Universidade Estadual de Maringá
Programa de Pós-graduação em Bioestatística

MARINA GANDOLFI

IMPUTAÇÃO MÚLTIPLA VIA ALGORITMO MICE E MÉTODO IMLD

Dissertação de Mestrado

Maringá-PR
2016

MARINA GANDOLFI

IMPUTAÇÃO MÚLTIPLA VIA ALGORITMO MICE E MÉTODO IMLD

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá como requisito parcial para obtenção do título de Mestre em Bioestatística.

Universidade Estadual de Maringá – UEM

Departamento de Estatística

Programa de Pós-Graduação em Bioestatística

Orientador: Prof. Dr. Eraldo Schunk Silva

Coorientador: Prof. Dr. Carlos Tadeu Dos Santos Dias

Maringá-PR

2016

Dados Internacionais de Catalogação na Publicação (CIP)
(Biblioteca Central - UEM, Maringá, PR, Brasil)

Gandolfi, Marina
G196i Imputação múltipla via algorítmo MICE e método IMLD / Marina Gandolfi. -- Maringá, 2016.
83,[16] f. : il. color., figs., tabs., quadros

Orientador: Prof. Dr. Eraldo Schunk Silva.
Coorientador: Prof. Dr. Carlos Tadeu Dos Santos Dias.
Dissertação (mestrado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Estatística, Programa de Pós-Graduação em Bioestatística, 2016.

1. Base de dados incompleta - Dados faltantes. 2. Matriz - Decomposição por valor singular. 3. Softwares estatísticos - Imputação Múltipla Livre de Distribuição (IMLD). 4. Bioestatística - Interdisciplinariedade. 5. Softwares estatísticos - Métodos de imputação. 6. Multivariate Imputation by Chained Equatoins (MICE) - Distribuição de dados. I. Silva, Eraldo Schunk, orient. II. Dias, Carlos Tadeu Dos Santos, coorient. III. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Estatística. Programa de Pós-Graduação em Bioestatística. IV. Título.

CDD 21.ed. 570.15195
AMMA-003084

MARINA GANDOLFI

Imputação Múltipla via Algoritmo Mice e Método IMLD

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de Mestre em Bioestatística.

BANCA EXAMINADORA

Prof. Dr. Eraldo Schunk Silva
Universidade Estadual de Maringá - UEM

Profª. Dra. Mariana Ragassi Urbano
Universidade Estadual de Londrina - UEL

Profª. Dra. Isolde Previdelli
Universidade Estadual de Maringá - UEM

Maringá, 18 de março de 2016.

Dedico,
com muito amor e gratidão,
à minha família.

Agradecimentos

A Deus, por ouvir minhas orações e iluminar o meu caminho.

Aos meus pais João e Ivani, por não medirem esforços na educação dada a mim e aos meus irmãos. Pelos exemplos de honestidade e força. Por me ampararem nos momentos difíceis. Por me inspirarem a ser uma pessoa melhor a cada dia. Por terem trabalhado muito, todos os dias, para que eu pudesse chegar até aqui.

Aos meus manos Everaldo, Mariza e Leonardo, por apoiarem minhas escolhas. Pela amizade, paciência e companheirismo. Pelo carinho que sempre me dedicam, com o qual recarrego minhas energias. Pelos conselhos e aprendizados . Deus não poderia ter me presenteado com irmãos melhores.

Ao meu namorado e confidente Leonardo Balena, por sempre se fazer presente nos meus dias. Pelas experiências e aventuras inesquecíveis. Por superar junto comigo momentos complicados. Pela ajuda com as traduções. Pela confiança e por todo o amor.

Ao meu orientador Professor Eraldo, por ter me recebido gentilmente e ter acreditado no meu potencial. Pelos conselhos, pela disponibilidade. Por incentivar meu crescimento profissional e por dividir comigo suas experiências e seus conhecimentos.

Ao meu coorientador Professor Carlos Tadeus Dos Santos Dias, por ter se disponibilizado a coorientar-me, pela forma como me recebeu na ESALQ, pelas dúvidas sanadas, pelas sugestões e pelos materiais fornecidos.

À Professora Isolde, coordenadora do programa, pelo acolhimento junto ao mestrado, pelo abraço nas horas difíceis, por sempre nos passar confiança e acalmar as marés agitadas. Por compartilhar ricas experiências e saberes. Por não medir esforços e abdicar do seu tempo em prol do crescimento da pós-graduação em Bioestatística. Por seu astral contagiante.

Aos demais professores, Vanderly Janeiro, Rosangela Santana, Terezinha Guedes, Eniuce Menezes, Josmar Mazucheli, Edson Martinez, Andrea Diniz e Robson Rossi, pela dedicação às disciplinas ministradas, pelas ajudas prestadas e por fazerem parte desta conquista.

Ao Sergio Arciniegas Alarcón, por ter respondido prontamente meus e-mails, pelos válidos esclarecimentos e pelos artigos compartilhados.

Aos companheiros de jornada, Rafaela, Kelly, Matheus, Marcos, Viviane, Guilherme, Paulo, Isabela, Omar, Emerson, Sérgio, Oilson, Jardel, por tornarem essa caminhada mais leve, dividindo frustrações e conhecimentos. Por todo apoio e companheirismo.

À todos os amigos e familiares, pelo estímulo durante este trajeto.

À CAPES, pelo apoio financeiro.

"Mármore, sim, mas mole. E vento, porque não?
Mármore capaz de tudo,
de tudo recolher.
e transmudar em nada. De transmudar o ouro
– alquimia ao contrário – na poeira que o vento
ao próprio vento espalha...
Mármore, sim, mas mole."
(David Mourão-Ferreira)

Resumo

Um problema comum em análises estatísticas é a ocorrência de bases de dados incompletas. Geralmente, nessas situações, restringe-se a análise aos sujeitos com dados completos nas variáveis. Esse procedimento reduz o tamanho da amostra e pode resultar em estimativas tendenciosas. O "preenchimento" dos dados faltantes pode ser feito por meio da imputação múltipla (IM), em que cada valor ausente é substituído por um conjunto de valores plausíveis, incorporando a incerteza sobre o valor a ser imputado. Atualmente a imputação múltipla está disponível nos principais *softwares* estatísticos, porém a maioria dos métodos implementados são paramétricos, e nestes casos há fortes suposições sobre a distribuição dos dados, o que na prática é difícil de se verificar. Com vistas a promover a interdisciplinaridade em Bioestatística, tratamos aqui de dois procedimentos para realizar imputação múltipla os quais oferecem maior flexibilidade quanto à distribuição dos dados: o algoritmo MICE - *Multivariate Imputation by Chained Equatoins*- e o método IMLD - Imputação Múltipla Livre de Distribuição. O algoritmo MICE, é aplicado a dados de um estudo transversal de recém-nascidos vivos residentes no estado de Paraná, no ano de 2012. Uma amostra aleatória, com registros completos, de 3380 casos foi obtida, um modelo de regressão logística foi ajustado para o desfecho baixo peso ao nascer. Por simulação, foram gerados três conjuntos de dados incompletos, com dados faltantes para o desfecho peso ao nascer, categorizado em baixo peso e peso normal. Os modelos foram ajustados nas três situações distintas para comparação com o modelo padrão. Percebe-se, por meio das estimativas, um melhor ajuste dos modelos com imputação, quando comparado ao caso em que analisamos os dados com registros faltantes. As estimativas dos erros padrão do modelo imputado se aproximam muito bem dos resultados obtidos com o modelo ajustado ao conjunto de dados completo (modelo padrão ouro). Uma aplicação usando o método IMLD é feita com uma matriz $\boldsymbol{Y}$ de dados referente a altura média de plantas (m) de 20 cultivares precoces e geneticamente modificadas de milho, avaliadas em 7 localidades no estado do Paraná (SHIOGA et al., 2015). Remoções aleatórias (5%, 15%, 30%) foram feitas na matriz original e posteriormente empregado o método IMLD para preenchimento destes valores faltantes. A implementação do método foi feita no *software R*, a qual é disponibilizada em anexo. Por meio de medidas de variabilidade e acurácia, o método mostrou-se eficaz. Com isso, temos indícios de que a imputação múltipla deve ser uma opção a ser utilizada quando se tem dados faltantes.

Palavras-chaves: Dados faltantes; Decomposição por valor singular; Imputação múltipla livre de distribuição; Interdisciplinariedade; Métodos de imputação.

Abstract

A common problem in statistical analyzes is the occurrence of incomplete databases. Generally, in these situations, it restricts the analysis to subjects with complete data on the variables. This reduces the size of the sample, and can result in unbiased estimates. The "filling" of the missing data can be done by multiple imputation (IM), wherein each missing value is replaced by a set of plausible values, incorporating the uncertainty about the amount to be imputed. Currently, multiple imputation is available in the main statistical software, but most of the implemented methods are parametric, and in these cases there are strong assumptions about the distribution of data, which in practice is difficult to verify. In order to promote interdisciplinarity in Biostatistics, we treat here two procedures to perform multiple imputation which offer greater flexibility in the distribution of the data: the MICE algorithm - Multivariate imputation by Chained Equations - and IMLD method - Multiple Imputation Distribution Free. The MICE algorithm is applied to data from a cross-sectional study of newborns live residents in the Parana state, in the year 2012. A random sample with complete records of 3380 cases was obtained, a logistic regression model was fitted to the outcome of low birth weight. By simulation, it was generated three sets of incomplete data, with missing data for weight outcome. The models were adjusted in three different situations for comparison with the standard model. It can be seen through the estimates, a better adjustment of the models with imputation when compared to the case where we analyze the data with missing records. The estimates of imputed model standard errors of approaches very well the results obtained with the gold standard model. An application to the IMLD method is made with a array $\boldsymbol{Y}$ of data regarding the average plant height (m) of 20 early and genetically modified corn cultivars, evaluated in seven locations in the Parana state (SHIOGA et al., 2015). Random removals (5%, 15%, 30%) were made in the original array and then used the method IMLD to fill these missing values. The implementation of the method was taken in the R software, which is provided in annex. Through variability and accuracy measurements, the method proved to be effective. With this, we have evidence that multiple imputation should be an option to be used when there is missing data.

Key-words: Decomposition by singular value; Distribution free multiple imputation; Interdisciplinarity; Imputation methods; Missing data.

Lista de ilustrações

Lista de Quadros

*

Lista de tabelas

Lista de abreviaturas e siglas

ACC	Análise de Casos Completos
BPN	Baixo Peso ao Nascer
DVS	Decomposição por Valores Singulares
IAPAR	Instituto Agronômico do Paraná
LOCF	Imputação via última observação realizada
IM	Imputação Múltipla
IMLD	Imputação Múltipla Livre de Distribuição
MAR	Perdas ao Acaso
MCAR	Perdas Completamente ao Acaso
NMAR	Perdas Não-Aleatórias
MICE	*Multivariate Imputation by Chained Equatoins*
MV	Máxima Verossimilhança
PNN	Peso Normal ao Nascer
OMS	Organização Mundial da Saúde
SINASC- PR	Sistema de Informações sobre Nascidos Vivos do Estado do Paraná

Sumário

INTRODUÇÃO

É frequente em pesquisas científicas a ocorrência de valores faltantes (*missing data*) nas bases de dados. *Missing data* é um dado que existe, e de interesse para o estudo, mas que por algum motivo não foi possível observá-lo. São várias as circunstâncias que podem gerar bancos incompletos, como a informação não ser fornecida pelo entrevistado, algumas medidas não estarem disponíveis no momento da coleta devido a morte de alguns animais ou plantas estarem danificadas (BERGAMO, 2007), ainda valores são perdidos em consequência de falhas advindas da etapa de mensuração das características de interesse (SCHAFER; GRAHAM, 2002). A condição ideal nestes casos seria a repetição do estudo para obter valores novos e suprir os dados ausentes, mas na prática isso geralmente é inviável, devido a recursos financeiros e/ou ao tempo limitado.

São diversas as complicações ocasionadas pela falta de dados, por exemplo, redução do tamanho da amostra, perda de eficiência, complicações na manipulação e análise dos dados e viés nas estimativas quando a análise com valores perdidos é feito inadequadamente (ROTH; SWITZER; SWITZER, 1999). Existem técnicas de análise específicas para que a inferência com *missing data* seja válida, porém, apesar do crescente desenvolvimento metodológico nesta área, é comum encontrar o uso de inadequadas metodologias para a análise de dados faltantes.

Os *softwares* estatísticos comumente utilizados possuem como padrão um procedimento chamado de exclusão *listwise* ou análise de casos completos (ACC) (BUUREN, 2012). O processo elimina todos os casos com um ou mais valores ausentes nas variáveis e assim, restringe-se a análise aos casos completamente observados (ENDERS, 2010). A grande vantagem do processo de análise completa é a conveniência, entretanto estimativas viesadas podem ser produzidas induzindo à tomadas de decisões errôneas (RAGHUNATHAN, 2004).

