distribuições *a priori* consideramos $\sigma^{-2} \sim W(\nu_1, \boldsymbol{\Lambda}_1)$ e $\boldsymbol{\Psi}^{-1} \sim W(\nu_2, \boldsymbol{\Lambda}_2)$, em que $W(\nu, \boldsymbol{\Lambda})$ denota uma variável Wishart com $\nu > 0$ graus de liberdade e média $\nu\boldsymbol{\Lambda} > 0$. Para $\boldsymbol{\beta}$ usa-se uma "densidade" uniforme imprópria sobre R^p. Uma descrição detalhada do algoritmo pode ser encontrada em Schafer (1997b).

5.2.2 Enfoque Frequentista

Após o ajuste do modelo (2.1) aos dados disponíveis e obtenção das estimativas de máxima verossimilhanca dos parâmetros, o procedimento adotado para imputação é apresentado a seguir (seguindo Rubin (1987), pp. 167).

5.2.2.1 Algoritmo para Imputação

1. Gerar uma variável $\chi^2_{df^*}$, digamos g, e fazer

$$\hat{\sigma}^{2*} = \hat{\sigma}^2 \times df^*/g.$$

Aqui df^* foi considerado igual ao número de graus de liberdade resultante da estimação por máxima verossimilhanca restrita, ou seja $\sum_{i=1}^{n} n_i - p^*$, em que p^* é o número de parâmetros estimados pelo modelo, ou seja, $\boldsymbol{\beta}$ e os componentes de variância.

2. Utilizando o fato que $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}^{-1}))$, gerar p desvios independentes $N(0,1)$ para formar o vetor $\boldsymbol{W}_1$ de dimensão p e calcular

$$\widehat{\boldsymbol{\beta}}^* = \widehat{\boldsymbol{\beta}} + \sigma^*(\boldsymbol{X}'\widehat{\boldsymbol{V}}\boldsymbol{X})^{-1/2}\boldsymbol{W}_1,$$

3. Gerar q desvios independentes $N(0,1)$ para formar o vetor $\boldsymbol{W}_2$ de dimensão q e calcular

$$\hat{\boldsymbol{b}}_i^* = \hat{\boldsymbol{b}}_i + \widehat{\boldsymbol{B}}^{-1/2}\boldsymbol{W}_2,$$

em que $\hat{\boldsymbol{b}}_i$ é o estimador BLUP de $\boldsymbol{b}_i$ (Fitzmaurice *et al.*, 2004) e $\hat{\boldsymbol{b}}_i \sim N(\boldsymbol{0}, \hat{\boldsymbol{B}})$.

4. Imputar cada valor ausente por

$$Y_{ij,mis}^* = \boldsymbol{X}_{ij}\widehat{\boldsymbol{\beta}}^* + \mathbf{Z}_i\hat{\mathbf{b}}_i^* + \hat{\sigma}^* Z_i,$$

sendo Z_i uma variável aleatória normal padrão.

Para imputação múltipla, os passos 1 a 4 foram repetidos k vezes independentemente para a criação de k bancos de dados completados.

5.3 IMPUTAÇÃO POR PAREAMENTO

Nesta seção apresentamos dois métodos usados para realizar imputação múltipla que consideram alguma forma de pareamento.

5.3.1 Imputação pelo Escore de Propensão

5.3.1.1 Motivação Histórica

No contexto de estudos observacionais, Rosenbaum & Rubin (1983) propuseram o escore de propensão como método de pareamento (ou estratificação) para redução do viés na estimação de efeitos de tratamento.

O pareamento (ou estratificação) pelo escore de propensão é uma forma de "corrigir"a estimação dos efeitos do tratamento controlando os fatores de confundimento, com base na ideia que o viés é reduzido quando a comparação dos resultados é realizada usando grupos tratados e controles que sejam tão similares quanto possível. Como o pareamento (ou estratificação) é impraticável num vetor p-dimensional de características, para um p grande, esse método propõe sumarizar as características pré-tratamento numa única variável (o escore de propensão) que torna o procedimento possível.

5.3.1.2 Definição

O escore de propensão, $p(\mathbf{X})$, foi definido originalmente como a probabilidade condicional de o indivíduo estar no grupo tratamento, dado o conjunto de covariáveis medidas pré-tratamento (Rosenbaum & Rubin, 1983):

$$p(\mathbf{X}) \equiv Pr(A = 1|\mathbf{X}) = E(A|\mathbf{X}),$$

em que $A = \{0, 1\}$ é o indicador de exposição ao tratamento e $\mathbf{X}$ é o vetor multidimensional de características pré-tratamento.

O escore de propensão reduz toda a informação do vetor $\mathbf{X}$ p-dimensional à apenas uma covariável.

5.3.1.3 Propriedades do Escore de Propensão

Sejam Y_1 e Y_0 as respostas potenciais nas situações de tratamento e não tratamento.

Proposição 1 (Propriedade de Balanceamento)

$$Pr(\mathbf{X}|A = 1, p(\mathbf{X})) = Pr(\mathbf{X}|A = 0, p(\mathbf{X}))$$

ou

$$A \perp \mathbf{X}|p(\mathbf{X}).$$

Proposição 2 (Não Confundimento Dado o Escore de Propensão) *Suponha que a atribuição ao tratamento é não confundida, isto é*

$$Y_1, Y_0 \perp A|\mathbf{X}.$$

Então a atribuição ao tratamento é não confundida dado o escore de propensão, isto é:

$$Y_1, Y_0 \perp A|p(\mathbf{X}).$$

A propriedade de balanceamento diz que as unidades tratadas, $A = 1$, e controles, $A = 0$, com o mesmo escore de propensão $p(\mathbf{X})$ tem a mesma distribuição das covariáveis observadas $\mathbf{X}$.

A propriedade de não confundimento estabelece que, se condicionado em $\mathbf{X}$ a atribuição ao tratamento é "puramente aleatória", será suficiente condicionar em $p(\mathbf{X})$ para obter estimativas não viesadas do efeito do tratamento. Essa condição é conhecida como "atribuição ao tratamento fortemente ignorável".

5.3.1.4 O Escore de Propensão na Imputação de Dados

Lavori *et al.* (1995) propuseram o uso de escore de propensão como método de imputação de dados em estudos longitudinais, especialmente quando há *dropout* (perda completa da informação para o indivíduo a partir de certo período). Esse método visa preencher cada valor ausente com um valor observado de um indivíduo que tenha a mesma probabilidade de apresentar aquele dado ausente.

Como vimos anteriormente, indivíduos com o mesmo escore de propensão têm, em média, a mesma distribuição de covariáveis. Assim, se todos os fatores que influenciam a probabilidade de um indivíduo apresentar dado ausente forem medidos (situação MAR), imputar uma resposta de um indivíduo com o mesmo escore de propensão significa que a distribuição dos valores das covariáveis medidas é a mesma, o que implica em imputação não viesada, pelos resultados das Proposições (1) e (2).

5.3.1.5 Algoritmo para Imputação

Seja D a indicadora para a resposta, com $D = 1$ se $\boldsymbol{Y}$ é observado e $D = 0$ se $\boldsymbol{Y}$ não é observado. Os índices indicam o tempo no qual as respostas são avaliadas. Considere, por exemplo, que haja *dropout*, no tempo $j = 3$. A propensão de o indivíduo permanecer no estudo nesse momento é dada por:

$$p(\mathbf{X}, \boldsymbol{Y}_{1,obs}, \boldsymbol{Y}_{2,obs}) = Pr(D_3 = 0|\mathbf{X}, \boldsymbol{Y}_{1,obs}, \boldsymbol{Y}_{2,obs}).$$

As covariáveis incluídas em $\mathbf{X}$ são aquelas que fazem plausível a suposição de que a não resposta seja *ignorável* (mecanismo MAR) condicionado em seus valores observados, isto é, que a não resposta depende apenas dos resultados nos tempos anteriores a j.

Dada a estratificação baseada nos quintis de $\hat{p}(\mathbf{X}, \boldsymbol{Y}_{1,obs}, \boldsymbol{Y}_{2,obs})$, são feitas imputações "próprias" (Little & Rubin, 1987) com base em um *bootstrap Bayesiano aproximado (ABB)*. Primeiramente é retirado aleatoriamente um conjunto 'potencial' de respostas observadas, com reposição, dentre as respostas observadas no quintil definido pelo escore de propensão estimado, e então são usados para imputação valores escolhidos (aleatoriamente) dentre aqueles observados dessa amostra 'potencial'. Isso assegura a variabilidade entre imputações, já que o passo do ABB é o análogo não paramétrico de retirar parâmetros da distribuição preditiva posteriori dos dados ausentes (Lavori *et al.*, 1995).

O procedimento para imputação, realizado para cada tempo j, é dado pelos seguintes passos.

1. Crie uma variável D_i com o valor 0 para observações ausentes e 1, caso contrário;

2. Ajuste um modelo logístico, da forma:

$$\log\left(\frac{p(\mathbf{X}_{ij})}{1 - p(\mathbf{X}_{ij})}\right) = \beta_0 + \beta_1 y_{i1} + \cdots + \beta_{j-1} y_{i,j-1} + \boldsymbol{\gamma}_j \mathbf{x}_i$$

   em que $\beta_0, \beta_1, \ldots, \beta_{j-1}, \boldsymbol{\gamma}_j$ são os coeficientes de regressão. Aqui, $\beta_1, \beta_2, \ldots, \beta_{j-1}$ estão associados com as respostas nos $j-1$ tempos anteriores, e $\boldsymbol{\gamma}_j \mathbf{x}_i = \gamma_1 x_{i1} + \gamma_p x_{ip}$

são os efeitos das p covariáveis (completamente observadas) do i-ésimo indivíduo.

3. A cada observação é atribuído um escore de propensão estimado:

$$\hat{p}(\mathbf{X}_{ij}) = \frac{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 y_{i1} + \ldots + \hat{\beta}_{j-1} y_{i,j-1} + \hat{\boldsymbol{\gamma}}_j \mathbf{x}_i\right\}}{1 + \exp\left\{\hat{\beta}_0 + \hat{\beta}_1 y_{i1} + \ldots + \hat{\beta}_{j-1} y_{i,j-1} + \hat{\boldsymbol{\gamma}}_j \mathbf{x}_i\right\}}.$$

   Ordene as observações com base no escore de propensão.

4. Divida as observações num número fixo de grupos com base no escore de propensão estimado. São geralmente considerados $q = 5$ grupos, ou seja, as observações são agrupadas pelos quintis do escore de propensão estimado.

5. Dentro de cada quintil é aplicado um bootstrap Bayesiano aproximado (ABB). No quintil q, seja $\boldsymbol{Y}_{obs}$ as n_1 observações com valores $\boldsymbol{Y}_j$ não ausentes ($D = 1$) e $\boldsymbol{Y}_{mis}$ as n_0 observações com valores $\boldsymbol{Y}_j$ ausentes ($D = 0$). A imputação ABB consiste em retirar n_1 observações aleatoriamente com reposição de $\boldsymbol{Y}_{obs}$ para criar um novo conjunto de dados $\boldsymbol{Y}^*_{obs}$. O processo então retira aleatoriamente n_0 valores para $\boldsymbol{Y}_{mis}$ com reposição de $\boldsymbol{Y}^*_{obs}$.