As primeiras técnicas estatísticas destinadas a repor dados faltantes por valores verossímeis a eles, surgiram no final da década de 70, e foram denominadas métodos de imputação (RUBIN, 1976). Dentre estas técnicas tem-se o preenchimento dos dados faltantes pela média ou pela mediana da variável, imputação por meio do vizinho mais próximo, imputação

hot deck, imputação pela regressão linear e imputação por meio da máxima verossimilhança. Nestas técnicas, chamadas de métodos de imputação simples, o dado ausente é preenchido uma única vez e então utiliza-se o banco de dados completo para as análises. O fato do *missing data* ser imputado somente uma vez, faz com que a incerteza associada ao procedimento não seja agregada às estimativas geradas pelo banco completo, trazendo uma grande limitação a estes métodos (ENDERS, 2010).

Ao encontro da necessidade de controlar o viés associado à imputação simples, Rubin (1978) propôs um novo método, a Imputação Múltipla (IM), a qual é composta de três etapas principais: imputação, análise e agrupamento. Cada valor ausente é substituído por um conjunto de valores plausíveis, o que representa a incerteza sobre o valor a ser imputado. As publicações de artigos (RUBIN, 1976),(RUBIN, 1978) e livros (RUBIN, 1987), (LITTLE; RUBIN, 1987), feitas pelo precursor Donald Bruce Rubin, serviram de apoio para que mais pesquisadores desenvolvessem trabalhos com a imputação múltipla, entretanto, apenas recentemente esta técnica vem sendo utilizada com mais tenacidade, devido aos desenvolvimentos computacionais para sua implementação.

Atualmente a imputação múltipla está disponível nos principais *softwares* estatísticos comerciais ou gratuitos, como *R*, *SAS* e *Stata*. Uma boa revisão dessas implementações é apresentada por Horton e Kleinman (2007), comparando resultados e fornecendo também para cada software, instruções e sintaxes utilizadas nas análises. Entretanto a maioria dos métodos de IM implementados são paramétricos, e acabam por incorporar suposições sobre a distribuição dos dados, que na prática com dados reais nem sempre são satisfeitas (BUUREN, 2012), criando uma limitação na aplicabilidade dos métodos de imputação (BUUREN, 2012).

Perante a dificuldade em se atender as suposições exigidas por métodos paramétricos, e com vistas a promover a interdisciplinaridade em Bioestatística, tratamos aqui de dois procedimentos para realizar imputação múltipla os quais oferecem maior flexibilidade quanto à distribuição dos dados: tratamos aqui de dois procedimentos para realizar imputação múltipla, propostos recentemente, que oferecem maior flexibilidade quanto a distribuição dos dados. O algoritmo MICE - *Multivariate Imputation by Chained Equatoins* e o método IMLD - Imputação Múltipla Livre de Distribuição.

O algoritmo MICE, Buuren e Oudshoorn (2000) e Buuren e Groothuis-Oudshoorn (2011), é um método de cadeia de Markov Monte Carlo (MCMC). Buuren (2012) apresenta várias possibilidades de imputação simples e múltipla com o algoritmo MICE , por meio do pacote `mice` do *software* R. Já a IMLD proposta por Bergamo (2007) trata-se de uma técnica para a primeira fase da imputação múltipla, sem suposição sobre a distribuição ou estrutura dos dados, utilizando a decomposição por valor singular (DVS) em matriz de interação.

Diante da recorrência frequente à questão de dados faltantes por pesquisadores de todas as áreas, é notória a importância de estudar, entender e dissipar a metodologia correta à ser adotada.

O trabalho está organizado em cinco partes. Na primeira é feita uma introdução ao temas abordados, na segunda são apresentados os principais conceitos da teoria de dados faltantes, uma breve apresentação de métodos de imputação simples e descrição da técnica de imputação múltipla. Na terceira parte é explorada a imputação múltipla por meio do algoritmo MICE, aplicando-o em dados de recém-nascidos do estado do Paraná. Na quarta parte é introduzida a Imputação Múltipla Livre de Distribuição (IMLD), fazendo uma aplicação em dados de alturas médias de plantas, avaliadas em 20 cultivares precoces geneticamente modificadas de milho. No quinto capítulo são apresentadas as considerações finais sobre o trabalho desenvolvido. Anexo (Anexo D), apresenta-se o artigo "Multiple Imputation in Logistic Regression Models: factors associated to the low weight at birth in the Parana State", submetido à Revista Acta Scientiarum. Health Sciences, ISSN 1679-9291 (impresso) e ISSN 1807-8648 (on-line), publicada semestralmente pela Editora da Universidade Estadual de Maringá-Eduem.

Capítulo 1

CONCEITOS E DEFINIÇÕES

Bancos de dados com *missing data* são recorrentes em todos os tipos de estudo, e de acordo com Buuren (2012) existe uma distinção entre dois tipos de dado ausente: falta de dados *intencionais* e *não intencionais*. Os dados em falta intencionais são planejadas pelo coletor de dados, como por exemplo, uma unidade pode estar em falta devido a exclusão desta unidade da amostra feita pelo próprio pesquisador ou também em dados de tempo de sobrevivência que são censurados em algum tempo (GONZALEZ; ELTINGE, 2007).

Os dados em falta não intencionais, mesmo que muitas vezes sejam previstos, não são planejados e o coletor de dados não possui controle sobre eles. Como exemplos pode-se ter um entrevistado pulando um item do questionário, erro na transcrição dos dados, indivíduos desistirem antes do estudo ser concluído ou ainda o sujeito foi amostrado mas se recusou a cooperar com a pesquisa.

Uma outra importante discriminação é feita quanto a item não "respondido" e "unidade não respondente". O primeiro diz respeito a uma situação em que o entrevistado ignora um ou mais itens do estudo. Enquanto que o segundo acontece com a recusa do entrevistado em participar, e assim todos os resultados ficam em falta para este indivíduo (BUUREN, 2012).

São vários os conceitos da teoria de dados faltantes que devem receber atenção ao se aplicar algum método de imputação, pois como alegado por Rubin e Little (2002) "uma imputação sem critérios pode criar mais problemas do que resolver, distorcendo estimativas, erros padrão e testes de hipóteses". Dentre os importantes conceitos destacam-se os padrões e os mecanismos dos dados ausentes. Os mecanismos descrevem possíveis relações entre as variáveis medidas e a probabilidade de dados em falta, enquanto que os padrões se referem à forma com que as unidades ausentes estão distribuídas em um conjunto de dados (ENDERS, 2010).

1.1 MECANISMOS DE DADOS FALTANTES

A ocorrência de *missing data* em bases de dados normalmente obedece a um mecanismo que aponta as condições de geração dos dados ausentes. A principal terminologia de classificação dos mecanismos foi criada por Rubin (1976), e define três mecanismos teóricos gerais extensamente utilizados na literatura :

- Perdas Completamente ao Acaso (*Missing Completely at Random* – MCAR);
- Perdas ao Acaso (*Missing at Random* – MAR);
- Perdas Não-Aleatórias (*Not Missing at Random* - NMAR).

O termo mecanismo é utilizado como sinônimo da função de distribuição dos dados ausentes, que define a probabilidade de cada valor ser ou não observado, e os fatores associados a essa probabilidade.

Para tratar de cada mecanismo, considere $\boldsymbol{Y}$ uma matriz de dados coletados com m linhas, as quais representam os indivíduos, n colunas, que representam as variáveis, com $y_{ij} = (y_{i1}, ..., y_{in})$, em que y_{ij} é o valor da variável j para o indivíduo i. Pode-se dividir $\boldsymbol{Y}$ em dois subconjuntos $\boldsymbol{Y} = \{\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}\}$ em que $\boldsymbol{Y}_{obs}$ são os dados observados (não-faltantes) e $\boldsymbol{Y}_{mis}$ são os dados faltantes. Correspondente a matriz de dados $\boldsymbol{Y}$, existe $\boldsymbol{R}$, uma matriz indicadora de respostas para cada item de $\boldsymbol{Y}$, $\boldsymbol{R}$ possui a mesma dimensão de $\boldsymbol{Y}$, sendo $r_{ij} = 1$, se y_{ij} é observado, e $r_{ij} = 0$, caso contrário (BUUREN, 2012).

A distribuição de $\boldsymbol{R}$ pode depender de $\boldsymbol{Y} = \{\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}\}$, seja pelo planejamento ou por acaso, e essa relação é descrita pelo modelo de dados faltantes (*missing data model*). A expressão geral do modelo do *missing data model* é:

$$Pr(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\psi})$$

sendo que $\boldsymbol{\psi}$ contém os parâmetros desconhecidos do modelo.

1.1.1 MCAR

Os dados são referidos como MCAR se $Pr(\boldsymbol{R} = 0|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\psi}) = Pr(\boldsymbol{R} = 0|\boldsymbol{\psi})$ o que significa que a probabilidade de um item ter respostas ausentes não depende de qualquer das quantidades,observadas ou não observadas. Implicando que a probabilidade de ocorrência do dado ausente é a mesma para todos os casos. Quando o mecanismo MCAR ocorre, os dados não observados constituem uma sub amostra aleatória dos dados observados.

1.1.2 MAR

Os dados são ditos MAR se $Pr(\boldsymbol{R} = 0|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\psi}) = Pr(\boldsymbol{R} = 0|\boldsymbol{Y}_{obs}, \boldsymbol{\psi})$ ou seja, os dados faltantes dependem apenas de informações observadas, disponíveis para análise e correlacionadas com a variável que possui *missing data*.

1.1.3 MNAR

Tem-se dados MNAR se $Pr(\boldsymbol{R} = 0|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\psi})$ nesta situação a probabilidade de *missing data* depende também das informações não observadas, isto é, a probabilidade de se ter um dado faltante varia por razões que são desconhecidas.

Uma situação que exemplifica cada mecanismo é dada por Bergamo (2007), tendo como base uma pesquisa que estuda o peso de pessoas. Se a ausência de resposta da varável peso, por exemplo, não está relacionada com o próprio peso do entrevistado e nem com qualquer outra variável, como a idade ou sexo, então o mecanismo de ausência de valores para o peso é MCAR. Se as pessoas com sobrepeso tendem a não informar seu peso, a ausência de resposta sobre o peso depende do próprio peso, caracterizando o mecanismo de ausência MNAR. No entanto, se a ausência de resposta sobre o peso não depende do próprio peso, mas pode depender de outras variáveis (pessoas do sexo feminino tendem a não informar seu peso), diz-se que o mecanismo de ausência de valores para o peso é MAR.

1.2 MECANISMO IGNOVÁREL E NÃO-IGNOVÁREL

O termo Ignorável é usado para indicar que não é necessário especificar um modelo para a não-resposta. A função da densidade conjunta de $\boldsymbol{Y}_{obs}$ e $\boldsymbol{R}$, $f(\boldsymbol{Y}_{obs}, \boldsymbol{R}|\boldsymbol{\theta}, \boldsymbol{\psi})$, depende dos parâmetros $\boldsymbol{\theta}$ para o banco completo $\boldsymbol{Y}$. A densidade conjunta é proporcional a verossimilhança de $\boldsymbol{\theta}$ e $\boldsymbol{\psi}$, isto é, $L(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}_{obs}, \boldsymbol{R}) \propto f(\boldsymbol{Y}_{obs}, \boldsymbol{R}|\boldsymbol{\theta}, \boldsymbol{\psi})$.

O mecanismo é ignorável para a inferência de verossimilhança se temos MAR ou MCAR, ou seja, os dados faltantes são ao acaso e se os parâmetros $\boldsymbol{\theta}$ e $\boldsymbol{\psi}$ são distintos, no sentido que o conjunto de espaços dos parâmetros $(\boldsymbol{\theta}, \boldsymbol{\psi})$ é o produto do espaço do parâmetro $\boldsymbol{\theta}$ e do espaço do parâmetro $\boldsymbol{\psi}$. Para uma inferência Bayesiana válida, esta última condição é mais rígida: $\boldsymbol{\theta}$ e $\boldsymbol{\psi}$ devem ser a *priori* independentes, $p(\boldsymbol{\theta}, \boldsymbol{\psi}) = p(\boldsymbol{\theta})p(\boldsymbol{\psi})$. O mecanismo (MCAR) possui uma suposição muito forte e raramente é satisfeito na prática, enquanto que MNAR é não-ignorável devido à falta de aleatoriedade da não-resposta. Portanto, nesta situação torna-se necessário especificar um modelo para a não-resposta.