Esse procedimento é repetido sequencialmente para cada variável com valores ausentes até o conjunto de dados ser completamente preenchido; múltiplas imputações são obtidas pela repetição desses passos k vezes.

5.3.2 Imputação por Pareamento Genético

Ainda no contexto de avaliação de efeito de tratamento em estudos observacionais, Sekhon (2007) propõe um algoritmo genético de pareamento que busca o balanceamento ótimo entre as distribuições das covariáveis para unidades tratadas e controles.

O pareamento genético é um método que automaticamente encontra o conjunto de pares que minimiza a discrepância entre a distribuição de potenciais confundidores entre os diferentes grupos, assim o balanceamento das covariáveis é maximizado.

O balanceamento ótimo é alcançado por pareamento multivariado no qual um algoritmo genético de busca determina o peso que é dado a cada covariável. Esse balancea-

mento é feito pelo exame de funções distribuição de probabilidade acumuladas com base em estatísticas padronizadas. Por padrão, essas estatísticas incluem testes t pareados e testes de Kolmogorov-Smirnov (KS). A maximização do balanceamento pode ser conduzida de acordo com várias funções de perda, como valores p dos testes de hipóteses. Também é possível usar estatísticas descritivas baseadas em QQ-plots empíricos.

O método é uma generalização da distância de Mahalanobis, muito usada em métodos de pareamento.

A distância de Mahalanobis entre dois vetores quaisquer é:

$$md(\boldsymbol{X}_{ij}, \boldsymbol{X}_{i'j}) = \left\{(\boldsymbol{X}_{ij} - \boldsymbol{X}_{i'j})'\boldsymbol{S}^{-1}(\boldsymbol{X}_{ij} - \boldsymbol{X}_{i'j})\right\}^{1/2},$$

em que $\boldsymbol{S}$ é a matriz de covariância de $\boldsymbol{X}$.

Pode-se combinar o escore de propensão com a distância de Mahalanobis, incluindo-o como uma covariável em $\boldsymbol{X}$. O pareamento (ou estratificação) pelo escore de propensão é particularmente bom em minimizar a discrepância em relação ao escore de propensão e a distância de Mahalanobis é particularmente boa em minimizar a distância entre as coordenadas individuais de $\boldsymbol{X}$ (ortogonais ao escore de propensão) (Rosenbaum & Rubin, 1985).

5.3.2.1 O Pareamento Genético

A ideia por trás do algoritmo é que se a distância de Mahalanobis não é ótima para produzir balanceamento num dado conjunto de dados, deve-se procurar no espaço de métricas de distância algo melhor. Uma forma de generalizar a métrica de Mahalanobis é incluir uma matriz de peso adicional:

$$d(\boldsymbol{X}_{ij}, \boldsymbol{X}_{i'j}) = \left\{(\boldsymbol{X}_{ij} - \boldsymbol{X}_{i'j})'(\boldsymbol{S}^{-1/2})'\boldsymbol{W}\boldsymbol{S}^{-1/2}(\boldsymbol{X}_{ij} - \boldsymbol{X}_{i'j})\right\}^{1/2}, \tag{5.16}$$

em que $\boldsymbol{W}$ é uma matriz $w \times w$ de pesos positiva definida e $\boldsymbol{S}^{1/2}$ é a decomposição de Cholesky de $\boldsymbol{S}$.

O algoritmo genético usa a distância $d()$ em (5.16) como medida de pareamento. A

matriz $\boldsymbol{W}$ é escolhida de forma a minimizar a maior discrepância de covariáveis entre os grupos pareados. Detalhes do algoritmo de pareamento genético são encontrados em (Sekhon & Mebane, 1998) e o método está implementado no software *R*, pacote *Matching*.

5.3.2.2 Algoritmo para Imputação

Para imputação por pareamento utilizando o algoritmo genético, o procedimento proposto é semelhante àquele adotado com o escore de propensão: para cada tempo é aplicado o algoritmo genético para encontrar de forma ótima a matriz de pesos em (5.16). Para cada indivíduo com resposta ausente é usada como imputação a resposta observada de seu respectivo par. O processo é repetido até que o banco de dados seja completamente preenchido.

5.4 Comentários Gerais sobre os Métodos de Imputação

A seguir alguns comentários gerais sobre os métodos de imputação discutidos nesse capítulo.

1. Para imputação considerando o Modelo Normal (Bayesiano), Pareamento Genético e Escore de Propensão existe a restrição de que os dados sejam balanceados. Ou seja, esses métodos exigem que as medidas sejam tomadas nas mesmas ocasiões para todos os indivíduos;

2. Na imputação com o Modelo Misto tem-se que estimar tal modelo quando o de interesse é o GEE. A justificativa recai na incerteza da modelagem da estrutura de covariância para os dados e na flexibilidade/robustez da imputação mesmo na presença de um modelo mal ajustado;

3. Os dois métodos de pareamento (Pareamento Genético e Escore de Propensão) ignoram a relação entre a resposta $\boldsymbol{Y}$ e as covariáveis $\boldsymbol{X}$.

CAPÍTULO 6

SIMULAÇÕES DE MONTE CARLO

6.1 Geração dos Cenários de Interesse

Neste capítulo apresentamos resultados de simulação de Monte Carlo para exemplificar o impacto dos dados ausentes na estimação dos parâmetros de um modelo GEE. A geração dos dados e da não resposta foram adaptadas de Hedeker & Gibbons (2006).

6.1.1 Geração do Banco de Dados

Consideramos duas formas de geração dos dados. Na primeira os dados são heterocedásticos e na segunda são homocedásticos. A distinção entre uma situação e outra é que na segunda a variância não é constante ao longo do tempo enquanto na primeira foi fixada a mesma variabilidade para cada período de tempo.

Para essas situações os dados foram gerados através do seguinte modelo

$$Y_{ij} = \beta_0 + \beta_1 Tempo_j + \beta_2 Grupo_i + \beta_3 (Grupo_i \times Tempo_j) + b_{0i} + b_{1i} Tempo_j + \varepsilon_{ij}, \quad (6.1)$$

para o caso heterocedástico; e

$$Y_{ij} = \beta_0 + \beta_1 Tempo_j + \beta_2 Grupo_i + \beta_3 (Grupo_i \times Tempo_j) + b_{0i} + \varepsilon_{ij}, \quad (6.2)$$

para o caso homocedástico. Em ambos, $Tempo_j$ foi codificado como $0, 1, 2, 3, 4$ para cinco medidas repetidas por indivíduo e $Grupo_i$ foi uma variável indicadora (isto é, 0 ou 1) com metade dos indivíduos em cada grupo. Os coeficientes de regressão foram fixados em: $\beta_0 = 25$, $\beta_1 = -1$, $\beta_2 = 0$ e $\beta_3 = -1$. Assim, as médias populacionais para os dois grupos nas cinco medidas repetidas, para os dois casos, foram:

- *Grupo 0*: 25, 24, 23, 22, 21; e

- *Grupo 1*: 25, 23, 21, 19, 17.

Os efeitos aleatórios b_{0i} e b_{1i} são normalmente distribuídos com média zero, variâncias $\sigma^2_{b_0} = 4$, $\sigma^2_{b_1} = 0,25$ e covariância $\sigma_{b_{01}} = -0,1$ (também igual a $-0,1$, expressos como correlação). Os erros ε_i foram gerados de uma distribuição normal com média zero e variância $\sigma^2 = 4$. A matriz de variância-covariância, $V(\boldsymbol{y}) = \boldsymbol{ZBZ}' + \sigma^2\boldsymbol{I}$, para o caso heterocedástico, foi então

$$V(\boldsymbol{y}) = \begin{bmatrix} 8,00 & 3,90 & 3,80 & 3,70 & 3,60 \\ 3,90 & 8,05 & 4,20 & 4,35 & 4,50 \\ 3,80 & 4,20 & 8,60 & 5,00 & 5,40 \\ 3,70 & 4,35 & 5,00 & 9,65 & 6,30 \\ 3,60 & 4,50 & 5,40 & 6,30 & 11,20 \end{bmatrix} \text{ ou } Cor(\boldsymbol{y}) = \begin{bmatrix} 1,00 & 0,49 & 0,46 & 0,42 & 0,38 \\ 0,49 & 1,00 & 0,50 & 0,49 & 0,47 \\ 0,46 & 0,50 & 1,00 & 0,55 & 0,55 \\ 0,42 & 0,49 & 0,55 & 1,00 & 0,61 \\ 0,38 & 0,47 & 0,55 & 0,61 & 1,00 \end{bmatrix};$$

enquanto, para o caso homocedástico, assumiu a forma

$$V(\boldsymbol{y}) = \begin{bmatrix} 8,00 & 4,00 & 4,00 & 4,00 & 4,00 \\ 4,00 & 8,00 & 4,00 & 4,00 & 4,00 \\ 4,00 & 4,00 & 8,00 & 4,00 & 4,00 \\ 4,00 & 4,00 & 4,00 & 8,00 & 4,00 \\ 4,00 & 4,00 & 4,00 & 4,00 & 8,00 \end{bmatrix} \text{ ou } Cor(\boldsymbol{y}) = \begin{bmatrix} 1,00 & 0,50 & 0,50 & 0,50 & 0,50 \\ 0,50 & 1,00 & 0,50 & 0,50 & 0,50 \\ 0,50 & 0,50 & 1,00 & 0,50 & 0,50 \\ 0,50 & 0,50 & 0,50 & 1,00 & 0,50 \\ 0,50 & 0,50 & 0,50 & 0,50 & 1,00 \end{bmatrix}.$$

Consideremos valores menores da resposta como *melhora*. Então, basicamente estamos comparando dois grupos que melhoram ao longo do tempo, embora um grupo ($Grupo = 1$) tenha uma taxa de melhora mais acentuada.

6.1.2 Geração da Não Resposta

Criamos um cenário em que a perda dos dados é MAR, em que a probabilidade da não resposta em cada tempo dependia apenas da resposta observada no tempo anterior. O esquema de geração da não resposta (a partir do segundo tempo) foi:

- **MAR**: Se o valor da variável dependente foi menor que 23, então o indivíduo saía da estudo no próximo período de tempo (isto é, temos dados ausentes para o próximo

e os subsequentes períodos) com probabilidade de 80%.

Esses valores foram escolhidos de forma a produzir, em média, 42% de dados ausentes. Nas seções a seguir são mostrados os ajustes de modelos GEE com quatro matrizes de trabalho: independente (IN), simetria composta (SC), não estruturada (NE) e auto regressiva de ordem 1 (AR). Os erros padrões apresentados neste trabalho são o da versão robusta.