Em um exemplo apresentado por Rubin (1987, p. 04), ficam claras as desvantagens de se fazer uma análise ignorando os dados faltantes ou unidades não respondestes do estudo. Em 1971 o Educational Testing Service (ETS) realizou uma pesquisa com amostra de 660 escolas, objetivando estudar os seus programas de leitura compensatórias. Informações sobre os tipos

de programas de leitura compensatórias e os níveis dos alunos nas escolas deviam ser obtidas a partir de um questionário enviado aos diretores. Ao final do inquérito, apenas 472 dos 660 diretores haviam retornado este questionário. Uma vez que os diretores sabiam que o objetivo da pesquisa era estudar os seus programas de leitura, a preocupação desenvolvida foi que as 188 escolas não-respondentes poderiam diferir das 472 respondentes. Talvez tendo os alunos com problemas de leitura mais graves ou programas de leitura que não foram tão eficazes. Com base nas informações de um censo que continha informações gerais de todas as escolas, verificaram que as escolas respondentes deferiam sistematicamente das não respondentes em características consideradas relevantes. Como a prática comum seria simplesmente descartar as 188 escolas com dados incompletos, nesse caso, os resultados seriam distorcidos, pois, por exemplo, se as escolas não respondentes possuíam alunos de baixo desempenho, análises baseadas somente em respondentes iriam superestimar o comportamento típico de estudantes em programas de leitura, podendo levar a tomadas de equivocadas e políticas públicas impróprias.

O impacto de cada mecanismo nas análises produzidas por diferentes métodos tem sido majoritariamente avaliado por estudos de simulação, como em Little (1992), Collins, Schafer e Kam (2001) e Schafer e Graham (2002). Para os mecanismo MCAR ou MAR muitos métodos de tratamento de dados ausentes têm sido aplicados, no entanto, para o padrão NMAR métodos apropriados estão em aberto.

Uma importante ressalva deve ser feita, a de que o mecanismo de dados ausentes que pode ser ignorado e não os dados ou as unidades com dados ausentes (PEREIRA, 2014).

1.3 TESTES PARA MECANISMOS

Identificar o mecanismo de ausência de dados não é uma tarefa simples. Vários testes tem sido propostos para testar MCAR versus MAR e em sua maioria possuem como base os dois testes apresentados a seguir.

Em Jamshidian, Jalal e Jansen (2014) é apresentado o pacote *MissMech* no software R, no qual estão implementados os métodos para testar a hipótese de MCAR, propostos por Jamshidian e Jalal (2010). O enfoque principal do pacote é testar MCAR, mas ele executa outras tarefas, testes de normalidade multivariada, imputação de *missing*, testes para homocedasticidade e normalidade para dados completos, obtenção de estimativas de máxima verossimilhança da média e covariância (incluindo erros padrão), para dados incompletos, utilizando o algoritmo EM, entre outras.

Jamshidian e Mata (2008) usam os dados disponíveis para examinar a sensibilidade de um determinado modelo ao mecanismo de dados em falta, para quando o pesquisador não possuir MCAR (ou MAR). Os autores fornecem um método específico para executar análise de sensibilidade *postmodeling* usando um teste estatístico e gráficos.

1.3.1 Teste t-Univariado

O método mais simples para avaliar MCAR é a utilização de uma série de Testes-t independentes para comparar os subgrupos de dados em falta, relatado por Brown e Dixon (1983). Esta abordagem separa os casos omissos e completos em uma determinada variável e usa o teste para examinar se existe diferença significativa dos grupos nas outras variáveis do conjunto de dados.

O mecanismo MCAR implica que os casos com *missing* pertencem à mesma população dos casos com dados completos e, portanto, dividem o mesmo vetor de média e a matriz de covariância.

Consequentemente, um Teste-t não significativo fornece evidências de que os dados são MCAR, enquanto que uma estatística-t significativa (ou uma grande diferença média) sugere que os dados são MAR ou MNAR.

1.3.2 Little's MCAR Teste

É uma extensão multivariada da abordagem do Teste-t proposta por Little (1988), que avalia simultaneamente diferenças médias de cada variável no conjunto de dados.

Como a abordagem do Teste-t, o teste de Little avalia diferenças médias entre os subgrupos de casos que compartilham o mesmo padrão de dados em falta. A estatística do teste é uma soma ponderada das diferenças padronizadas entre as médias dos subgrupos e as grandes médias, como segue:

$$d^2 = \sum_{j=1}^{J} n_j(\hat{\mu}_j - \hat{\mu}_j{}^{(ML)})^T \hat{\sum}_j^{-1} (\hat{\mu}_j - \hat{\mu}_j{}^{(ML)})$$

em que n_j é o número de casos com falta de dados com padrão j, $\hat{\mu}_j$ contém a média da variável para os casos com falta de dados padrão j, $\hat{\mu}_j{}^{(ML)}$ contém estimativas de máxima verossimilhança das grandes médias, e $\hat{\sum}_j$ é a estimativa de máxima verossimilhança da matriz de covariância.

A estatística d^2 é aproximadamente distribuída como uma distribuição Qui-quadrado (χ^2) com $\sum k_j - k$ graus de liberdade, onde k_j é o número de variáveis completas para o padrão j, e k é o número total de variáveis. Coerente com a abordagem do Teste-t univariado, uma estatística significativa d^2 fornece evidências contra MCAR.

A função *LittleMCAR(x)* no software R, parte do pacote *BaylorEdPsych*, Beaujean e Beaujean (2012), usa o "Teste de Little" para avaliar a falta completamente aleatória (MCAR) em dados multivariados.

1.4 PADRÕES DE DADOS AUSENTES

É muito importante verificar a forma com que ocorre a ausência dos dados, isso pode ser feito por meio dos padrões de comportamento dos dados faltantes, os quais descrevem a localização dos valores em falta (SILVA, 2012). Os padrões mais frequentes estão descritos a seguir e representados na Figura 1.

Padrão univariado: apresenta uma falta de dados isoladamente em uma variável, o que é comum em estudos experimentais.

Padrão de item não respondido: ocorre muito em pesquisas realizadas por meio de questionários, em que alguns itens são respondidos pelos indivíduos e outros são recusados, causando valores em falta para questionários com item sem resposta.

Padrão monótono: bastante presente em pesquisas clínicas, onde os indivíduos participantes da pesquisa em algum momento não podem continuar no estudo devido à alguns fatores, por exemplo, reação de alguma droga em análise. Este tipo de padrão de dados em falta é característico de experimentos longitudinais.

Padrão geral: padrão conhecido como arbitrário que consiste numa dispersão de unidades ausentes por toda a matriz de dados. Aparentemente é aleatório, porém pode existir uma relação entre a falta de valores de uma variável e a tendência da falta de dados referente à outra variável medida.

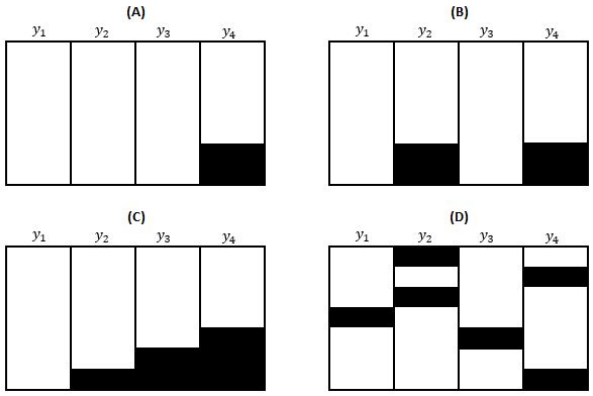

Figura 1 – Padrões de *missing*. As áreas sombreadas representam a localização dos valores em falta num conjunto de dados com quatro variáveis em estudo. (A) Padrão univariado; (B) Padrão de item não respondido; (C) Padrão monótono; (D) Padrão geral.

1.5 MÉTODOS DE IMPUTAÇÃO

A imputação é o preenchimento dos dados ausentes com valores plausíveis para uma posterior análise dos dados completos. Ela pode ser simples, quando somente um valor é colocado para cada dado ausente, ou múltipla, quando há mais de um valor para cada dado ausente.

Constantemente, artigos de revisão e tutoriais vêm sendo publicados, apoiando os pesquisadores no tratamento de dados faltantes. Schafer e Graham (2002) e Little e Rubin (2014) apresentam uma ampla revisão dos métodos existentes para lidar com valores perdidos, apontando as condições em que os mesmos produzem resultados válidos. Outros trabalhos de revisão são encontrados Myers (2000) e Tsikriktsis (2005). Em Scheffer (2002), Acock (2005) e Buhi, Goodson e Neilands (2008) são comparadas algumas das técnicas disponíveis nos principais pacotes estatísticos.

Os estudos comparativos apontam os métodos de Máxima Verossimilhança(MV) e Imputação Múltipla(IM) como preferenciais para a análise de bases incompletas, já que ambos, além de utilizar todas as informações coletadas, devem produzir resultados válidos sob condições menos restritas que os demais (SCHAFER; GRAHAM, 2002),(RAGHUNATHAN, 2004). Esses estudos discutem ainda vantagens da IM sobre o método de MV em relação à praticidade de aplicação e disponibilidade, já que apenas o primeiro encontra-se implementado na maioria dos softwares de análise tradicionais (HORTON; KLEINMAN, 2007).

Na literatura, são vários os autores que fazem uma reunião dos diversos métodos de imputação (BRAND, 1999; ENDERS, 2010; LITTLE; RUBIN, 1987; LITTLE; RUBIN, 2014; MOLENBERGHS; KENWARD, 2007; SCHAFER, 1997). A seguir são apresentados os mais usuais.

1.5.1 Exclusão *listwise*

Análise de caso completo *listwise* é o método padrão para lidar com dados incompletos em muitos pacotes estatísticos, incluindo SPSS, SAS e Stata, S-PLUS e R. O procedimento elimina todos os casos com um ou mais valores ausentes nas variáveis de análise.

A grande vantagem do processo de análise completa é a conveniência. Sob MCAR, exclusão *listwise* produz erros padrão e níveis de significância que estão corretos para a redução do subconjunto de dados, mas que são muitas vezes maiores em relação a todos os dados disponíveis.

1.5.2 Exclusões *Pairwise*

As exclusões *Pairwise*, também conhecidas como análise de casos disponíveis, são tentativas para resolver o problema das perda de dados das exclusões *listwise*. O método calcula

as médias e as variâncias em todos os dados observados. Assim, a média da variável X é baseada em todos os casos com os dados observados em X, a média da variável Y utiliza todos os casos com valores de Y observados, e assim por diante. Para as correlações e covariância, todos os dados são tomados em que ambos X e Y possuem escores não ausentes.

1.5.3 Imputação pela média

Imputação pela média é uma solução rápida e simples para os dados em falta. No entanto, alguns problemas podem ocorrer: subestimação da variância; perturbação nas relações entre as variáveis; viés em quase qualquer estimativa diferente da média. A imputação pela média só deverá ser utilizada como uma correção rápida quando um pequeno número de valores estiverem faltando, e deve ser evitada em geral.

1.5.4 Imputacão via Regressão

A imputação via regressão incorpora conhecimento de outras variáveis com a ideia de produzir imputações mais inteligentes. A primeira etapa envolve a construção de um modelo a partir dos dados observados. Previsões para os casos incompletos são, então, calculadas de acordo com o modelo ajustado, e servem como substitutos para os dados faltantes.

1.5.5 Imputação via última observação realizada (LOCF)

A última observação realizada (LOCF) requer dados longitudinais . A ideia é levar o último valor observado como um substituto para os dados em falta. LOCF é conveniente porque gera um conjunto de dados completo .

1.5.6 Imputação via método do indicador

Suponha que se quer utilizar uma regressão, mas faltam valores em uma das variáveis explicativas. O método do indicador substitui cada valor em falta por um zero e prolonga-se o modelo de regressão pelo indicador de resposta. O procedimento é aplicado a cada variável incompleta. Este método é popular em saúde pública e epidemiologia. Uma vantagem é que o método do indicador conserva o conjunto de dados completo.

1.5.7 Método de Imputação Múltipla

A imputação múltipla, Rubin (1987), é um tema em grande expansão na estatística. Sobre a imputação múltipla Buuren (2012) afirma:

> A técnica é simples, elegante e poderosa. É simples porque preenche os buracos nos dados com valores plausíveis. É elegante, pois a incerteza sobre os dados desconhecidos está codificada nos dados propriamente ditos. E

> é poderoso porque pode resolver "outros"problemas que são efetivamente problemas de dados faltantes disfarçados.

Na aplicação deste método dois importantes pressupostos precisam ser atendidos. Primeiro, os dados em falta devem seguir um mecanismo de ausência aleatório e segundo, é necessário que exista correlação entre a variável a ser imputada e o vetor de covariáveis que será utilizado para modelar os dados a serem preenchidos.

1.5.7.1 As fases da Imputação Múltipla

O método de imputação múltipla se resume a três etapas principais, imputação, análise e agrupamento. Um esquema resumido destas etapas é apresentado na Figura (2).

(i) A análise começa com dados observados e dados incompletos. A imputação múltipla cria $m > 1$ versões completas dos dados, substituindo os valores em falta por valores de dados plausíveis. Estes valores plausíveis são extraídos de uma distribuição modelada especificamente para cada entrada de dados faltantes. Os conjuntos de dados imputados são idênticos para as entradas de dados observados, mas diferem nos valores imputados. A magnitude destas diferenças reflete nossa incerteza sobre o valor a imputar.