6.2 Resultados MAR, Caso Heterocedástico

Nessa situação os dados foram gerados segundo o modelo (6.1), ou seja, o modelo correto para análise deveria incluir uma estrutura de covariância não constante.

Para cada um dos cinco métodos considerados são apresentados os resultados da imputação. Os valores mostrados são as médias de 5.000 repetições do processo de geração e perda de dados segundo o mecanismo MAR, descrito anteriormente. No corpo das tabelas a seguir os valores estão na forma: estimativa (erro padrão). Para o caso da imputação com o escore de propensão e matriz de trabalho GEE não estruturada houve problemas computacionais no ajuste, gerando resultados suspeitos; neste caso mostramos as medianas das estimativas dos coeficientes, assim como de seus respectivos erros padrões. A simulação foi repetida independentemente para cada método, e o nosso objetivo foi comparar como cada um deles se comporta, em média, nas 5.000 repetições do processo. O tamanho de cada banco criado foi de $n = 500$, totalizando 2.500 observações.

Consideramos imputação múltipla com $k = 5$ bancos de dados imputados. Os erros padrões finais apresentados são os combinados pela regra de Rubin.

6.2.1 Modelo Normal (Bayesiano)

Na Tabela 6.1 estão os resultados de simulação considerando a imputação pelo modelo normal bayesiano. Como distribuição *a priori* para o modelo normal foi usada a não informativa discutida na Seção 5.1.2; esta é a distribuição *a priori default* no pacote *norm*, software *R*. Foram observados 41,72% de dados ausentes.

Tabela 6.1: Imputação pelo Modelo Normal (Bayesiano), n=500

		β_0 (i)	β_1 (t)	β_2 (g)	β_3 $(g \times t)$
	Simulado	25	-1	0	-1
COMP	GEE-IN	25,001 (0,160)	-1,001 (0,051)	-0,002 (0,226)	-0,999 (0,072)
	GEE-SC	25,001 (0,160)	-1,001 (0,051)	-0,002 (0,226)	-0,999 (0,072)
	GEE-NE	25,001 (0,160)	-1,001 (0,051)	-0,003 (0,227)	-0,999 (0,074)
	GEE-AR	25,001 (0,166)	-1,001 (0,053)	-0,003 (0,236)	-0,999 (0,076)
MAR	GEE-IN	24,928 (0,162)	-0,455 (0,080)	-0,042 (0,230)	-0,884 (0,133)
	GEE-SC	24,934 (0,166)	-0,970 (0,074)	0,010 (0,237)	-1,015 (0,119)
	GEE-NE	24,902 (0,164)	-0,635 (0,076)	-0,008 (0,233)	-0,957 (0,124)
	GEE-AR	24,984 (0,175)	-1,216 (0,083)	0,007 (0,248)	-1,083 (0,128)
IMP	GEE-IN	24,986 (0,160)	-0,986 (0,050)	-0,007 (0,226)	-0,992 (0,071)
	GEE-SC	24,986 (0,160)	-0,986 (0,050)	-0,007 (0,226)	-0,992 (0,071)
	GEE-NE	25,069 (0,161)	-1,009 (0,053)	0,012 (0,229)	-0,998 (0,075)
	GEE-AR	24,994 (0,166)	-0,989 (0,053)	-0,006 (0,236)	-0,982 (0,074)

O método de imputação que considera o modelo normal para a variável dependente teve um desempenho de modo geral bastante satisfatório. Os coeficientes estimados agora estão muito próximos daqueles gerados para todas as matrizes de trabalho consideradas.

Em termos médios, o erro padrão de alguns coeficientes pós imputação foi menor que o de dados completos. Esse é um comportamento não esperado, pois a variância dos estimadores advinda dos dados imputados é a da média das variâncias dos dados completados (imputados) acrescida da variância entre as cinco imputações. Uma explicação é o fato de o método de imputação produzir em alguns momentos bancos de dados com estimativas tendo variabilidade menor que aquelas de dados completos e a variabilidade entre as estimativas dos bancos imputadas ser pequena.

6.2.2 Modelo Misto (Bayesiano)

Na Tabela 6.2 estão os resultados de simulação considerando a imputação pelo modelo misto com enfoque bayesiano. Foram usadas como distribuições *a priori* para a matriz de covariância dos efeitos aleatórios uma distribuição Wishart invertida com parâmetros $\nu_2 = 1$ e $\mathbf{\Lambda}$ uma matriz identidade de dimensão 2×2; para erro residual foi adotada uma distribuição Wishart invertida com ambos parâmetros iguais a 1; para o vetor de parâmetros de regressão usou-se a *priori* uniforme na reta. A escolha dessas distribuições *a priori* foi feita de forma a minimizar sua influência nos dados. Foram observados 41,75% de dados ausentes.

Tabela 6.2: Imputação pelo Modelo Misto (Bayesiano), n=500

		β_0 (i)	β_1 (t)	β_2 (g)	β_3 $(g \times t)$
	Simulado	25	-1	0	-1
COMP	GEE-IN	25,000 (0,160)	-1,000 (0,051)	0,001 (0,226)	-1,000 (0,072)
	GEE-SC	25,000 (0,160)	-1,000 (0,051)	0,001 (0,226)	-1,000 (0,072)
	GEE-NE	25,000 (0,161)	-1,000 (0,052)	0,000 (0,227)	-1,000 (0,074)
	GEE-AR	24,999 (0,167)	-1,000 (0,054)	0,001 (0,236)	-1,000 (0,076)
MAR	GEE-IN	24,926 (0,162)	-0,454 (0,080)	-0,038 (0,230)	-0,886 (0,133)
	GEE-SC	24,931 (0,166)	-0,969 (0,074)	0,015 (0,237)	-1,018 (0,119)
	GEE-NE	24,899 (0,164)	-0,634 (0,076)	-0,004 (0,234)	-0,961 (0,125)
	GEE-AR	24,982 (0,175)	-1,214 (0,083)	0,013 (0,248)	-1,087 (0,128)
IMP	GEE-IN	25,003 (0,160)	-1,006 (0,052)	0,001 (0,226)	-1,003 (0,073)
	GEE-SC	25,003 (0,160)	-1,006 (0,052)	0,001 (0,226)	-1,003 (0,073)
	GEE-NE	25,008 (0,164)	-1,008 (0,056)	0,000 (0,234)	-1,002 (0,079)
	GEE-AR	25,001 (0,167)	-1,005 (0,054)	0,001 (0,236)	-1,003 (0,077)

As imputações pelo modelo misto com enfoque bayesiano removeram completamente o vício (diferença entre os valores gerados e estimados) das estimativas entre as diversas matrizes de trabalho adotadas aqui. Após a imputação as estimativas dos coeficientes de regressão para essas diferentes estruturas de correlação tornaram-se bastante semelhantes. Além disso, as estimativas de variabilidade combinadas obtidas com os dados imputados são agora muito próximas daquelas obtidas com os dados completos.

6.2.3 Modelo Misto (Frequentista)

Na Tabela 6.3 estão os resultados de imputação com o modelo misto frequentista. O percentual médio de dados ausentes foi de 41,73%.

Tabela 6.3: Imputação pelo Modelo Misto (Frequentista), n=500

		β_0 (i)	β_1 (t)	β_2 (g)	β_3 $(g \times t)$
	Simulado	25	-1	0	-1
COMP	GEE-IN	25,002 (0,159)	-1,001 (0,051)	0,004 (0,226)	-0,998 (0,072)
	GEE-SC	25,002 (0,159)	-1,001 (0,051)	0,004 (0,226)	-0,998 (0,072)
	GEE-NE	25,002 (0,160)	-1,001 (0,052)	0,004 (0,227)	-0,998 (0,074)
	GEE-AR	25,003 (0,166)	-1,001 (0,053)	0,006 (0,236)	-0,997 (0,076)
MAR	GEE-IN	24,929 (0,162)	-0,457 (0,080)	-0,044 (0,230)	-0,881 (0,133)
	GEE-SC	24,934 (0,166)	-0,972 (0,074)	0,009 (0,237)	-1,013 (0,119)
	GEE-NE	24,903 (0,164)	-0,638 (0,076)	-0,010 (0,234)	-0,954 (0,125)
	GEE-AR	24,986 (0,174)	-1,219 (0,083)	-0,004 (0,248)	-1,079 (0,128)
IMP	GEE-IN	25,003 (0,162)	-1,005 (0,067)	-0,004 (0,230)	-0,998 (0,097)
	GEE-SC	25,003 (0,162)	-1,005 (0,067)	-0,004 (0,230)	-0,998 (0,097)
	GEE-NE*	25,005 (0,305)	-0,998 (0,185)	-0,017 (0,445)	-0,996 (0,269)
	GEE-AR	25,003 (0,166)	-1,005 (0,068)	-0,006 (0,236)	-0,998 (0,099)

* Por problemas de convergência são apresentadas os valores medianos

Após a imputação pelo modelo misto o vício das estimativas GEE foi totalmente removido. Todos os coeficientes agora são muito próximos dos reais valores gerados, preservando todos os efeitos, inclusive o de interação.

A matriz de correlação de trabalho não estruturada apresentou problemas em boa parte das repetições, resultando em estimativas dos coeficientes e erros padrões irreais. Por isso são mostrados os valores medianos entre as repetições.

6.2.4 Escore de Propensão

Para imputação com o escore de propensão o procedimento adotado foi separado por grupo. Assim, foi feita a imputação para os dois grupos de forma independente, sendo incluídos no modelo logístico as variáveis observadas nos tempos anteriores (incluindo as imputadas). Foi observado 41,74% de dados ausentes.

Tabela 6.4: Imputação pelo Escore de Propensão, n=500

		β_0 (i)	β_1 (t)	β_2 (g)	β_3 $(g \times t)$
	Simulado	25	-1	0	-1
COMP	GEE-IN	25,000 (0,160)	-1,000 (0,051)	-0,001 (0,226)	-1,000 (0,072)
	GEE-SC	25,000 (0,160)	-1,000 (0,051)	-0,001 (0,226)	-1,000 (0,072)
	GEE-NE	25,001 (0,160)	-1,000 (0,052)	-0,001 (0,227)	-1,000 (0,074)
	GEE-AR	25,000 (0,166)	-1,000 (0,053)	0,001 (0,235)	-1,000 (0,076)
MAR	GEE-IN	24,927 (0,162)	-0,455 (0,080)	-0,040 (0,230)	-0,885 (0,133)
	GEE-SC	24,932 (0,166)	-0,969 (0,074)	0,013 (0,237)	-1,016 (0,119)
	GEE-NE	24,900 (0,164)	-0,635 (0,076)	-0,005 (0,233)	-0,959 (0,124)
	GEE-AR	24,983 (0,175)	-1,214 (0,083)	0,011 (0,248)	-1,085 (0,128)
IMP	GEE-IN	24,943 (0,160)	-0,907 (0,048)	-0,042 (0,227)	-0,949 (0,069)
	GEE-SC	24,943 (0,160)	-0,907 (0,048)	-0,042 (0,227)	-0,949 (0,069)
	GEE-NE*	25,100 (0,159)	-0,941 (0,050)	0,006 (0,225)	-0,960 (0,070)
	GEE-AR	24,958 (0,165)	-0,904 (0,050)	-0,033 (0,234)	-0,945 (0,072)

* Por problemas de convergência são apresentadas os valores medianos

A imputação pelo escore de propensão apresentou resultados viesados para o efeito de tempo, em torno de 10%, em média. Para os outros coeficientes de regressão o vício foi menor. De forma bastante geral, em média, a imputação trouxe ganhos à análise, já que as estimativas com os dados incompletos são bastante viesadas, exceto para a matriz de trabalho simetria composta. Observe, neste caso, que a imputação pelo escore de propensão aumenta o vício de $\hat{\beta}_1$.