(ii) O segundo passo é estimar os parâmetros de interesse de cada conjunto de dados imputados, por meio da aplicação de métodos de análises padrão para dados completos. Os resultados serão diferentes porque os seus dados de entrada são diferentes, o que é apenas resultado da incerteza sobre o valor a imputar.

(iii) No último passo os m resultados são agrupados em uma estimativa pontual final acrescidos do desvio padrão, por regras de agrupamento simples, conhecidas como "Regras de Rubin"(RUBIN, 1987).

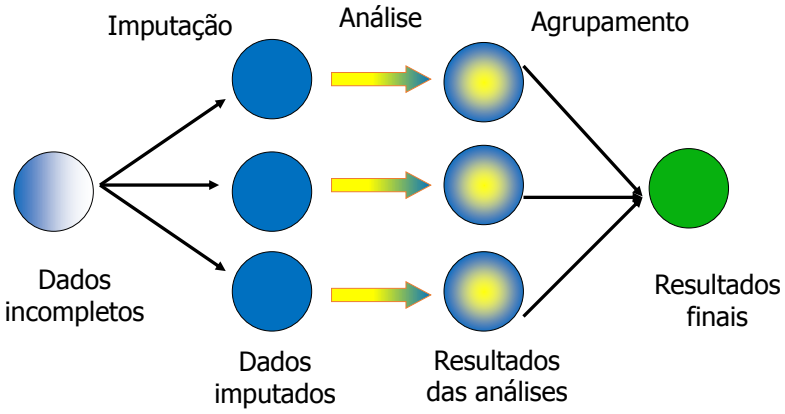

Figura 2 – Principais etapas da imputação múltipla.

O primeiro passo requer uma atenção especial, pois é decisivo para a validade dos resultados produzidos pelas análises posteriores. O pesquisador deve definir as variáveis que farão parte do modelo de imputação, e o tipo de modelo que melhor se ajusta à distribuição da variável com *missing* (X). Buuren et al. (1999) propõe uma estratégia geral para a seleção das variáveis, que consiste em: incluir todas as variáveis que serão utilizadas em análises conjuntas com X; incluir variáveis associadas com as perdas; incluir variáveis preditoras da variável X; excluir aquelas variáveis que apresentam uma elevada proporção de perdas onde X é ausente. O modelo utilizado na etapa de imputação não precisa ser o mesmo da etapa de análise, pois nem sempre o modelo utilizado para imputar é o mais adequado para analisar (BARACHO, 2003).

Outro ponto a ser considerado na imputação múltipla é a escolha do número de imputações (m). A escolha de um m (bancos completos gerados) pequeno pode inflacionar o intervalo de confiança das estimativas e consequentemente reduzir o poder das análises (GRAHAM; OLCHOWSKI; GILREATH, 2007). Rubin (1978) quantifica essa inflação para diferentes frações de informação ausente e escolhas de m. O conceito de fração de informação ausente não é o mesmo da proporção de dados ausentes, mas seu valor tende a ser igual ou inferior à proporção de dados ausentes (RUBIN, 1987). Molenberghs e Verbeke (2006) relatam a alta eficácia da imputação múltipla, que mesmo com valores pequenos de m, como por exemplo, de 3 a 5 imputações são suficientes para bons resultados serem obtidos.

1.5.7.2 Fases de Análise e Agrupamento

Na fase de análise vários conjuntos de estimativas de parâmetros e erros padrão são produzidos. O objetivo posterior é combinar todas as análises dos m conjuntos de dados completos, em um único conjunto de resultados. Fórmulas simples para esta etapa foram descritas por Rubin (1987).

O agrupamento das estimativas é dado por:

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i \tag{1.1}$$

em que $\hat{Q}_i$ é a estimativa do i -ésimo parâmetro considerado, correspondente aos seu conjunto(m) de dados imputados.

A combinação dos erros padrão envolve duas fontes de variação:

- **A variância dentro das imputações**: Definida como média aritmética das m variâncias amostrais descrita como:

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m}\hat{U}_i \tag{1.2}$$

sendo $\hat{U}_i$ a variância do m-ésimo conjunto de dados imputados.

- **A variação entre imputações:** Que quantifica a variabilidade de uma estimativa em todas as m imputações, sendo simplesmente, a variância do parâmetro estimado em todas as m imputações.

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q})^2. \tag{1.3}$$

Após o cálculo das estimativas combinadas $\bar{Q}, \bar{U}$ e B, o próximo passo é a obtenção da variância combinada total descrita por:

$$T = \bar{U} + (1 + \frac{1}{m})B \tag{1.4}$$

sendo $(1 + \frac{1}{m})$ a correção de números infinitos de imputações.

Em seguida pode-se realizar testes de hipóteses e construir intervalos de confiança para a média $(\bar{Q})$ por meio de uma aproximação para a distribuição $t - Student$, ou seja,

$$\frac{(\bar{Q} - Q)}{SE} \sim t_{v_m} \tag{1.5}$$

com $v_m = (m-1)\left[\frac{1+\bar{U}}{(1+m)^{-1}B}\right]^2$ graus de liberdades e Q o valor esperado da variável em estudo.

Uma medida para determinar o incremento relativo da variância, devido a presença das unidades ausentes também é apresentada por Rubin (1987),

$$r = \frac{(1+m)^{-1}B}{\bar{U}} \tag{1.6}$$

e uma taxa de unidades ausentes que se aproxima de

$$\lambda = \frac{r}{(1+r)}, \tag{1.7}$$

conhecida uma fração de falta de dados alta, o número de conjuntos de dados a serem imputados deve ser determinado com maior atenção.

Capítulo 2

O ALGORITMO MICE

O algoritmo MICE "*Multivariate Imputation by Chained Equation*" (BUUREN; OUDSHOORN, 2000, 2011), é um método de cadeia de Markov Monte Carlo (MCMC), onde o espaço de estado é o conjunto de todos os valores imputados. Mais especificamente, se as condicionais são compatíveis, o algoritmo MICE é um amostrador de *Gibbs*, uma técnica de simulação Bayesiana que amostra as distribuições condicionais a fim de obter amostras da distribuição conjunta. Em aplicações convencionais do amostrador *Gibbs*, as distribuições condicionais completas são derivadas da distribuição de probabilidade conjunta. No algoritmo MICE, as distribuições condicionais estão sob o controle direto do usuário, e assim a distribuição conjunta só é implicitamente conhecida, o que parece indesejável de um ponto de vista teórico (já que não sabemos a distribuição conjunta para que o algoritmo converge), mas na prática, não parece prejudicar as aplicações úteis do método (BUUREN, 2012).

Implementado no *software* R, o pacote `mice`, é de tal forma que o usuário pode especificar um método de imputação para cada coluna de dados incompletos. O método de imputação leva um conjunto de preditores completos, e retorna uma única imputação para cada entrada em falta na coluna incompleta. Várias imputações são criadas por chamadas repetidas para a função (BUUREN; OUDSHOORN, 2000). Uma vantagem da abordagem variável por variável é que, para cada uma das variáveis, um modelo de imputação diferente pode ser utilizado. Portanto, um conjunto de dados pode ter variáveis tanto contínuas como categóricas (OUDSHOORN; BUUREN; RIJCKEVORSEL, 1999).

A biblioteca `mice` fornece uma série de métodos de imputação univariados embutidos, que são estão indicados no Quadro 2.1. Apesar da indicação de métodos padrões para cada situação, em casos especiais escolher um método diferente pode ser melhor (BUUREN; GROOTHUIS-OUDSHOORN, 2011).

A biblioteca `mice` define três classes de dados:

- `mids`: conjunto de dados do resultado da imputação múltipla;

Quadro 2.1 – Métodos de imputação embutidos no pacote `mice`

Método	**Descrição**	**Tipo de variável**
`norm`	Regressão linear Bayesiana (*Bayesian linear regression*)	Quantitativa
`norm.predict`	Valores preditos (*Predicted value*)	Quantitativa
`norm.nob`	Regressão estocástica (*Stochastic regression*)	Quantitativa
`norm.boot`	Imputação normal com *bootstrap* (*Normal imputation with bootstrap*)	Quantitativa
`2L.norm`	Modelo normal multinível (*Multilevel normal model*)	Quantitativa
`pmm`	Média preditiva correspondente (*Predictive mean matching*)	Quantitativa
`mean`	Incondicional imputação média (*Unconditional mean imputation*)	Quantitativa
`logreg`	Regressão logística (*Logistic regression*)	Qualitativa/Binária
`logreg.boot`	Regressão logística com *bootstrap* (*Logistic regression with*)	Qualitativa/Binária
`polyreg`	Regressão logística multinomial (*Polytomous logistic regression*)	Qualitativa/Nominal
`lda`	Análises discriminantes (*Discriminant analysis*)	Qualitativa/Nominal
`sample`	Amostra aleatória a partir dos valores observados (*Random sample from the available observed values*)	Quantitativa ou Qualitativa
`polr`	Modelo logito ordenado (*Ordered logit model*)	Qualitativa/Ordinal

- `mira`: análises dos dados completos imputados;
- `mipo`: agrupamento das análises da imputação múltipla;

O Quadro (2.2) contém uma breve descrição das principais funções na biblioteca *mice*.

Quadro 2.2 – Descrição das principais funções da biblioteca `mice`

Função	**Entrada**	**Saída**	**Descrição**
`md.pattern`	`data.frame`	`matrix`	resume o padrão dos dados faltantes
`mice`	`data.frame`	`mids`	cria conjunto de dados da imputação múltipla
`complete`	`mids`	`data.frame`	converte `mids` em dados completos
`lm.mids`	`mids`	`mira`	regressão linear para os dados imputados
`glm.mids`	`mids`	`mira`	modelo linear generalizado para dados imputados
`pool`	`mira`	`mipo`	agrupamento das análises repetidas

O algoritmo MICE é uma maneira conceitualmente simples, flexível e prática para gerar imputações múltiplas. Para cada variável incompleta o usuário pode escolher um conjunto

de preditores que serão utilizados para a imputação. Isto é útil para a imputação de grandes conjuntos de dados (PIRDAWD, 2007).

A ideia por trás do pacote `mice` é a que segue:

Começa com um sorteio a partir dos dados observados, e imputa os dados incompletos variável por variável. Uma iteração consiste de um ciclo passando por todo Y_j. O número de iterações T frequentemente é baixo, 5 ou 10. O algoritmo MICE gera várias imputações executando o processo a seguir M vezes paralelas.

Passos do algoritmo MICE para a imputação em dados multivariados

1. Especifica um modelo de imputação $P(Y_j^{mis}|Y_j^{obs}, Y_{-j}, R)$ para a variável Y_j com $j = 1, ..., p$;
2. Para cada j, começa o preenchimento das imputações Y_j^0 por aleatórias retiradas de Y_j^{obs};
3. Repete para $t = 1, ..., T$;
4. Repete para $j = 1, ..., p$;
5. Define $Y_{-j}^t = (Y_1^t, ..., Y_{j-1}^t, Y_{j+1}^{t-1}, ..., Y_p^{t-1})$ como os dados atualmente completos exceto Y_j.
6. Retira $\phi_j^t \sim P(\phi_j^t|Y_j^{obs}, Y_{-j}^t, R)$;
7. Retira imputações $Y_j^t \sim P(Y_j^{mis}|Y_j^{obs}, Y_{-j}^t, R, \phi_j^t)$;
8. Repete j;
9. E repete t.

2.1 APLICAÇÃO

Os dados são de um estudo tipo transversal que procurou identificar fatores de risco de baixo peso ao nascer, utilizando registros disponíveis no Sistema de Informações sobre Nascidos Vivos do Estado do Paraná (SINASC- PR), para o ano de 2012. Para manipular os registros foi necessário descompactar os arquivos por meio do programa *TABWIN*. Registros de 153.945 recém-nascidos vivos foram obtidos, destes retirou-se uma amostra aleatória de 3380 registros completos.

O tamanho da amostra foi calculado por meio de uma amostra piloto com n=3000, utilizada para obter uma aproximação das proporções do baixo peso ao nascer (BPN) e peso normal ao nascer (PNN), encontrada a proporção de $0,10$ para BPN estabeleceu-se um erro máximo tolerado de 1% e nível de confiança de 95%, obtendo n= 3381.

A partir desta amostra foram criados, por simulação, três bancos de dados incompletos, em que se excluiu, aleatoriamente pelo *software* R, 5%, 10% e 20% das observações da variável resposta "Peso ao nascer", categorizada conforme apresentado no Quadro 2.3. Como a perda de dados foi por meio de amostras aleatórias, a probabilidade dos dados da variável "Peso ao nascer" serem faltantes é a mesma para todos os indivíduos e, independe dos dados, o que caracteriza o mecanismo da não-resposta aleatório.

Para analisar os resultados da utilização do algoritmo MICE serão comparados os diferentes ajustes feitos usando-se o modelo logístico, com dados faltantes e dados imputados. Adotou-se como modelo padrão ouro o modelo logístico ajustado com a amostra original, formada somente por registros completos. As variáveis incluídas como preditoras da variável resposta, peso ao nascer, são em conformidade com alguns estudos em outras regiões do Brasil (CASCAES et al., 2008) (GIGLIO et al., 2005), (ZHUOFAN, 2011).

2.1.1 Modelo

O modelo logístico, (HILBE, 2009), é utilizado quando a variável resposta é qualitativa com dois resultados possíveis, tomando os valores 1 e 0, com probabilidades π e $1-\pi$, respectivamente. Y é uma variável de *Bernoulli* com parâmetro $E(Y)=\pi$. O modelo na sua forma usual é dado por:

$$Y_i = E(Y_i) + \epsilon_i$$

em que

$$E(Y_i) = \pi_i = \frac{exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}{1 + exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}$$

Uma transformação essencial no estudo dos modelos de regressão logística é a transformação *logit* cujo objetivo é linearizar o modelo, aplicando o logaritmo. Essa transformação

define-se como:

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$$

$$g(x) = ln\left(\frac{\frac{exp(\beta_0+\beta_1X_1+...+\beta_pX_p)}{1+exp(beta_0+\beta_1X_1+...+\beta_pX_p)}}{1-\frac{exp(\beta_0+\beta_1X_1+...+\beta_pX_p)}{1+exp(\beta_0+\beta_1X_1+...+\beta_pX_p)}}\right) = ln\left(\frac{\frac{exp(\beta_0+\beta_1X_1+...+\beta_pX_p)}{1+exp(\beta_0+\beta_1X_1+...+\beta_pX_p)}}{\frac{1}{1+exp(\beta_0+\beta_1X_1+...+\beta_pX_p)}}\right)$$

$$g(x) = ln(exp(\beta_0 + \beta_1X_1 + ... + \beta_pX_p)) = \beta_0 + \beta_1X_1 + ... + \beta_pX_p$$

Esta transformação é chamada *transformação logit* de $\pi(x)$ e a razão $\frac{\pi(x)}{1-\pi(x)}$ é chamada *Odds ration* (razão de chances), definida com a razão entre a chance de um evento ocorrer no grupo exposto e a chance de ocorrer no grupo não exposto.

A regressão logística foi utilizada nesta aplicação, pelo fato do desfecho ter sido considerado binário, 1 para Baixo peso ao nascer ($< 2500g$) e 0 para Peso normal ($\geq 2500g$), ponto de corte adotado com base na definição de baixo peso ao nascer da Organização Mundial da Saúde (OMS). Das variáveis estudadas, algumas delas foram categorizadas de acordo com a estrutura fornecida pelo SINASC, Quadro 2.3. No Quadro 2.4, estão indicadas as categorias tomadas como referências (*baseline*).

2.1.2 Metodologia

O modelo final obtido, para a variável desfecho Peso ao nascer, incluiu as seguintes variáveis preditoras: idade da mãe, estado civil da mãe, escolaridade da mãe, tipo de gravidez (número de bebês), semanas de gestação, número de consultas pré-natal e a mãe ser primípara (primeiro filho).

A imputação múltipla foi realizada por meio do pacote *Multivariate Imputation by Chained Equations* (MICE) do software *R*. O método de imputação escolhido foi o *polyreg*, por ser adequado para variáveis binárias com dois níveis, como no caso da variável "Peso"(BUUREN; OUDSHOORN, 1999).

Após a realização da imputação múltipla, os dados completados foram analisados por regressão logística, por meio da função *glm* disponível no pacote *stat* que já vem com a instalação básica do *R*. A função *glm* utiliza os métodos de máxima verossimilhança e escore de Fisher para a estimação dos parâmetros do modelo. As estimativas de todas as análises dos m conjuntos de dados completos foram combinadas em um único conjunto de resultados de acordo com as "regras de agrupamento de Rubin".

2.1.3 Resultados e discussões

Os resultados obtidos do modelo logístico, para a variável desfecho peso ao nascer,com a amostra de dados completos, com simulação de 5% de dados faltantes e 5% de imputação,

Quadro 2.3 – Descrição das variáveis utilizadas no modelo de regressão logística.

Variável	**Categoria**	**Descrição**
Peso	1	Baixo Peso ao Nascer (BNP): o recém-nascido pesou menos que 2500 g
	0	Peso Normal ao Nascer (PNN): o recém-nascido pesou 2500 g ou mais
Idade	(10 -15)	Mãe com idade entre 10 e 15 anos
	(16 - 20)	Mãe com idade entre 16 e 20 anos
	(21-29)	Mãe com idade entre 21 e 29 anos
	(30 – 39)	Mãe com idade entre 30 e 39 anos
	≥ 40	Mãe com 40 anos de idade ou mais
Estado Civil (Est.civ)	Solteira	Mãe com estado civil solteira
	Separada	Mãe com estado civil separada
	Casada	Mãe com estado civil casada
	União. Estável	Mãe com estado civil união estável
Primípara/Mais filhos (Prim/Mais Filhos)	Primípara	Primípara (1^{o} filho): mãe que não possuía filhos vivos e também não possuía filhos mortos
	Não Primípara	Mãe que já possuía filhos, vivos ou mortos
Tipo de Gravidez (Grav)	Única	Gravidez única
	Dupla	Gravidez dupla
	Tripla $(+)$	Gravidez tripla ou mais
Tipo de Gestação (Gest)	$<$31	Menos de 31 semanas de gestação
	32 a 36	De 32 a 36 semanas de gestação
	≥ 36	36 semanas de gestação ou mais
Escolaridade (Escol)	Nenhuma	Nenhum ano de estudo concluído
	1 a 3 anos	e 1 a 3 anos de estudos concluídos
	4 a 7 anos	De 4 a 7 anos de estudos concluídos
	8 a 11 anos	De 8 a 11 anos de estudos concluídos
	≥ 12	12 anos de estudos concluídos ou mais
Número de consultas durante o pré-natal	Nenhuma	Nenhuma consulta de pré-natal
	(01 - 03)	De 1 a 3 consultas de pré-natal
	(04 – 06)	De 4 a 6 consultas de pré-natal
	≥ 7	7 consultas de pré-natal ou mais

Quadro 2.4 – Categorias de referência (*baseline*).

Variáveis	**Categoria de Referência**
Idade	(21-29)
Estado Civil (Est.civ)	Casada
Escolaridade (Escol)	≥ 12
Gestação (Gest)	≥ 36
Consultas pré-natal (Cons)	≥ 7
Gravidez (Grav)	Única
Primípara/Mais Filhos (Prim Mais Filhos)	1^{o} Filho

estão apresentados na Tabela 1, com 10% na Tabela 2 e com 20% na Tabela 3.

O modelo logístico foi ajustado em 7 situações distintas, resultando assim, em 133 coeficientes estimados (sem considerar os 7 interceptos). Quando comparados os ajustes com *missing data* e dos dados com as imputações aos dados originais, observa-se, que de um modo geral, os valores das estimativas e seus respectivos erros padrão se assemelham mais ao modelo de referência quando realizada a imputação. Nas três situações de simulação de falta de dados, os modelos logísticos utilizando a imputação de dados apresentaram, em sua grande maioria, menor erro padrão, quando comparado com as estimativas de referência (amostra completa).

A variável preditora, Idade (16-19) no primeiro cenário, por exemplo, foi significativa a 5% no modelo padrão e deixa de ser no modelo em que há presença de dados faltantes, porém volta a ser significativa quando ajustado no modelo com imputação. Ainda no primeiro cenário, podemos destacar a variável "Est.civ Uniao.Est", para a qual tanto a estimativa, como o erro padrão e o valor-p do modelo com imputação são muito similares aos do modelo padrão, mostrando uma melhora em relação ao modelo com *missing*.

Em termos da *Odds ratio*, para a variável "Est.civ Uniao.Est", no modelo padrão, modelo com *missing data* e modelo com imputação, as *OR* estimadas são 1,1759; 1,2864 e 1,1817, respectivamente. Por se tratar de uma medida muito utilizada na área da saúde, em estudos que visam identificar fatores de risco para agentes ou patógenos que afetam a saúde da população, destaca-se o fato de que, mesmo com uma pequena porcentagem de valores faltantes (5%), em uma amostra consideravelmente grande, melhoras com a aplicação do método de imputação são perceptíveis.

Para mais exemplos da eficácia do método de imputação utilizado citam-se as variáveis: Idade ≥ 40, Escol Nenhuma, Escol (1 a 3), Prim Mais Filhos, no cenário de 10% de *missing data* e Idade(16-19), Gest $<$32, no cenário de 20%. Os resultados obtidos corroboram com os apresentados por Nunes, Klück e Fachel (2009) e Camargos et al. (2011).

Em algumas situações, como no caso da variável " Escol (4 a 7) ", no cenário de 5% de *missing data*, e "Escol (4 a 7) ", cenário com 20% de dados faltantes, não se verifica um bom ajuste com a imputação. Como as variáveis foram categorizadas, algumas ficaram com a quantidade muito baixa de casos, o que pode justificar a imprecisão das estimativas e erros padrão elevados.

Restringir a análise aos casos que possuem observações completas pode fazer com que preditoras importantes deixem de ser identificadas. Além disso, pode resultar em tamanhos de amostras menores que o planejado e ainda gerar modelos menos preditivos do que o caso com dados completos, com erros padrão maiores nos estimadores.

O método de imputação via algoritmo MICE está disponível no *software R*, além deste, outras técnicas de imputação múltipla já estão implementadas em softwares estatísticos convencionais. Portanto, recomenda-se que os pesquisadores ao analisarem seus dados não

Tabela 1 – Estimativas, erros padrão e valores-p dos modelos logísticos para dados completos, incompletos com 5% de dados faltantes e com imputação.

Variáveis preditoras	Estimativas, Erros Padrão e valores-p dos modelos logísticos ajustados		
	Modelo Padrão	Com 5% de dados faltantes	Com 5% de imputação
Idade (10-15	-0,2848 0,4753 0,5489	-0,3015 0,4815 0,5311	-0,2694 0,4835 0,5773
Idade (16-19)	-0,4439 0,2161 0,0399	-0,3620 0,2235 0,1052	-0,4804 0,2147 0,0252
Idade (30-39)	-0,3478 0,1917 0,0696	-0,2783 0,1991 0,1622	-0,2297 0,1934 0,2349
Idade ≥ 40	-0,6173 0,4532 0,1732	-0,6420 0,4586 0,1615	-0,5919 0,4543 0,1926
Est.civ Solteira	0,2243 0,1921 0,2430	0,1778 0,1973 0,3674	0,1955 0,1922 0,3091
Est.civ Separada	-0,5452 0,5180 0,2925	-0,5685 0,5194 0,2737	-0,5785 0,5172 0,2633
Est.civ Uniao.Est	0,1621 0,2168 0,4546	0,2518 0,2256 0,2644	0,1670 0,2181 0,4436
Escol Nenhuma	-2,2437 0,6757 0,00089	-2,0457 0,7548 0,0067	-2,1720 0,6767 0,0013
Escol (1 a 3)	0,0373 0,4756 0,9373	0,1426 0,5062 0,7781	0,1568 0,4778 0,7427
Escol (4 a 7)	-0,0319 0,2568 0,9009	0,0565 0,2679 0,8329	0,1993 0,2581 0,4399
Escol (8 a 11)	0,2405 0,2139 0,2610	0,2410 0,2202 0,2736	0,3547 0,2121 0,0944
Grav dupla	-2,3176 0,3347 <0,0001	-2,5591 0,3583 <0,0001	-2,6378 0,3439 <0,0001
Grav tripla (+)	-3,2366 1,0127 0,0013	-3,26599 1,0089 0,0012	3,2623 1,0197 0,0013
Gest <32	-4,0226 0,3846 <0,0001	-3,9522 0,3914 <0,0001	-3,8989 0,3787 <0,0001
Gest [32 a 36]	-2,4398 0,1593 <0,0001	-2,4432 0,1636 <0,0001	-1,9183 0,6435 0,0028
Cons Nenhuma	-0,6199 0,7122 0,3841	-1,2910 0,8365 0,1227	-1,9183 0,6435 0,0028
Cons (1-3)	-0,9687 0,3043 0,0014	-0,8596 0,3199 0,0072	-0,8971 0,3070 0,0034
Cons (4-6)	-0,6094 0,1803 0,0007	-0,6526 0,1848 0,0004	-0,6123 0,1811 0,0007
Prim Mais Filhos	0,4930 0,1741 0,0046	0,4661 0,1798 0,0095	0,4688 0,1740 0,0070

Tabela 2 – Estimativas, erros padrão e valores-p dos modelos logísticos para dados completos, incompletos com 10% de dados faltantes e com imputação.