6.2.5 Pareamento

Para a imputação pelo método de pareamento com o algoritmo genético, o procedimento de imputação foi separado por grupo. Assim, unidades dentro do primeiro grupo foram pareadas com unidades dentro do seu próprio grupo, o mesmo acontecendo com o segundo grupo. Foi observado 41,77% de dados ausentes.

Tabela 6.5: Imputação por Pareamento, n=500

		β_0 (i)	β_1 (t)	β_2 (g)	β_3 $(g \times t)$
	Simulado	25	-1	0	-1
COMP	GEE-IN	24,997 (0,160)	-1,001 (0,051)	-0,005 (0,226)	-0,998 (0,072)
	GEE-SC	24,997 (0,160)	-1,001 (0,051)	-0,005 (0,226)	-0,998 (0,072)
	GEE-NE	24,997 (0,161)	-1,001 (0,052)	-0,005 (0,227)	-0,998 (0,074)
	GEE-AR	24,996 (0,167)	-1,001 (0,054)	-0,008 (0,236)	-0,998 (0,076)
MAR	GEE-IN	24,924 (0,162)	-0,456 (0,080)	-0,035 (0,230)	-0,880 (0,133)
	GEE-SC	24,927 (0,166)	-0,971 (0,074)	0,020 (0,237)	-1,014 (0,119)
	GEE-NE	24,896 (0,164)	-0,637 (0,076)	0,001 (0,233)	-0,955 (0,125)
	GEE-AR	24,987 (0,175)	-1,219 (0,083)	0,019 (0,248)	-1,084 (0,128)
IMP	GEE-IN	24,884 (0,165)	-0,845 (0,051)	-0,039 (0,233)	-0,945 (0,072)
	GEE-SC	24,884 (0,164)	-0,845 (0,051)	-0,039 (0,233)	-0,945 (0,072)
	GEE-NE	25,038 (0,164)	-0,881 (0,053)	0,031 (0,230)	-0,963 (0,074)
	GEE-AR	24,917 (0,167)	-0,840 (0,053)	-0,028 (0,237)	-0,941 (0,075)

A imputação pelo pareamento também não teve um bom desempenho. Assim como acontece com o escore de propensão esse método de pareamento usa a informação disponível (observada) de indivíduos com históricos semelhantes. Dependendo de como ocorre a perda ou do número de indivíduos na amostra, pode ser muito difícil encontrar um valor adequado para ser imputado.

Um problema com métodos de pareamento é que os erros padrões resultantes geralmente são inadequados, pelo fato de o valor imputado carregar pouca variabilidade. Isso geralmente acarreta erros padrões combinados das estimavas após imputação muitas vezes menores que aqueles obtidos com o banco completo. Nesse caso, a incerteza associada com o processo de imputação não é corretamente refletida.

6.3 Resultados MAR, Caso Homocedástico

Nessa situação, os dados foram gerados segundo o modelo (6.2). Aqui o modelo correto para análise assume variabilidade constante entre os tempos.

Novamente os valores mostrados são as médias de 5.000 repetições do processo de geração e perda de dados segundo o mecanismo MAR, descrito anteriormente. No corpo das tabelas a seguir os valores estão na forma: estimativa (erro padrão). A simulação foi repetida independentemente para cada método, e o nosso objetivo foi comparar como cada um deles se comporta, em média, nas 5.000 repetições do processo. Em alguns casos houve problemas de convergência no ajuste GEE com matriz de trabalho não estruturada; nessas situações os valores mostrados são a mediana das estimativas. O tamanho de cada banco criado foi agora de $n = 100$, totalizando 500 observações, situação mais comum de se verificar na prática.

A imputação múltipla foi conduzida para $k = 5$ bancos de dados imputados.

6.3.1 Modelo Normal (Bayesiano)

Na Tabela 6.6 estão os resultados de simulação considerando a imputação pelo modelo normal bayesiano. Como distribuição *a priori* para o modelo normal foi usada a não informativa discutida na Seção 5.1.2. Foram observados 41,82% de dados ausentes.

Tabela 6.6: Imputação pelo Modelo Normal (Bayesiano), n=100

		β_0 (i)	β_1 (t)	β_2 (g)	β_3 $(g \times t)$
	Simulado	25	-1	0	-1
COMP	GEE-IN	24,993 (0,353)	-0,999 (0,089)	0,010 (0,502)	-1,002 (0,126)
	GEE-SC	24,993 (0,353)	-0,999 (0,089)	0,010 (0,502)	-1,002 (0,126)
	GEE-NE	24,993 (0,351)	-0,999 (0,089)	0,008 (0,499)	-1,002 (0,127)
	GEE-AR	24,991 (0,370)	-0,998 (0,097)	0,013 (0,526)	-1,003 (0,137)
MAR	GEE-IN	24,983 (0,358)	-0,551 (0,152)	-0,027 (0,512)	-0,899 (0,254)
	GEE-SC	24,980 (0,368)	-1,065 (0,133)	0,022 (0,527)	-1,019 (0,217)
	GEE-NE	24,927 (0,363)	-0,706 (0,143)	0,011 (0,519)	-0,976 (0,235)
	GEE-AR	24,986 (0,389)	-1,294 (0,162)	0,021 (0,554)	-1,065 (0,255)
IMP	GEE-IN	24,976 (0,356)	-1,004 (0,091)	0,006 (0,505)	-0,998 (0,129)
	GEE-SC	24,976 (0,356)	-1,004 (0,091)	0,006 (0,505)	-0,998 (0,129)
	GEE-NE*	25,300 (0,348)	-1,060 (0,096)	0,095 (0,496)	-1,020 (0,135)
	GEE-AR	24,974 (0,370)	-1,010 (0,099)	0,004 (0,526)	-0,997 (0,140)

* Por problemas de convergência são apresentadas os valores medianos

Na situação que considera variabilidade constante entre os tempos o método de imputação com o modelo normal teve um desempenho bom. O grande vício que era observado para o coeficiente de tempo foi corrigido e, agora, as estimativas dos coeficientes são bastante próximas, independente da escolha da matriz de trabalho. Além disso, nota-se que a variabilidade final das estimavas, expressa em termos de erro padrão, está mais próxima da obtida com os dados completos.

6.3.2 Modelo Misto (Bayesiano)

Na Tabela 6.7 estão os resultados de simulação considerando a imputação pelo modelo misto bayesiano. Tanto para o efeito aleatório quanto para o erro residual foram usadas a Wishart invertida como distribuição *a priori* com ambos os parâmetros iguais a 1, de forma a minimizar seu impacto nos resultados finais. Foram observados 41,88% de dados ausentes.

Tabela 6.7: Imputação pelo Modelo Misto (Bayesiano), n=100

		β_0 (i)	β_1 (t)	β_2 (g)	β_3 $(g \times t)$
	Simulado	25	-1	0	-1
COMP	GEE-IN	24,998 (0,354)	-0,999 (0,089)	0,008 (0,503)	-1,003 (0,126)
	GEE-SC	24,998 (0,354)	-0,999 (0,089)	0,008 (0,503)	-1,003 (0,126)
	GEE-NE	24,999 (0,352)	-0,999 (0,089)	0,007 (0,500)	-1,002 (0,127)
	GEE-AR	25,000 (0,370)	-0,999 (0,096)	0,004 (0,526)	-1,002 (0,137)
MAR	GEE-IN	24,988 (0,359)	-0,549 (0,153)	-0,030 (0,513)	-0,900 (0,254)
	GEE-SC	24,987 (0,368)	-1,067 (0,134)	0,017 (0,527)	-1,021 (0,217)
	GEE-NE	24,935 (0,364)	-0,706 (0,143)	-0,005 (0,520)	-0,977 (0,235)
	GEE-AR	24,996 (0,389)	-1,299 (0,162)	0,010 (0,554)	-1,065 (0,255)
IMP	GEE-IN	25,030 (0,354)	-1,050 (0,107)	0,001 (0,510)	-1,000 (0,154)
	GEE-SC	25,030 (0,358)	-1,050 (0,107)	0,001 (0,510)	-1,000 (0,154)
	GEE-NE*	25,100 (0,362)	-1,050 (0,111)	-0,004 (0,517)	-0,993 (0,159)
	GEE-AR	25,017 (0,372)	-1,050 (0,113)	-0,002 (0,569)	-0,996 (0,162)

* Por problemas de convergência são apresentadas os valores medianos

Considerando o modelo misto com enfoque bayesiano, as estimativas finais dos coeficientes ficaram próximas daquelas geradas, mas levemente maiores para o efeito de tempo.

Novamente houve problemas com a matriz de trabalho não estruturada em algumas situações, o que impede que essa estrutura seja usada em todos os casos.

6.3.3 Modelo Misto (Frequentista)

Na tabela 6.8 estão os resultados de imputação com o modelo misto frequentista. O percentual médio de dados ausentes foi de 41,85%.

Tabela 6.8: Imputação pelo Modelo Misto (Frequentista), n=100

		β_0 (i)	β_1 (t)	β_2 (g)	β_3 $(g \times t)$
	Simulado	25	-1	0	-1
COMP	GEE-IN	24,996 (0,354)	-0,999 (0,089)	0,005 (0,502)	-1,003 (0,126)
	GEE-SC	24,996 (0,354)	-0,999 (0,089)	0,005 (0,502)	-1,003 (0,126)
	GEE-NE	24,995 (0,352)	-0,998 (0,090)	0,007 (0,499)	-1,003 (0,127)
	GEE-AR	24,997 (0,371)	-0,998 (0,097)	0,003 (0,526)	-1,003 (0,137)
MAR	GEE-IN	24,990 (0,358)	-0,551 (0,153)	-0,040 (0,512)	-0,895 (0,254)
	GEE-SC	24,986 (0,368)	-1,066 (0,134)	0,010 (0,527)	-1,020 (0,217)
	GEE-NE	24,934 (0,363)	-0,706 (0,143)	-0,001 (0,519)	-0,974 (0,235)
	GEE-AR	24,993 (0,389)	-1,299 (0,162)	0,007 (0,555)	-1,065 (0,255)
IMP	GEE-IN	24,999 (0,357)	-1,006 (0,102)	0,004 (0,508)	-1,008 (0,148)
	GEE-SC	24,999 (0,357)	-1,006 (0,102)	0,004 (0,508)	-1,008 (0,148)
	GEE-NE	25,008 (0,358)	-1,008 (0,105)	-0,001 (0,509)	-1,006 (0,152)
	GEE-AR	25,000 (0,371)	-1,006 (0,108)	0,002 (0,527)	-1,007 (0,156)

Em termos de estimativas para os coeficientes, o vício foi novamente removido ao imputarmos os dados ausentes com o modelo misto. Para toda matriz de trabalho adotada os valores estimados estão, em média, muito próximos dos gerados.