Variáveis preditoras	Estimativas, Erros Padrão e valores-p dos modelos logísticos ajustados		
	Modelo Padrão	Com 10% de dados faltantes	Com 10% de Imputação
Idade (10-15	-0,2848 0,4753 0,5489	-0,1815 0,5053 0,7194	-0,2977 0,4984 0,9523
Idade (16-19)	-0,4439 0,2161 0,0399	-0,2964 0,2308 0,1991	-0,37681 0,2157 0,0807
Idade (30-39)	-0,3478 0,1917 0,0696	-0,2091 0,2058 0,3096	-0,2524 0,1948 0,1950
Idade ≥ 40	-0,6173 0,4532 0,1732	-0,5279 0,4920 0,2833	-0,6713 0,4529 0,1383
Est.civ Solteira	0,2243 0,1921 0,2430	0.1635 0.2029 0.4203	0,1056 0,1912 0,5806
Est.civ Separada	-0,5452 0,5180 0,2925	-0,6139 0,5526 0,2666	-0,6255 0,5184 0,2275
Est.civ Uniao.Est	0,1621 0,2168 0,4546	0,2818 0,2332 0,2269	0,2173 0,2222 0,3281
Escol Nenhuma	-2,2437 0,6757 0,0008	-2,1518 0,7577 0,0045	-2,3872 0,6754 0,0004
Escol (1 a 3)	0,0373 0,4756 0,9373	0,0162 0,5436 0,9760	-0,0513 0,4782 0,9144
Escol (4 a 7)	-0,0319 0,2568 0,9009	-0,0033 0,2771 0,9904	0,0633 0,2650 0,8111
Escol (8 a 11)	0,2405 0,2139 0,2610	0,17182 0,2284 0,4519	0,1268 0,2171 0,5591
Grav dupla	-2,3176 0,3347 <0,0001	-2,5273 0,3694 <0,0001	-2,6983 0,3418 <0,0001
Grav tripla (+)	-3,2366 1,0127 0,0013	-3,3194 1,0112 0,0010	-3,3167 1,0008 0,0009
Gest <32	-4,0226 0,3846 <0,0001	-4,0263 0,4082 <0,0001	-3,9302 0,3790 <0,0001
Gest [32 a 36]	-2,4398 0,1593 <0,0001	-2,4651 0,1690 <0,0001	-2,4137 0,1605 <0,0001
Cons Nenhuma	-0,6199 0,7122 0,3841	-0,1564 1,3998 0,9110	-1,1376 0,6875 0,0979
Cons (1-3)	-0,9687 0,3043 0,0014	-0,7901 0,3298 0,0165	-0,7682 0,3136 0,0142
Cons (4-6)	-0,6094 0,1803 0,0007	-0,6751 0,1905 0,0003	-0,6585 0,1804 0,0002
Prim Mais Filhos	0,4930 0,1741 0,0046	0,5205 0,1853 0,0049	0,5531 0,1756 0,0016

Tabela 3 – Estimativas, erros padrão e valor-p do modelo logístico para dados completos, incompletos com 20% de dados faltantes e com imputação.

Variáveis preditoras	Estimativas, Erros Padrão e valores-p dos modelos logísticos ajustados		
	Modelo Padrão	Com 20% de dados faltantes	Com 20% de imputação
Idade (10-15)	-0,2848 0,4753 0,5489	0,03182 0,5516 0,9540	0,1043 0,5260 0,8428
Idade (16-19)	-0,4439 0,2161 0,0399	-0,3582 0,2452 0,1439	-0,4140 0,2156 0,0548
Idade (30-39)	-0,3478 0,1917 0,0696	-0,1886 0,2156 0,3817	-0,2512 0,1873 0,1799
Idade ≥ 40	-0,6173 0,4532 0,1732	-0,3792 0,5334 0,4771	-0,2922 0,4704 0,5344
Est.civ Solteira	0,2243 0,1921 0,2430	0,1764 0,2125 0,4062	0,0483 0,1860 0,7949
Est.civ Separada	-0,5452 0,5180 0,2925	-0,6823 0,5703 0,2315	-0,7090 0,4846 0,1434
Est.civ Uniao.Est	0,1621 0,2168 0,4546	0,3136 0,2468 0,2038	0,3585 0,2225 0,1071
Escol Nenhuma	-2,2437 0,6757 0,0008	-2,2795 0,7909 0,0039	-1,6655 0,7239 0,0214
Escol (1 a 3)	0,0373 0,4756 0,9373	-0,0961 0,5488 0,8609	-0,3677 0,4080 0,3673
Escol (4 a 7)	-0,0319 0,2568 0,9009	0,0689 0,2898 0,8118	0,3259 0,2537 0,1990
Escol (8 a 11)	0,24050 0,2139 0,2610	0,2226 0,2361 0,3457	0,4405 0,2060 0,0324
Grav dupla	-2,3176 0,3347 <0,0001	-2,4844 0,3870 <0,0001	-2,77156 0,3428 <0,0001
Grav tripla (+)	-3,2366 1,0127 0,0013	-3,2490 1,0090 0,0012	-3,1164 0,9956 0,0017
Gest <32	-4,0226 0,3846 <0,0001	-4,3316 0,4612 <0,0001	-4.08781 0,3879 <0,0001
Gest [32 a 36]	-2,4398 0,1593 <0,0001	-2,4299 0,1786 <0,0001	-2,4428 0,1577 <0,0001
Cons Nenhuma	-0,6199 0,7122 0,3841	0,0005 1,4978 0,9997	-1,4509 0,6655 0,0292
Cons (1-3)	-0,9687 0,3043 0,0014	-0,7303 0,3469 0,0352	-0,3845 0,3288 0,2422
Cons (4-6)	-0,6094 0,1803 0,0007	-0,6164 0,2035 0,0024	-0,6815 0,1765 0,0001
Prim Mais Filhos	0,4930 0,1741 0,0046	0,4688 0,1943 0,0158	0,3643 0,1716 0,0338

ignorem simplesmente o problema de dados faltantes. A imputação de dados pode aumentar, consideravelmente, o tamanho efetivo do conjunto de dados e assim dar mais confiabilidade para os resultados obtidos.

Capítulo 3

MÉTODO DE IMPUTAÇÃO MÚLTILPA LIVRE DE DISTRIBUIÇÃO (IMLD)

O método de Imputação Múltipla Livre de Distribuição (IMLD), proposto por Bergamo (2007), estima os valores a serem imputados por meio de uma mudança no procedimento de imputação simples desenvolvido por Krzanowski (1988). Este método parte, inicialmente, da afirmação feita por Good (1969), que qualquer matriz $\boldsymbol{Y}_{(n,p)}$ pode ser decomposta por valor singular na forma

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T. \tag{3.1}$$

em que $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{I}_p$ e $\boldsymbol{D} = diag(d_1, ..., d_p)$ com $d_1 \geq d_2 \geq, ..., \geq d_p \geq 0$. As matrizes $\boldsymbol{Y}^T\boldsymbol{Y}$ e $\boldsymbol{Y}\boldsymbol{Y}^T$ têm os mesmos autovalores não nulos, e os elementos d_i são a raiz quadrada destes autovalores. A i-ésima coluna $\boldsymbol{v}_i = (v_{i1}, ..., v_{ip})^T$ da matriz $\boldsymbol{V}_{p\times p}$ é o autovetor correspondente ao i-ésimo maior autovalor d_i^2 de $\boldsymbol{Y}^T\boldsymbol{Y}$; enquanto a j-ésima coluna $\boldsymbol{u}_j = (u_{1j}, ..., u_{nj})^T$ da matriz $\boldsymbol{U}_{n\times p}$ é o autovetor correspondente ao i-ésimo maior autovalor d_i^2 de $\boldsymbol{Y}\boldsymbol{Y}^T$.

A decomposição (3.1) tem sua representação elementar como

$$y_{ij} = \sum_{h=1}^{p} u_{ih} d_h v_{jh}. \tag{3.2}$$

Krzanowski (1988) usou esta representação como uma base para determinar a dimensionalidade de um conjunto de dados multivariados. Se a estrutura dos dados é essencialmente H-dimensional ($H < p$) então a variação na dimensão resultante ($p - H$) pode ser tratada como ruído aleatório. As características principais dos dados estarão supostamente no espaço dos H primeiros componentes principais. A correspondência entre as quantidades do lado direito de (3.2) e os eixos principais da configuração dos dados sugere o modelo de

H-componentes

$$y_{ij} = \sum_{h=1}^{H} u_{ih} d_h v_{jh} + \epsilon_{ij}, \tag{3.3}$$

em que ϵ_{ij} é o ruído.

Supondo o modelo (3.3) para um valor específico de H, com uma única observação y_{ij} ausente na matriz de dados, tem-se y_{ij} estimado por

$$\hat{y_{ij}}^{(H)} = \sum_{h=1}^{H} u_{ih} d_h v_{jh}, \tag{3.4}$$

em que u_{ih}, d_h, v_{jh}, devem ser estimados com o restante dos dados. As melhores estimativas destes valores estão baseadas na maior quantidade possível de dados. Simbolizada, por $\boldsymbol{Y}^{(-i)}$ a matriz dos dados obtida, retirando-se a i-ésima linha de $\boldsymbol{Y}$, e por $\boldsymbol{Y}^{(-j)}$ a matriz dos dados obtida, retirando-se a j-ésima coluna de $\boldsymbol{Y}$, a decomposição de valor singular dessas matrizes fica

$$\boldsymbol{Y}^{(-i)} = \bar{\boldsymbol{U}} \bar{\boldsymbol{D}} \bar{\boldsymbol{V}}^T, \quad \bar{\boldsymbol{U}} = (\bar{u}_{sh}), \quad \bar{\boldsymbol{V}} = (\bar{v}_{sh}), \quad \bar{\boldsymbol{D}} = (\bar{d}_1, ..., \bar{d}_p), \tag{3.5}$$

e

$$\boldsymbol{Y}_{(-j)} = \tilde{\boldsymbol{U}} \tilde{\boldsymbol{D}} \tilde{\boldsymbol{V}}^T, \quad \tilde{\boldsymbol{U}} = (\tilde{u}_{sh}), \quad \tilde{\boldsymbol{V}} = (\tilde{v}_{sh}), \quad \tilde{\boldsymbol{D}} = (\tilde{d}_1, ..., \tilde{d}_{p-1}). \tag{3.6}$$

A estimativa de u_{ih} e v_{jh} em (3.6), obtida com o máximo dos dados de $\boldsymbol{Y}$, é $\tilde{u}_{ih}$ e $\bar{v}_{jh}$, respectivamente, enquanto d_h pode ser estimado por $\bar{d}_h$, $\tilde{d}_h$ ou por alguma combinação dos dois. Uma forma adequada parece ser $\sqrt{\bar{d}_h}\sqrt{\tilde{d}_h}$ em que uma estimativa do valor ausente y_{ij} é dada por

$$\hat{y_{ij}}^{(H)} = \sum_{h=1}^{H} (\tilde{u}_{ih}\sqrt{\tilde{d}_h})(\bar{v}_{jh}\sqrt{\bar{d}_h}). \tag{3.7}$$

Seguindo o preceito da máxima informação dos dados, usa-se o valor mais elevado disponível de H. De (3.7), este valor é, evidentemente, $p-1$, então o valor imputado a y_{ij} será

$$\hat{y_{ij}} = \sum_{h=1}^{p-1} (\tilde{u}_{ih}\sqrt{\tilde{d}_h})(\bar{v}_{jh}\sqrt{\bar{d}_h}). \tag{3.8}$$

As estimativas iniciais dos valores y_{ij} ausentes são feitas pela média $\bar{y_j}$ da j-ésima coluna. Para evitar qualquer influência de possíveis variações entre as colunas, por exemplo, a escala das variáveis, e para tornar o algoritmo mais estável, é recomendado aplicar uma padronização em $\boldsymbol{Y}$. Para os valores y_{ij}, inclusive os ausentes já substituídos pela média ($\bar{y_j}$), é calculada uma nova média ($\bar{y_j}'$) e um desvio padrão (dp_j) para cada coluna j, então y_{ij} é padronizado por $y'_{ij} = \frac{y_{ij} - \bar{y_j}'}{dp_j}$. Padronização semelhante também é feita nas matrizes $\boldsymbol{Y}^{(-i)}$ e $\boldsymbol{Y}_{(-j)}$.

As estimativas de cada valor ausente são recalculadas usando (3.8) matrizes padronizadas. Para cada estimativa são necessárias duas decomposições de valores singulares, isto é, uma para i e outra para j. Finalmente, na matriz $\boldsymbol{Y}$ completada (*observados* + *imputados*) é aplicada uma operação para retorno á sua escala original, ou seja, se $y_{ij}^{(c)}$ representa cada

valor da matriz $\boldsymbol{Y}$ completada, calcula-se novamente a média da coluna j ($\bar{y}_j^{(c)}$) e o seu desvio padrão ($s_j^{(c)}$). Cada valor da matriz $\boldsymbol{Y}$ completada, na escala original, é então obtido por, $y_{ij} = \bar{y}_j^{(c)} + s_j^{(c)} y_{ij}^{(c)}$.