Os erros padrões combinados após imputação são maiores que aqueles obtidos com os dados completos, refletindo a incerteza inerente ao processo de imputação.

6.3.4 Escore de Propensão

Para imputação com o escore de propensão o procedimento adotado foi separado por grupo. Assim, foi feita a imputação para os dois grupos de forma independente, sendo incluídos no modelo logístico as variáveis observadas nos tempos anteriores (incluindo as imputadas). Foi observado 41,90% de dados ausentes.

Tabela 6.9: Imputação pelo Escore de Propensão, n=100

		β_0 (i)	β_1 (t)	β_2 (g)	β_3 $(g \times t)$
	Simulado	25	-1	0	-1
COMP	GEE-IN	24,998 (0,353)	-0,999 (0,089)	-0,003 (0,502)	-1,001 (0,126)
	GEE-SC	24,998 (0,353)	-0,999 (0,089)	-0,003 (0,502)	-1,001 (0,126)
	GEE-NE	24,998 (0,351)	-1,000 (0,090)	-0,003 (0,499)	-1,001 (0,127)
	GEE-AR	24,997 (0,370)	-0,999 (0,097)	-0,003 (0,526)	-1,001 (0,137)
MAR	GEE-IN	24,988 (0,358)	-0,551 (0,153)	-0,036 (0,512)	-0,896 (0,254)
	GEE-SC	24,986 (0,368)	-1,067 (0,134)	0,013 (0,527)	-1,019 (0,217)
	GEE-NE	24,934 (0,363)	-0,707 (0,143)	0,002 (0,519)	-0,974 (0,235)
	GEE-AR	24,992 (0,389)	-1,298 (0,162)	0,010 (0,554)	-1,065 (0,255)
IMP	GEE-IN	24,906 (0,361)	-0,820 (0,097)	-0,158 (0,520)	-0,825 (0,140)
	GEE-SC	24,906 (0,361)	-0,820 (0,097)	-0,158 (0,520)	-0,825 (0,140)
	GEE-NE*	25,300 (0,360)	-0,942 (0,107)	-0,016 (0,511)	-0,886 (0,152)
	GEE-AR	24,924 (0,366)	-0,817 (0,099)	-0,132 (0,525)	-0,812 (0,143)

* Por problemas de convergência são apresentadas os valores medianos

Após imputação pelo método do escore de propensão notamos vício nas estimativas. Esse vício agora é maior por estarmos com tamanho de amostra menor (100 indivíduos) que no caso anterior (500 indivíduos); e como a imputação é feita por pareamento há menos indivíduos disponíveis para encontrar um valor plausível, resultando em imputação viesada.

Também o ajuste com a matriz de trabalho GEE não estruturada apresentou problemas em alguns casos.

6.3.5 Pareamento

Para a imputação pelo método de pareamento com o algoritmo genético o procedimento de imputação foi separado por grupo. Assim, unidades dentro do primeiro grupo foram pareadas com unidades dentro do primeiro grupo, o mesmo acontecendo com o segundo grupo. Foi observado 41,83% de dados ausentes.

Tabela 6.10: Imputação por Pareamento, n=100

		β_0 (i)	β_1 (t)	β_2 (g)	β_3 $(g \times t)$
	Simulado	25	-1	0	-1
COMP	GEE-IN	25,010 (0,354)	-1,001 (0,089)	-0,011 (0,503)	-0,997 (0,126)
	GEE-SC	25,010 (0,354)	-1,001 (0,089)	-0,011 (0,503)	-0,997 (0,126)
	GEE-NE	25,010 (0,352)	-1,001 (0,090)	-0,009 (0,501)	-0,998 (0,127)
	GEE-AR	25,009 (0,372)	-1,002 (0,097)	-0,013 (0,528)	-0,995 (0,137)
MAR	GEE-IN	24,998 (0,359)	-0,551 (0,152)	-0,048 (0,513)	-0,893 (0,253)
	GEE-SC	24,997 (0,369)	-1,069 (0,133)	-0,002 (0,529)	-1,019 (0,217)
	GEE-NE	24,944 (0,364)	-0,708 (0,142)	-0,012 (0,520)	-0,971 (0,234)
	GEE-AR	25,006 (0,390)	-1,303 (0,161)	-0,007 (0,556)	-1,065 (0,254)
IMP	GEE-IN	24,886 (0,363)	-0,779 (0,098)	-0,142 (0,521)	-0,839 (0,141)
	GEE-SC	24,886 (0,363)	-0,779 (0,098)	-0,142 (0,521)	-0,839 (0,141)
	GEE-NE*	25,222 (0,356)	-0,876 (0,104)	-0,010 (0,505)	-0,909 (0,146)
	GEE-AR	24,917 (0,368)	-0,777 (0,101)	-0,123 (0,526)	-0,830 (0,144)

* Por problemas de convergência os valores apresentadas são a mediana

Assim como com a imputação pelo escore de propensão, a imputação pelo pareamento genético também não apresentou bons resultados. Novamente a quantidade pequena de informação disponível dentre os indivíduos com resposta observada é a justificativa. Aliás, qualquer método que use pareamento ou alguma combinação de regressão e pareamento – no qual o valor imputado é de um indivíduo da amostra – sofre do mesmo problema.

6.4 Discussão

Para o caso MAR, em que a probabilidade de o indivíduo apresentar dado ausente depende de alguma covariável medida ou respostas anteriores, os coeficientes estimados podem ser bastante viesados, como pudemos observar nas Seções 6.2 e 6.3. Nesse caso, as estimavas obtidas ao utilizar o modelo GEE podem levar a inferências enganosas quanto as tendências longitudinais. Os maiores vícios foram observados assumindo uma matriz de trabalho independente ou não estruturada, e o menor vício aconteceu quando foi escolhida uma estrutura de correlação da forma simetria composta. Para essa última praticamente não foi observado vício. No caso da matriz não estruturada houve vários problemas de convergência entre as repetições da simulação; nesse caso não houve convergência do algoritmo que ajusta o modelo GEE, assim essa estrutura não pode ser utilizada em todas as ocasiões.

Quanto aos métodos de imputação, são uma forma interessante de remover o vício nas estimativas do modelo GEE. Após a imputação, a escolha da matriz de trabalho tem pouca influência nos coeficientes de regressão, diferente do que ocorre na análise de dados disponíveis, no mecanismo MAR. À parte de todas as vantagens que fazer imputação múltipla num cenário MAR quando se opta pelo método de análise semiparamétrico GEE, a escolha inadequada do método de imputação pode acarretar muitos perigos.

Um bom método de imputação deve ser capaz de manter a relação de dependência longitudinal entre as variáveis no modelo. No caso do escore de propensão, por exemplo, isso não ocorre. Esse método pode apresentar grande vício quando as variáveis incluídas na regressão logística não são explicativas da probabilidade de *dropout* (Allison, 2000), que acontece na situação em que a perda é completamente ao acaso.

O outro método de pareamento – que utiliza uma generalização da distância de Mahalanobis para encontrar o indivíduo mais "semelhante" cuja resposta observada será usada como imputação – sofre do mesmo problema dos métodos de pareamento: a limitação da quantidade de informação disponível na amostra! Alguns métodos (não aplicados aqui)

combinam regressão e pareamento, mas a plausibilidade do valor imputado depende fortemente de como ocorreu a perda. Por exemplo, se todos os indivíduos com resposta acima de determinado valor no primeiro tempo abandonarem o estudo no próximo tempo então esses métodos não encontrarão um valor adequado para usar como imputação. Ainda, há o problema de subestimarem a variabilidade entre as imputações (Rubin, 1987). Esses métodos de pareamento foram criados para estudos tranversais; na aplicação a dados longitudinais existe a restrição de que os indivíduos sejam medidos nos mesmos instantes.

Quanto aos métodos de imputação baseados em modelos de regressão, o desempenho foi bastante satisfatório. O vício das estimativas foi removido, mesmo com um número pequeno de imputações e uma grande porcentagem de não resposta. Para a imputação com o modelo normal existe a restrição de que hajam tempos comuns de observação, o que não ocorre em muitos estudos. Modelos flexíveis, como o modelo misto adotado aqui, não fazem essa restrição e, portanto, são aplicáveis a uma ampla variedade de situações. No enfoque frequentista, o modelo misto teve um desempenho bastante bom ao imputar valores coerentes com os que formam perdidos.

CAPÍTULO 7

APLICAÇÃO A DADOS REAIS

7.1 Introdução

Neste capítulo apresentamos uma aplicação dos métodos de imputação aqui considerados a um conjunto de dados reais. Os dados são referentes a um estudo com pacientes internados em CTI no Hospital de Clínicas da Universidade Federal de Minas Gerais com quadro de insuficiência respiratória, na época da pandemia de influenza H1N1. Uma das características do estudo é a perda de informação por *dropout*.

7.1.1 Os Dados de H1N1

O estudo envolveu 35 pacientes dos quais 12 tiveram o diagnóstico confirmado de H1N1 (Grupo 1), 6 tiveram diagnóstico de influenza sazonal (ou seja, outros tipos de influenza) (Grupo 2), e 17 tiveram diagnósticos alternativos (embolia pumonar, etc) (Grupo 3).

Foram estudadas duas respostas que correspondem a duas moléculas inflamatórias, produzidas em caso de infecções e outros tipos de agressão (trauma, etc), medidas no sangue dos pacientes. A resposta que nos interessa aqui corresponde à proteína C reativa (C-reactive protein), uma proteína produzida principalmente no fígado. O valor normal vai até 10. Assim, para muitos indivíduos essa resposta elevou-se dezenas de vezes em relação aos valores normais. Isso significa inflamação mais intensa.

O período de coleta foi de agosto a novembro de 2009 (a epidemia no Brasil começou em abril daquele ano). Os pacientes foram acompanhados por 28 dias (ou até o óbito ou alta do CTI, quando ocorriam antes dos 28 dias). As medidas foram tomadas nos pacientes nos dias 0, 3, 5 e 7 de acompanhamento.

Houve perda de 19 dentre os 140 (35×4) valores programados, o que corresponde a 13,57% de dados ausentes. Uma das razões para a não resposta foi o óbito para dois

pacientes – não se sabe se essas perdas são decorrentes dos valores observados da resposta, o que caracterizaria um mecanismo MAR. Nos demais casos a justificativa foi alta hospitalar ou problemas técnicos na coleta de sangue. Para apenas um dos indivíduos (no de diagnósticos alternativos) a perda foi intermitente.

Na Tabela 7.1 são mostradas estatísticas descritivas para os três grupos.

Tabela 7.1: Descrição da Resposta por Tempo e Grupo

		Tempo			
Grupo	Estatística	Dia 0	Dia 3	Dia 5	Dia 7
	n	12	12	11	9
1	Média	197,2	193,8	152,9	118,6
	Desvio Padrão	139,5	114,4	89,3	70,8
	n	6	5	5	4
2	Média	83,2	94,4	65,0	104,3
	Desvio Padrão	71,5	71,0	54,7	91,4
	n	17	15	12	13
3	Média	122,9	91,8	127,5	101,8
	Desvio Padrão	84,1	73,9	93,4	78,3

A Figura 7.1 apresenta a representação gráfica dos perfis dos 35 pacientes, por grupo.

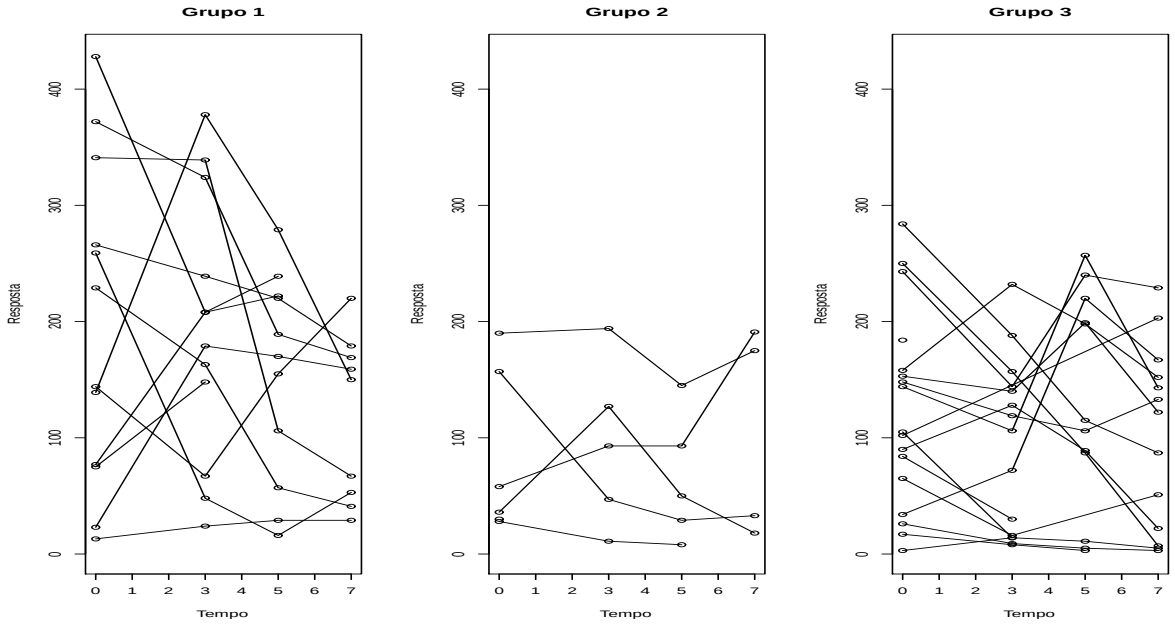

Figura 7.1: Gráfico do Perfil dos Pacientes por Grupo

O que podemos notar é uma grande variabilidade nos dados, com possível tendência de queda dos valores da resposta ao longo do período de avaliação.

A seguir o gráfico com os perfis alisados.

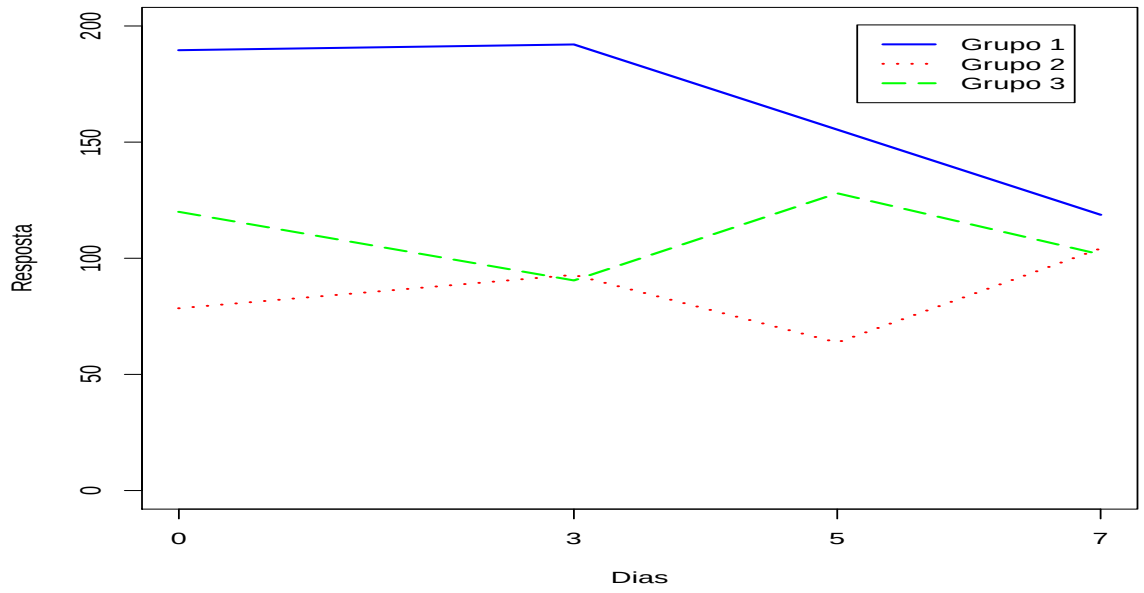

Figura 7.2: Gráfico do Perfil Alisado por Grupo

7.1.2 O Modelo de Análise

O modelo GEE utilizado na análise teve a seguinte forma para a média populacional:

$$E(y_{ij}) = \beta_0 Dias + \beta_1 I(Grupo1) + \beta_2 I(Grupo2) + \beta_3 I(Grupo3)$$

ou seja, consideramos intercepto diferentes mas o mesmo efeito do tempo para todos os grupos. O efeito de interação, que permitiria inclinações diferentes para cada grupo, foi testado mas não se mostrou significativo. Como os tempos entre as medições não foram igualmente espaçados a matriz de trabalho AR-1 não pode ser utilizada. A análise foi então conduzida com as matrizes de trabalho independente, simetria composta e não estruturada.

Na Tabela 7.2 encontram-se os coeficientes ajustados para as três matrizes de trabalho,

apresentados na forma: estimativa (erro padrão).

Tabela 7.2: Ajuste GEE aos Dados Disponíveis, dados H1N1

	β_0	β_1	β_2	β_3
GEE-IN	-4,568 (3,226)	185,081 (26,481)	101,184 (24,900)	126,605 (19,905)
GEE-SC	-5,537 (3,220)	188,387 (26,483)	97,820 (24,907)	127,534 (19,300)
GEE-NE	-4,084 (3,253)	181,651 (26,324)	96,777 (25,499)	123,527 (19,851)

Mesmo com um percentual pequeno de dados ausentes notamos que as estimativas GEE são diferentes para alguns coeficientes, como o efeito do intercepto, por exemplo.

A seguir aplicamos os cinco métodos de imputação múltipla adotados nesse trabalho e comparamos as estimativas com aquelas obtidas a partir dos dados disponíveis.

7.2 Resultados dos Métodos de Imputação Múltipla

Para imputação múltipla adotamos $k = 10$ bancos de dados imputados. O número maior de imputações comparado com a simulação é justificado pela maior variabilidade observada entre as imputações. No corpo das tabelas a seguir os valores estão na forma: estimativa (erro padrão).

7.2.1 Modelo Normal (Bayesiano)

Os resultados da imputação pelo modelo normal encontram-se na Tabela 7.3. A distribuição *a priori* adotada foi a não informativa, como discutido na Seção 5.1.2.

Tabela 7.3: Imputação pelo Modelo Normal (Bayesiano), dados H1N1

	β_0	β_1	β_2	β_3
GEE-IN	-8,453 (2,934)	188,025 (27,186)	103,531 (38,267)	125,455 (19,228)
GEE-SC	-8,453 (2,934)	188,025 (27,185)	103,531 (38,267)	125,455 (19,228)
GEE-NE	-8,321 (2,972)	181,057 (27,741)	105,379 (41,572)	127,106 (19,353)

A imputação pelo modelo normal traz alguns resultados interessantes: enquanto as estimativas de grupo permanecem em geral inalteradas quando comparadas com a Tabela 7.2, o efeito do tempo praticamente dobra e o erros padrões associados com esses coeficientes são menores que no ajuste com os dados disponíveis. Para os demais efeitos, embora as estimavas dos coeficientes não tenham sofrido alteração a incerteza associada a estes foi aumentada.

7.2.2 Modelo Misto (Bayesiano)

Na Tabela 7.4 estão os resultados da imputação pelo modelo linear misto com enfoque Bayesiano. Foi adotado o modelo (2.1) com efeitos aleatórios no intercepto e tempo. As distribuições *a priori* usadas são as mesmas das simulações de Monte Carlo, conforme discutido na Subseção 6.2.2.

Após a imputação as estimativas dos coeficientes para as diferentes matrizes de trabalho estão mais próximas entre si. O que mais se destaca é a incerteza da imputação na

Tabela 7.4: Imputação pelo Modelo Misto (Bayesiano), dados H1N1

	β_0	β_1	β_2	β_3
GEE-IN	-5,214 (3,192)	185,782 (31,917)	101,755 (73,144)	126,734 (25,149)
GEE-SC	-5,214 (3,192)	185,782 (31,917)	101,755 (73,144)	126,734 (25,149)
GEE-NE	-5,359 (3,183)	184,941 (37,733)	104,003 (84,663)	126,564 (24,952)

estimação de β_2, cuja variabilidade praticamente triplicou.

7.2.3 Modelo Misto (Frequentista)

Adotou-se o modelo linear misto também com efeitos aleatórios tanto no intercepto quanto no tempo. A escolha dessa estrutura de componentes de variância foi justificada por um teste de razão de verossimilhança que apontou este como o melhor modelo para análise.

Na Tabela 7.5 encontram-se os resultados da imputação com o modelo linear misto, enfoque frequentista.

Tabela 7.5: Imputação pelo Modelo Misto (Frequentista), dados H1N1

	β_0	β_1	β_2	β_3
GEE-IN	-5,812 (3,055)	186,609 (30,067)	107,429 (64,282)	131,788 (21,633)
GEE-SC	-5,909 (3,052)	186,940 (30,085)	107,093 (64,387)	131,871 (19,266)
GEE-NE	-5,286 (3,069)	181,025 (30,159)	103,642 (49,421)	129,350 (21,622)

Assim, como o enfoque bayesiano, a imputação introduziu grande variabilidade nas estimativas do coeficiente associado ao Grupo 2.