Para gerar as imputações ($m = 1, ..., M$) na primeira etapa da imputação múltipla, Bergamo (2007) propôs uma mudança nos expoentes dos radicandos $\tilde{d}_h$ e $\bar{d}_h$ em (3.8), ou seja, de uma maneira genérica, se $\sqrt[b]{d^a}$ for representada como uma potência fracionária $d^{\frac{a}{b}}$, o procedimento requer a mudança no numerador do expoente, tanto de $\tilde{d}_h^{\frac{\tilde{a}}{b}}$ como de $\bar{d}_h^{\frac{\bar{a}}{b}}$, de modo que a soma dos expoentes seja igual a 1 ($\frac{\tilde{a}+\bar{a}}{b} = 1$). Krzanowski (1988) sugere como estimativas para d_h em (3.4) uma combinação entre $\bar{d}_h$ e $\tilde{d}_h$ de (3.5) e (3.6), respectivamente, resultando na forma $\sqrt{\bar{d} - h}\sqrt{\tilde{d} - h}$, a qual admite influências iguais de (3.6) e (3.7). Assim, variando os expoentes de $\tilde{d}_h$ e $\bar{d}_h$, admite-se um peso maior para (3.6) ou (3.7) na estimativa final de y_{ij} em (3.8).

Cada mudança em $\tilde{a}$ e, consequentemente em $\bar{a}$, gera uma nova matriz $\boldsymbol{Y}$ completada, caracterizando, assim, um processo de geração dos M conjuntos de dados completados da primeira etapa da imputação múltipla.

O número de imputações fica condicionado às mudanças nos expoentes. Assim, com um número de 5 mudanças nos expoentes há uma variação entre 40% e 60% nos pesos dados a (3.6) e (3.7), ou seja, partindo de um denominador fixo ($b = 20$, por exemplo), os valores assumidos por $\tilde{a}$(8, 9, 10, 11 e 12) e respectivamente por $\bar{a}$(12, 11, 10, 9 e 8) levam a uma variação (40%, 45%, 50%, 55% e 60%) nas proporções de (3.6) e (3.7) em

$$\hat{y}_{ij} = \sum_{h=1}^{p-1} (\tilde{u}_{ih} \tilde{d}_h^{\frac{\tilde{a}}{b}})(\bar{v}_{jh} \bar{d}_h^{\frac{\bar{a}}{b}}). \tag{3.9}$$

Para avaliar a exatidão do método de imputação, medidas de acurácia, podem ser calculadas com as expressões seguintes,

$$acc_l = \frac{\sum_{m=1}^{M} (\hat{y}_{ij_m} - VO_l)^2}{M - 1}, \tag{3.10}$$

em que M é o número de imputações, VO é o valor original retirado aleatoriamente na posição $l = 1, 2, ..., na$, em que na representa o número total de valores retirados correspondentes à i-ésima linha e j-ésima coluna $(i; j)$, e $\hat{y}_{ij}$ seu respectivo valor imputado (3.9). Esta expressão pode ser separada em duas partes,

$$acc_l = \frac{\sum_{m=1}^{M} (\hat{y}_{ij_m} - \bar{Y}_l)^2}{M - 1} + \frac{M(\bar{Y}_l - VO_l)^2}{M - 1}, \tag{3.11}$$

em que $\bar{Y}_l$ é a média das imputações na posição l. A primeira parte representa uma variância e a segunda um viés, dos M valores imputados em cada posição.

Uma medida geral da acurácia T_{acc} pode ser calculada por meio da média das acc_l, ou seja,

$$T_{acc} = \frac{\sum_{l=1}^{na} acc_l}{na}, \tag{3.12}$$

em que, $na = g \times e \times porc$, g é o total de genótipos, e o total de ambientes e $porc$ a porcentagem de ausência dos dados.

T_{acc} possui dois componentes,

$$T_{acc} = V_E + VQM, \tag{3.13}$$

em que

$$V_E = \frac{1}{na}\sum_{l=1}^{na}\left[\frac{\sum_{m=1}^{M}(\hat{y}_{ij_m} - \bar{Y}_l)^2}{M-1}\right] \tag{3.14}$$

e

$$VQM = \frac{1}{na}\sum_{l=1}^{na}\frac{M(\bar{Y}_l - VO_l)^2}{M-1} \tag{3.15}$$

O primeiro representa a variância entre imputações (V_E), portanto, quanto maior o seu valor, menor é a precisão do método de imputação múltipla, mas um valor reduzido desta variância também não significa um bom método de imputação, pois o método pode ser tendencioso. O segundo componente representa o viés quadrático médio (VQM) entre os valores de $\bar{Y}$ e VO, quanto menor for o viés, mais as imputações assemelham-se aos valores originais. Assim, melhor será o método de imputação múltipla quanto menores forem V_E e VQM (BERGAMO, 2007),(BERGAMO; DIAS; KRZANOWSKI, 2008).

A metodologia IMLD, foi proposta para aplicação em dados de ensaios multiambientais (BERGAMO, 2007), usados em estudos de melhoramento genético de plantas para testar a adaptação das cultivares em diferentes ambientes, conhecida como interação genótipo por ambiente ($G \times E$) (DIAS; KRZANOWSKI, 2003).

Vários trabalhos têm utilizado e testado a IMLD em matrizes de interação $G \times E$ com informação incompleta. Arciniegas-Alarcón e Dias (2009) avaliam, usando os modelos de efeitos aditivos com interação multiplicativa (AMMI), o método IMLD comparando-o com outros procedimentos de imputação ((CALINSKI et al., 1992) e (DENIS; BARIL, 1992)) que já tiveram êxito em experimentos genótipo-ambiente com dados faltantes.

Em outro artigo, Arciniegas-Alarcón e Dias (2010) avaliam a adequação de definir o número de componentes multiplicativos do modelo AMMI em experimentos de interação genótipo $\times$ ambiente de algodão com matrizes de dados que contêm imputações por IMLD ou apenas a informação observada (matriz incompleta). Os autores recomendam o uso da imputação de dados para a estimação dos parâmetros de um modelo AMMI sob ocorrência de dados ausentes, mas que para determinar o número de componentes multiplicativos é preferível basear-se apenas nas informações observadas.

Em Silva (2012) é comparada a eficiência da imputação múltipla livre da distribuição (IMLD) com o método de imputação múltipla com Monte Carlo via cadeia de Markov (IMMCMC), na imputação de unidades ausentes presentes em experimentos na cultura de *Eucaluptus grandis* de interação genótipo $\times$ ambiente. De acordo com os resultados obtidos

pela autora, a eficiência relativa em ambas as porcentagens manteve-se acima de 90% quando imputado com a IMLD.

No trabalho de Arciniegas-Alarcón et al. (2014) são comparados quatro algoritmos (imputação *biplot* , EM+DVS, imputação GabrielEigen e imputação múltipla livre de distribuição – IMLD), recentemente relatados na literatura, para imputação baseada na decomposição por valores singulares de uma matriz (DVS). A metodologia usada como padrão-ouro foi a EM-AMMI (JR; ZOBEL, 1990). Os termos gerais, os autores concluem que com o métodos EM+DVS os resultados obtidos são competitivos aos do padrão-ouro. Os métodos IMLD e GabrielEigen tiveram desempenho intermediário, sendo a imputação *biplot* a menos eficiente.

A exemplo dos métodos de imputação presentes na pesquisa de Arciniegas-Alarcón et al. (2014), recém mencionada, existem na literatura várias propostas metodológicas de imputação com base na decomposição por valores singulares de uma matriz (DVS). Arciniegas-Alarcón, Dias e García-Peña (2014) propõem um novo algoritmo de imputação múltipla livre de distribuição em tabelas incompletas de dupla entrada, por meio de modificações no método de imputação simples recentemente desenvolvido por Yan (2013) para contornar o problema de desbalanceamento de experimentos. O método utiliza DVS e não depende de pressuposições distribucionais ou estruturais dos dados, bem como, não tem restrições quanto ao padrão ou mecanismo dos dados faltantes.

3.1 APLICAÇÃO

A aplicação é feita com dados fornecidos no Boletim Técnico Nº 86 (SHIOGA et al., 2015), do Instituto Agronômico do Paraná (IAPAR-Londrina), que avalia o comportamento agronômico de cultivares de milho convencionais e geneticamente modificadas, na segunda safra de milho 2015, no Estado do Paraná.

A fim de garantir diferentes condições edafoclimáticas das principais regiões produtoras de milho da segunda safra no Estado, os ensaios avaliados pelo IAPAR foram implantados em oito municípios, seguindo a época de plantio e o sistema de cultivo da região. A avaliação das cultivares foi dividida em três grupos: cultivares superprecoces geneticamente modificadas, cultivares precoces geneticamente modificadas e, cultivares convencionais.

O delineamento experimental foi o aleatorizado em blocos com três repetições. As parcelas foram constituídas por duas fileiras de cinco metros de comprimento, espaçadas $0,80m$ entre linhas e com cinco plantas por metro linear após o desbaste. Na coleta dos dados as duas fileiras foram integralmente consideradas como área útil ($8,00\ m^2$).

A matriz $\boldsymbol{Y}$ de dados se refere a valores de altura média das plantas de 20 cultivares precoces de milho e geneticamente modificadas, algumas ainda em fase experimental, avaliadas apenas em 7 localidades, pois para uma localidade experimental os dados de altura não estão

disponíveis. As localidades, as cultivares e suas respectivas alturas médias estão na Tabela 4.

Tabela 4 – Altura média (m) das cultivares de milhos precoces geneticamente modificadas nas 7 diferentes localidades

Cultivar \ Localidade	Londrina	Sertanópolis	Cambará	Floresta	Campo Mourão	Palotina	Santa Helena
AS 1633 PRO2	2,43	2,85	2,35	2,55	2,55	2,17	2,30
P 30S31 YH	2,65	2,83	2,35	2,75	2,72	2,22	2,43
MG 699 PW	2,48	2,63	2,20	2,62	2,58	1,93	2,48
RB 9006 PRO	2,50	2,73	2,32	2,63	2,55	1,98	2,37
2B610 PW	2,37	2,72	2,35	2,43	2,55	2,20	2,28
MG 580 PW	2,30	2,53	2,32	2,40	2,43	2,00	2,27
2B633 PW	2,45	2,67	2,15	2,47	2,45	2,17	2,40
CD 3715 PRO2	2,35	2,67	2,32	2,45	2,38	2,10	2,32
MG 652 PW	2,50	2,53	2,42	2,55	2,55	2,20	2,32
BG 7432 H	2,45	2,68	2,35	2,55	2,60	2,13	2,38
RB 9005 PRO	2,55	2,80	2,37	2,58	2,53	2,23	2,45
P 4285 YH	2,53	2,65	2,40	2,55	2,58	2,20	2,47
RB 9004 PRO	2,45	2,60	2,25	2,43	2,52	2,05	2,40
CD 3770 PW	2,28	2,52	2,18	2,42	2,52	2,05	2,27
CD 3765 PW	2,18	2,55	2,25	2,42	2,37	2,02	2,13
EXP 90159 VIP3	2,40	2,68	2,23	2,50	2,47	2,00	2,27
SHS 7990 PRO2	2,43	2,68	2,25	2,43	2,70	2,08	2,23
CD 384 PW	2,43	2,77	2,33	2,25	2,48	2,05	2,43
P 30F53 YH	2,47	2,78	2,32	2,57	2,50	1,93	2,37
22M12 VIP	2,28	2,38	1,97	2,13	2,28	1,90	2,00
Média	2,42	2,66	2,28	2,48	2,51	2,08	2,33

3.1.1 Metodologia

Os procedimentos de imputação foram feitos em três cenários distintos, com 5%, 15% e 30% de perdas aleatórias geradas, de mesmo valor inicial como semente. O método de Imputação Múltipla Livre de Distribuição, primeira passo da IM, foi implementado no *software R*. Com a execução do programa obtém-se um arquivo com os $M = 5$ conjuntos de dados completos gerados pela imputação, esses dados já estão prontos para serem utilizados na segunda etapa da Imputação Múltipla. Os códigos com a geração das perdas aleatórias e com o método IMLD é apresentado no Anexo A, para que possa ser utilizado, avaliado e melhorado.

Para a segunda fase do processo de imputação, isto é, a análise individual dos 5 conjuntos de dados completados, utilizou-se a *Proc UNIVARIATE* do aplicativo *SAS* , Anexo B, sendo o *SAS* disponível no departamento de Estatística/UEM. Como resultado desse programa

tem-se um conjunto de dados com a média de alturas de cada localidade e seu respectivo erro padrão, medidas usadas na terceira fase da Imputação Múltipla.