7.2.4 Escore de Propensão

Devido ao tamanho amostral reduzido e ao efeito de interação entre tempo e grupo não ser significativo em análises preliminares, a imputação pelo escore de propensão não foi conduzida de forma separada por grupo. O modelo logístico utilizado para estimação da probabilidade de *dropout* incluiu as respostas nos tempos anteriores a imputação, assim como o indicador de grupo.

Na Tabela 7.6 encontram-se os resultados da imputação sob esse enfoque.

Tabela 7.6: Imputação pelo Escore de Propensão, dados H1N1

	β_0	β_1	β_2	β_3
GEE-IN	-3,615 (3,091)	188,637 (33,316)	86,678 (44,672)	126,102 (21,029)
GEE-SC	-3,615 (3,091)	188,637 (33,316)	86,678 (44,672)	126,102 (21,029)
GEE-NE	-2,249 (1,974)	184,470 (46,823)	79,906 (36,722)	122,751 (23,888)

Em comparação com os métodos anteriores, as estimativas dos coeficientes de tempo e grupo ficaram menores. Nota-se, também, a incerteza na estimação do efeito dos dois primeiros grupos.

7.2.5 Pareamento

Por fim, a Tabela 7.7 contem os resultados dos bancos de dados imputados pelo método de pareamento genético. A imputação por esse método foi feita de forma semelhante ao escore de propensão, ou seja, a imputação não foi separada por grupo.

Tabela 7.7: Imputação por Pareamento, dados H1N1

	β_0	β_1	β_2	β_3
GEE-IN	-5,104 (2,989)	183,999 (30,479)	96,629 (55,209)	129,829 (25,757)
GEE-SC	-5,104 (2,989)	183,999 (30,479)	96,629 (55,209)	129,829 (25,757)
GEE-NE	-5,581 (2,989)	181,048 (34,363)	100,359 (55,795)	123,775 (21,068)

Os coeficientes de regressão combinados após imputação estão próximos daqueles obtidos com os dados disponíveis, se levarmos em conta a grande incerteza das estimativas.

7.3 Discussão

Pelo tamanho amostral reduzido não foi possível uma comparação precisa entre os métodos de imputação. Mas podemos destacar alguns pontos.

No processo de imputação pelo escore de propensão os coeficientes estimados pela regressão logística foram não significativos, isso pode ser um indicativo que o mecanismo de perda de dados é MCAR. Nesse caso, a imputação pelo escore de propensão não apresentará bons resultados. O problema é que os quintis construídos de grupo "similares" são essencialmente agrupamentos aleatórios das observações. Embora o método de imputação que usa o pareamento genético não sofra desse problema, ainda assim sua utilidade fica limitada pelo número de observações disponíveis. Em situações como essa pode não ser vantajoso conduzir imputação múltipla por algum desses métodos.

Percebemos que os resultados das análises após a imputação podem diferir de forma bastante significativa, mesmo para uma pequena proporção de dados ausentes. A grande mudança nas estimativas ocorreu em termos dos erros padrões estimados, que agora passaram a ser muito maiores que na análise de dados disponíveis. Isso é reflexo da variabilidade entre as imputações.

Pela discrepância entre as estimativas combinadas após imputação percebe-se que é de fundamental importância a escolha do modelo. O modelo de imputação e de análise podem ser distintos, o que pode ser uma vantagem para a imputação múltipla, por permitir a incorporação de informação relevante na predição do *dropout* mas que não é de interesse no modelo final de análise. Schafer (2003) apresenta uma discussão sobre esse e outros casos.

CAPÍTULO 8

CONCLUSÕES

A presente dissertação mostrou como dados ausentes podem apresentar grande impacto na estimação de quantidades de interesse em estudos longitudinais. O impacto além do vício das estimativas também está na precisão destas.

Coerente com os resultados da literatura vimos, através de simulação que o método GEE – que não requer especificação de distribuição para a variável dependente – apresenta estimativas bastante viesadas dos coeficientes de regressão na situação MAR. Nesse caso, diferente do que ocorre com os dados completos a escolha da matriz de correlação de trabalho tem fundamental importância na estimativa final.

Para as diferentes estruturas de correlação consideradas os parâmetros estimados foram bastante diversos. A matriz de trabalho simetria composta foi a que apresentou o melhor desempenho para as situações das simulações, com resultados praticamente não viesados em todas as situações geradas. Já a independente apresentou os maiores vícios.

Como alternativa para tratar o problema no caso GEE adotamos a imputação múltipla. Nesse caso, a imputação pelo modelo misto de regressão, tanto no enfoque bayesiano como no frequentista, se mostrou um método bastante eficaz. Considerar o modelo normal, que usa como imputação um valor gerado da distribuição dos dados ausentes condicional aos dados observados, também é uma boa escolha quando a variável resposta é normal e o delineamento é balanceado. Métodos que consideram alguma forma de pareamento, como o escore de propensão e o pareamento genético, não tiveram um desempenho tão bom quanto os primeiros. Uma possível razão para esse comportamento é a não utilização da associação entre as respostas e as covariáveis na construção das medidas de pareamento. Ressaltamos que em amostras pequenos esse tipo de imputação pode não ser adequado. Provavelmente a principal razão para este mal desempenho é ignorar a relação funcional entre as respostas e as covariáveis.

Como conclusão notamos que à custa da flexibilidade do método GEE para análise de dados longitudinais deve-se tomar bastante cuidado na escolha da estrutura de correlação a ser adotada na presença da dados ausentes. Nessas situações a imputação de dados é uma ferramenta adequada que além de remover o vício das estimativas também recupera parte da variabilidade perdida com os dados ausentes.

CAPÍTULO 9

FUNÇÕES UTILIZADAS NAS SIMULAÇÕES

9.1 Códigos para Geração dos Dados

O código a seguir foi utilizado para geração dos dados completos e da não resposta no cenário MAR, caso heterocedástico. Para o caso homocedástico foi alterado apenas o tamanho amostral e retirado o afeito aleatório na variável *tempo*.

```
library(MASS)
####################################
# Geração do Banco de Dados
n=500
Z=cbind(c(1,1,1,1,1),c(0,1,2,3,4))
s0=4 #variância do e.a. intercepto
s1=0.25 #variância do e.a. tempo
s10=s01=-0.1 #covariância do e.a.
s2=4 #variância do erro
Sigma=matrix(c(s0,s10,s01,s1),2,2) #matriz cov. efeitos aleatórios
S=diag(x=1,5,5) #matriz identidade para o erro
grupo=rbinom(n,1,p=0.5) #geração dos grupos de tamanho iguais
erro=mvrnorm(n=n,mu=c(0,0,0,0,0),Sigma=diag(x=4,5,5),empirical=TRUE)
#geração do erro residual
B=mvrnorm(n=n,mu=c(0,0),Sigma=Sigma,empirical=TRUE) #geração dos efeitos
aleatórios correlacionados
v0=B[,1] #efeito aleatório intercepto
v1=B[,2] #efeito aleatório tempo
b0=25 #beta efeito intercepto
```

```
b1=-1 #beta efeito tempo
b2=0  #beta efeito grupo
b3=-1 #beta efeito grupo*tempo
tempo=c(0,1,2,3,4)
y=matrix(0,n,5)
id=1:n
for(i in 1:n){ #geração das respostas
y[i,]=b0+b1*tempo+b2*grupo[i]+b3*(grupo[i]*tempo)+v0[i]+v1[i]*tempo+erro[i,]
}
colnames(y)=c("t0","t1","t2","t3","t4")
dados=as.data.frame(cbind(id,y,grupo))
Y=c(y[,1],y[,2],y[,3],y[,4],y[,5])
GRUPO=rep(grupo,5)
TEMPO=c(rep(0,n),rep(1,n),rep(2,n),rep(3,n),rep(4,n))
ID=rep(id,5)
dados1=as.data.frame(cbind(ID,Y,TEMPO,GRUPO))
dados1<-dados1[order(dados1$ID),] #banco de dados completos
####################################
# Geração da Não Resposta
R=matrix(0,n,5) #R==1 é missing
colnames(R)=c("r0","r1","r2","r3","r4")
mar=cbind(dados,R)
naux=length(mar$r1[mar$t0<23 & mar$grupo==1])
raux=rbinom(n=naux,1,p=0.8)
mar$r1[mar$t0<23 & mar$grupo==1]=raux
naux=length(mar$r2[mar$t1<23 & mar$grupo==1])
raux=rbinom(n=naux,1,p=0.8)
mar$r2[mar$t1<23 & mar$grupo==1]=raux
naux=length(mar$r3[mar$t2<23 & mar$grupo==1])
```

```
raux=rbinom(n=naux,1,p=0.8)
mar$r3[mar$t2<23 & mar$grupo==1]=raux
naux=length(mar$r4[mar$t3<23 & mar$grupo==1])
raux=rbinom(n=naux,1,p=0.8)
mar$r4[mar$t3<23 & mar$grupo==1]=raux
naux=length(mar$r1[mar$t0<23 & mar$grupo==0])
raux=rbinom(n=naux,1,p=0.8)
mar$r1[mar$t0<23 & mar$grupo==0]=raux
naux=length(mar$r2[mar$t1<23 & mar$grupo==0])
raux=rbinom(n=naux,1,p=0.8)
mar$r2[mar$t1<23 & mar$grupo==0]=raux
naux=length(mar$r3[mar$t2<23 & mar$grupo==0])
raux=rbinom(n=naux,1,p=0.8)
mar$r3[mar$t2<23 & mar$grupo==0]=raux
naux=length(mar$r4[mar$t3<23 & mar$grupo==0])
raux=rbinom(n=naux,1,p=0.8)
mar$r4[mar$t3<23 & mar$grupo==0]=raux
mar$r2[mar$r1==1]=1
mar$r3[mar$r2==1]=1
mar$r4[mar$r3==1]=1
mar$t1[mar$r1==1]=NA
mar$t2[mar$r2==1]=NA
mar$t3[mar$r3==1]=NA
mar$t4[mar$r4==1]=NA
Y.m=c(mar$t0,mar$t1,mar$t2,mar$t3,mar$t4)
mar1=as.data.frame(cbind(ID,Y.m,TEMPO,GRUPO))
mar1<-mar1[order(mar1$ID),] #Banco de dados com missings
```

9.2 Função de Imputação pelo Escore de Propensão

A seguir função usada para imputação com o método do escore de propensão.

Parâmetros de Entrada

p: probabilidade predita de o dado ser ausente (estimada por regressão logística)

y: variável resposta a ser imputada

```
####################################
# Função para Imputação
ps.imput=function(y,p){
dados=as.data.frame(cbind(y,p))
yim=y
quintil=rep(0,length(y))
q=quantile(p,  probs = c(0.2,0.4,0.6,0.8)) #quintis
Q1=y[p<=q[1]] #primeiro quintil
Q2=y[p>=q[1] & p<q[2]] #segundo quintil
Q3=y[p>=q[2] & p<q[3]] #terceiro quintil
Q4=y[p>=q[3] & p<q[4]] #quarto quintil
Q5=y[p>=q[4]] #quinto quintil
B11=B22=B33=B44=B55=1
mat=abs(outer(dados$p,dados$p,FUN="-"))
diag(mat)=500
mat[,which(is.na(dados$y))]=500
resample <- function(x, size, ...)
       if(length(x) <= 1) { if(!missing(size) && size == 0) x[FALSE] else x
       } else sample(x, size, ...)
{if(sum(!is.na(Q1))==0){B11=0} else
B1=resample(Q1[!is.na(Q1)],replace=TRUE)} #bootstrap no primeiro quintil
{if(sum(!is.na(Q2))==0){B22=0} else
B2=resample(Q2[!is.na(Q2)],replace=TRUE)} #bootstrap no segundo quintil
```

```
{if(sum(!is.na(Q3))==0){B33=0} else
B3=resample(Q3[!is.na(Q3)],replace=TRUE)} #bootstrap no primeiro quintil
 {if(sum(!is.na(Q4))==0){B44=0} else
B4=resample(Q4[!is.na(Q4)],replace=TRUE)} #bootstrap no quarto quintil
{if(sum(!is.na(Q5))==0){B55=0} else
B5=resample(Q5[!is.na(Q5)],replace=TRUE)} #bootstrap no quinto quintil
for(i in 1:length(y)){
if(p[i]<=q[1]) {quintil[i]=1} else
if(p[i]>=q[1] & p[i]<q[2]) {quintil[i]=2} else
if(p[i]>=q[2] & p[i]<q[3]) {quintil[i]=3} else
if(p[i]>=q[3] & p[i]<q[4]) {quintil[i]=4} else
if(p[i]>=q[4]) {quintil[i]=5}
}
for(i in 1:length(y)){#mínimo ps como valor imputado se não
houver dados no quintil
if(is.na(yim[i]) &  quintil[i]==1 & B11==0)
{yim[i]=dados$y[which.min(mat[i,])]} else
if(is.na(yim[i]) &  quintil[i]==2 & B22==0)
{yim[i]=dados$y[which.min(mat[i,])]} else
if(is.na(yim[i]) &  quintil[i]==3 & B33==0)
{yim[i]=dados$y[which.min(mat[i,])]} else
if(is.na(yim[i]) &  quintil[i]==4 & B44==0)
{yim[i]=dados$y[which.min(mat[i,])]} else
if(is.na(yim[i]) &  quintil[i]==5 & B55==0)
{yim[i]=dados$y[which.min(mat[i,])]}
}
for(i in 1:length(y)){#imputação pelos quintis...
if(is.na(yim[i]) & quintil[i]==1) {yim[i]=resample(B1,1,replace=TRUE)} else
if(is.na(yim[i]) & quintil[i]==2) {yim[i]=resample(B2,1,replace=TRUE)} else
```

```
if(is.na(yim[i]) & quintil[i]==3) {yim[i]=resample(B3,1,replace=TRUE)} else
if(is.na(yim[i]) & quintil[i]==4) {yim[i]=resample(B4,1,replace=TRUE)} else
if(is.na(yim[i]) & quintil[i]==5) {yim[i]=resample(B5,1,replace=TRUE)}
}
#cat("Tamanho Amostra:",length(y),"\n")
#cat("Casos Missings:", sum(is.na(y)),"\n")
#cat("Casos Completo:", sum(!is.na(y)),"\n")
#cat("Percentual Missings:", 100*(sum(is.na(y)))/length(y),"\n")
res=data.frame(cbind(y,p,quintil,yim))
}
```

9.3 Função de Imputação por Pareamento Genético

A seguir função utilizada para imputação pelo método do pareamento genético.

```
library(Matching)
####################################
# Função para Imputação
for(i in 1:5) { #cinco imputações
nam <- paste("w",i, sep=".")
M=mar
M1=subset(M,grupo==1) #imputação por grupo
n1=dim(M1)[1]
cov=with(M1,t0)
mout=with(M1,GenMatch(Tr = r1, X = cov, ties=F,replace=TRUE))
m=with(M1,Match(Tr = r1, X = cov,replace=TRUE,ties=F,M=1,,Weight.matrix=mout))
M1$t1[m$index.treated]=M1$t1[m$index.control] #imputação no primeiro tempo
cov=with(M1,cbind(t0,t1))
mout=with(M1,GenMatch(Tr = r2, X = cov, ties=F,replace=TRUE))
m=with(M1,Match(Tr = r2, X = cov,replace=TRUE,ties=F,M=1,Weight.matrix=mout))
```

```
M1$t2[m$index.treated]=M1$t2[m$index.control] #imputação no segundo tempo
cov=with(M1,cbind(t0,t1,t2))
mout=with(M1,GenMatch(Tr = r3, X = cov,  ties=F,replace=TRUE))
m=with(M1,Match(Tr = r3, X = cov,replace=TRUE,ties=F,M=1,Weight.matrix=mout))
M1$t3[m$index.treated]=M1$t3[m$index.control] #imputação no terceiro tempo
cov=with(M1,cbind(t0,t1,t2,t3))
mout=with(M1,GenMatch(Tr = r4, X = cov,  ties=F,replace=TRUE))
m=with(M1,Match(Tr = r4, X = cov,replace=TRUE,ties=F,M=1,Weight.matrix=mout))
M1$t4[m$index.treated]=M1$t4[m$index.control] #imputação no quarto tempo
Y1=c(M1$t0,M1$t1,M1$t2,M1$t3,M1$t4)
GRUPO1=rep(M1$grupo,5)
TEMPO1=c(rep(0,n1),rep(1,n1),rep(2,n1),rep(3,n1),rep(4,n1))
ID1=rep(M1$id,5)
matching1=as.data.frame(cbind(ID1,Y1,TEMPO1,GRUPO1))
names(matching1)=c("ID","Y","TEMPO","GRUPO")
####################################
M0=subset(M,grupo==0)  #imputação por grupo
n0=dim(M0)[1]
cov=with(M0,t0)
mout=with(M0,GenMatch(Tr = r1, X = cov, ties=F,replace=TRUE))
m=with(M0,Match(Tr = r1, X = cov,replace=TRUE,ties=F,M=1,,Weight.matrix=mout))
M0$t1[m$index.treated]=M0$t1[m$index.control]
cov=with(M0,cbind(t0,t1))
mout=with(M0,GenMatch(Tr = r2, X = cov, ties=F,replace=TRUE))
m=with(M0,Match(Tr = r2, X = cov,replace=TRUE,ties=F,M=1,Weight.matrix=mout))
M0$t2[m$index.treated]=M0$t2[m$index.control]
cov=with(M0,cbind(t0,t1,t2))
mout=with(M0,GenMatch(Tr = r3, X = cov,  ties=F,replace=TRUE))
m=with(M0,Match(Tr = r3, X = cov,replace=TRUE,ties=F,M=1,Weight.matrix=mout))
```

```
M0$t3[m$index.treated]=M0$t3[m$index.control]
cov=with(M0,cbind(t0,t1,t2,t3))
mout=with(M0,GenMatch(Tr = r4, X = cov,  ties=F,replace=TRUE))
m=with(M0,Match(Tr = r4, X = cov,replace=TRUE,ties=F,M=1,Weight.matrix=mout))
M0$t4[m$index.treated]=M0$t4[m$index.control]
Y0=c(M0$t0,M0$t1,M0$t2,M0$t3,M0$t4)
GRUPO0=rep(M0$grupo,5)
TEMPO0=c(rep(0,n0),rep(1,n0),rep(2,n0),rep(3,n0),rep(4,n0))
ID0=rep(M0$id,5)
matching0=as.data.frame(cbind(ID0,Y0,TEMPO0,GRUPO0))
names(matching0)=c("ID","Y","TEMPO","GRUPO")
pareado=rbind(matching1,matching0)
pareado<-pareado[order(pareado$ID),]
assign(nam, pareado) #atribuição dos bancos aos objetos da forma w.1, w.2,...
#banco final com as 5 imputações
}
```

REFERÊNCIAS BIBLIOGRÁFICAS

Allison, P. D. 2000. Multiple Imputation for Missing Data: A Cautionary Tale. *Sociological Methods and Research*, **28**, 301–309.

DeGroot, M. H. 1970. *Optimal Statistical Decisions*. McGraw-Hill.

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.

Goodnight, J. H. 1979. A Tutorial on the Sweep Operator. *American Statistician*, **33**, 149–158.

Hedeker, D., & Gibbons, R. D. 2006. *Longitudinal Data Analyses*. Wiley-Interscience.

Horton, N. J., & Lipsitz, S. R. 2001. Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables. *Journal of the American Statistical Association*, **55**, 244–254.

Laird, M., & Ware, James H. 1982. Random-Effects Models for Longitudinal Data. *Biometrics*, **38**, 963–974.

Lavori, Philip W., Dawson, Ree, & Shera, David. 1995. A Multiple Imputation Strategy for Clinical Trials with Truncation of Patient Data. *Statistics in Medicine*, **14**, 1913–1925.

Liang, K. Y., & Zeger, S. L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Little, Roderick. J. A., & Rubin, Donald. R. 1987. *Statistical Analysis with Missing Data*. Wiley.

McCullagh, P., & Nelder, J.A. 1989. *Generalized linear models.* Chapman and Hall.

Muirhead, Robb J. 1982. *Aspects of Multivariate Statistical Theory.* Wiley-Interscience.

Patterson, H.D., & Thompson, R. 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

Rosenbaum, Paul R., & Rubin, Donald B. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, **70**, 41–55.

Rosenbaum, Paul R., & Rubin, Donald B. 1985. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician*, **39**, 33–38.

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

Schafer, J.L. 1997a. *Analysis of Incomplete Multivariate Data.* Chapman & Hall.

Schafer, J.L. 1997b. *Imputation of missing covariates under a multivariate linear mixed model.* Tech. rept. Dept. of Statistics, The Pennsylvania State University.

Schafer, Joseph L. 2003. Multiple Imputation in Multivariate Problems When the Imputation and Analyses Models Differ. *Statistica Neerlandica*, **57**, 19–35.

Schafer, Joseph L., & Yucel, Recai M. 2002. Computational Strategies for Multivariate Linear Mixed-Effects Models With Missing Values. *Journal of Computational and Graphical Statistics*, **11**, 437–457.

Sekhon, Jasjeet S. 2007. Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R. *Journal of Statistical Software.*

Sekhon, J.S., & Mebane, Jr. W. R. 1998. Genetic Optimization Using Derivatives: Theory and Application to Nonlinear Models. *Political Analysis*, **7**, 189–203.

Tanner, M. A., & Wong, W. H. 1987. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, **82**, 528–540.

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

Verbeke, Geerk, & Molenberghs, Geerk. 2000. *Linear Mixed Models for Longitudinal Data*. Springer.