Na terceira etapa, realizada na *Proc MIANALYZE* do *SAS* e apresentada no Anexo C, as estimativas da média de alturas ($\hat{\beta_m}$) de cada localidade, dos 5 bancos gerados, são combinadas em uma só média ($\hat{\beta}^*$). Nesta etapa é efetuado também o cálculo da estatística t-Student, testando a hipóteses de que a média ($\hat{\beta}^*$) é igual a média original de alturas em cada ambiente (última linha da Tabela 4). Por fim, para cada porcentagem de valores ausentes, é calculada uma medida de acurácia, definida pela expressão 3.13.

3.1.2 Resultados e discussões

Os primeiros resultados apresentados são para a retirada aleatória de 5% dos dados da matriz original $Y_{c\times l}$, isto é, 7 valores foram retirados da Tabela 4.

Uma comparação entre os valores originais, os valores gerados em cada imputação, média e variância das 5 imputações é apresentada na Tabela 5.

Tabela 5 – Valores originais (VO), média ($\bar{M}$) e variância (Var) das imputações, para as alturas, na retirada aleatória de 5% dos valores da Tabela 4.

Posição retirada	VO	Imputação (M) 1	2	3	4	5	$\bar{M}$	Var
(1;1)	2,43	2,4357	2,4356	2,4356	2,4355	2,4354	2,4356	$1,07 \times 10^{-8}$
(11;1)	2,55	2,4674	2,4672	2,467	2,4668	2,4666	2,4670	$9,48 \times 10^{-8}$
(7;2)	2,67	2,6570	2,6570	2,6570	2,6570	2,6570	2,6570	$3,62 \times 10^{-10}$
(16;2)	2,68	2.6520	2,6521	2,6521	2,6521	2,6521	2,6521	$1,92 \times 10^{-9}$
(18;3)	2,33	2,3281	2,3280	2,3280	2,3280	2,3280	2,3280	$2,25 \times 10^{-9}$
(5;6)	2,2	2,1443	2,1442	2,1441	2,1440	2,1439	2,1441	$2,04 \times 10^{-8}$
(14;7)	2,27	2,2839	2,2839	2,2840	2,2841	2,2841	2,2840	$1,26 \times 10^{-8}$

Pela posição da coluna dos valores originais referidos na Tabela 5, nota-se que as localidades 4 e 5 não possuem valores faltantes e que as localidades 1 e 2 apresentam dois dados ausentes cada, enquanto as localidades 3,6 e 7 possuem apenas 1 *missing data*.

Nas posições (1;1) e (18;3) as imputações geraram valores maiores do que a medida original da altura das plantas, enquanto que, nas outras posições as imputações foram menores. As variâncias próximas de zero, em todas as posições de retirada, expressam a pequena variabilidade dos valores imputados, com relação à média das imputações ($\bar{M}$).

Na segunda etapa da IM, *Proc UNIVARIATE do SAS* (Anexo B), são calculadas as médias das localidades e os seus respectivos erros padrão, para cada conjunto completado

pelas imputações, Tabela 6. Por não possuírem valores omissos, as localidade 4 e 5 não são apresentadas.

Tabela 6 – Médias e erros padrão de alturas das localidades completadas pelas imputações, com 5% de dados faltantes.

	Localidade (média/ erro padrão)				
Imputações	1	2	3	6	7
1	2,420152	2,660452	2,283904	2,077713	2,329193
	0,023412	0,026610	0,023210	0,022972	0,026349
2	2,420139	2,660453	2,283902	2,077708	2,329196
	0,023411	0,026610	0,023210	0,022971	0,026348
3	2,420126	2,660453	2,283901	2,077704	2,329200
	0,023410	0,026610	0,023210	0,022970	0,026348
4	2,420113	2,660454	2,283899	2,077699	2,329203
	0,023409	0,026610	0,023209	0,022970	0,026348
5	2,420100	2,660455	2,283898	2,077695	2,329207
	0,023408	0,026610	0,023209	0,022969	0,026347

Por meio da *Proc MIANALYZE* do *SAS* (Anexo C), combinaram-se as estimativas das médias de alturas das 5 localidades (1, 2, 3, 6 e 7), provenientes dos 5 conjuntos de dados completados pelas imputações. Os resultados desta etapa estão resumidamente apresentados na Tabela 7.

Pelo teste t-Student, com 5% de significância, não houve diferença significativa das médias estimadas ($\hat{\beta}^*$) para as médias das alturas nos dados originais.

A variabilidade (T) das estimativas ($\hat{\beta}^*$), o aumento relativo na variância, devido à ausência de dados (r), bem como a precisão da estimativa se nenhum dado estivesse ausente (λ), foram valores próximos de zero, indicando pouca influência da falta de dados nas estimativas. A eficiência relativa (ER), o número de imputações utilizadas foi 100% de uma estimativa feita com infinitas imputações.

O banco de dados completado, com 5% de *missing*, considerando a média das alturas dos 5 bancos gerados pela imputação, são expostos na Tabela 8.

Tabela 7 – Estimativa média $\hat{\beta}^*$ das médias de alturas, medidas associadas a sua variabilidade (B,$\bar{W}$ e T) e Teste t-Student, com v_{obs} graus de liberdade, para comparação da média original nas localidades, com 5% de dados faltantes.

	Localidades				
Estimativas	1	2	3	6	7
$\hat{\beta}^*$	$2,420126$	$2,660453$	$2,283901$	$2,077704$	$2,329200$
B	$4,2 \times 10^{-10}$	$1,5 \times 10^{-12}$	$5,6 \times 10^{-12}$	$5,1 \times 10^{-11}$	$3,1 \times 10^{-11}$
$\bar{W}$	$0,000548$	$0,000708$	$0,000539$	$0,000528$	$0,000694$
T	$0,000548$	$0,000708$	$0,000539$	$0,000528$	$0,000694$
r	$9,2 \times 10^{-7}$	$2,6 \times 10^{-9}$	$1,2 \times 10^{-8}$	$1,1 \times 10^{-7}$	$5,4 \times 10^{-8}$
λ	$9,2 \times 10^{-7}$	$2,6 \times 10^{-9}$	$1,2 \times 10^{-8}$	$1,1 \times 10^{-7}$	$5,4 \times 10^{-8}$
ER	$1,00$	$1,00$	$1,00$	$1,00$	$1,00$
v_{obs}	$17,273$	$17,273$	$17,273$	$17,273$	$17,273$
$\bar{Y}$	$2,4240$	$2,6625$	$2,2840$	$2,0805$	$2,3285$
t_{calc}	$-0,17$	$-0,08$	$0,00$	$-0,12$	$0,03$
$valor-p$	$0,8705$	$0,9396$	$0,9966$	$0,9045$	$0,9791$

Quando retirados aleatoriamente 15% dos dados da Matriz original, 21 valores, todas as localidades apresentaram *missing data*, com as localidades 1 e 7 concentrando 5 dados faltantes cada, Tabela 9.

Tabela 8 – Matriz de dados completada pela imputação com 5% de valores ausentes.

	Localidade						
Cultivar	1	2	3	5	5	6	7
1	2,44	2,85	2,35	2,55	2,55	2,17	2,30
2	2,65	2,83	2,35	2,75	2,72	2,22	2,43
3	2,48	2,63	2,20	2,62	2,58	1,93	2,48
4	2,50	2,73	2,32	2,63	2,55	1,98	2,37
5	2,37	2,72	2,35	2,43	2,55	2,14	2,28
6	2,30	2,53	2,32	2,4	2,43	2,00	2,27
7	2,45	2,66	2,15	2,47	2,45	2,17	2,40
8	2,35	2,67	2,32	2,45	2,38	2,10	2,32
9	2,50	2,53	2,42	2,55	2,55	2,20	2,32
10	2,45	2,68	2,35	2,55	2,60	2,13	2,38
11	2,47	2,8	2,37	2,58	2,53	2,23	2,45
12	2,53	2,65	2,40	2,55	2,58	2,20	2,47
13	2,45	2,60	2,25	2,43	2,52	2,05	2,40
14	2,28	2,52	2,18	2,42	2,52	2,05	2,28
15	2,18	2,55	2,25	2,42	2,37	2,02	2,13
16	2,40	2,65	2,23	2,5	2,47	2,00	2,27
17	2,43	2,68	2,25	2,43	2,70	2,08	2,23
18	2,43	2,77	2,33	2,25	2,48	2,05	2,43
19	2,47	2,78	2,32	2,57	2,5	1,93	2,37
20	2,28	2,38	1,97	2,13	2,28	1,90	2,00
Média	2,42	2,66	2,28	2,48	2,52	2,08	2,33

Os valores gerados pelas 5 imputações continuaram com pouca oscilação em relação aos valores originais, ou seja, as variâncias continuaram próximas de zero (Tabela 9). As estimativas para os dados ausentes tiveram pequenas alterações, se comparadas com a imputação feita com 5% de $missing$, como por exemplo, a posição (1;1), com uma estimativa de 2,4356 para 5% caindo para 2,4186 com 15%. Já para a posição (14;7), ouve um aumento na estimativa, de 2,328 com 5% para 2,3453 com 15%.

Tabela 9 – Valores originais (VO), média ($\bar{M}$) e variância (Var) das imputações, para as alturas, na retirada aleatória de 15% dos valores da Tabela 4.

Posição retirada	VO	Imputação (M)					$\bar{M}$	Var
		1	2	3	4	5		
(1;1)	2,43	2,4188	2,4187	2,4186	2,4185	2,4184	2,4186	$3,03 \times 10^{-8}$
(3;1)	2,48	2,4621	2,4621	2,4621	2,4622	2,4622	2,4621	$5,89 \times 10^{-10}$
(4;1)	2,5	2,446	2,446	2,4459	2,4459	2,4459	2,4459	$4,33 \times 10^{-9}$
(11;1)	2,55	2,44	2,4398	2,4396	2,4394	2,4391	2,4396	$1,18 \times 10^{-7}$
(14;1)	2,28	2,3992	2,3994	2,3995	2,3997	2,3998	2,3995	$5,28 \times 10^{-8}$
(7;2)	2,67	2,663	2,663	2,663	2,663	2,663	2,663	$3,11 \times 10^{-10}$
(8;2)	2,67	2,6574	2,6574	2,6574	2,6575	2,6575	2,6574	$1,36 \times 10^{-9}$
(16;2)	2,68	2,6622	2,6623	2,6624	2,6624	2,6625	2,6624	$1,38 \times 10^{-8}$
(19;2)	2,78	2,6878	2,6878	2,6878	2,6878	2,6877	2,6878	$1,86 \times 10^{-9}$
(20;2)	2,38	2,5622	2,5626	2,5631	2,5636	2,5641	2,5631	$5,61 \times 10^{-7}$
(18;3)	2,33	2,297	2,297	2,2971	2,2971	2,2972	2,2971	$5,19 \times 10^{-9}$
(12;4)	2,55	2,5131	2,5129	2,5127	2,5125	2,5123	2,5127	$9,25 \times 10^{-8}$
(16;4)	2,5	2,4662	2,4663	2,4664	2,4665	2,4666	2,4664	$2,55 \times 10^{-8}$
(20;5)	2,28	2,4698	2,4701	2,4704	2,4706	2,4709	2,4704	$1,99 \times 10^{-7}$
(3;6)	1,93	2,0363	2,036	2,0358	2,0355	2,0352	2,0358	$1,73 \times 10^{-7}$
(5;6)	2,2	2,1247	2,1248	2,1249	2,125	2,125	2,1249	$1,77 \times 10^{-8}$
(10;7)	2,38	2,3822	2,3822	2,3821	2,382	2,382	2,3821	$8,37 \times 10^{-9}$
(14;7)	2,27	2,345	2,3451	2,3453	2,3454	2,3455	2,3453	$4,01 \times 10^{-8}$
(15;7)	2,13	2,3409	2,3409	2,341	2,3411	2,3411	2,341	$1,07 \times 10^{-8}$
(16;7)	2,27	2,3631	2,3631	2,3631	2,3632	2,3632	2,3631	$3,42 \times 10^{-9}$
(20;7)	2,00	2,2922	2,2926	2,2929	2,2933	2,2936	2,2929	$3,24 \times 10^{-7}$

Os resultados do *Proc UNIVARIATE do SAS* (Anexo B), para cada uma das localidades completadas pelas imputações, estão apresentados na Tabela 10.

As estimativas das localidades combinadas por meio da *Proc MIANALYZE* do *SAS* (Anexo C), estão apresentados na Tabela 11.

Tabela 10 – Médias e erros padrão de alturas das localidades completadas pelas imputações, com 15% de dados faltantes.

	Localidade (média/erro padrão)						
Imputações	1	2	3	4	5	6	7
1	2,420310	2,665129	2,282349	2,480462	2,524989	2,082050	2,362167
	0,021733	0,022121	0,023106	0,030200	0,019955	0,021632	0,015526
2	2,420300	2,665157	2,282351	2,480458	2,525003	2,082041	2,362194
	0,021732	0,022115	0,023106	0,030200	0,019953	0,021634	0,015521
3	2,420289	2,665185	2,282353	2,480453	2,525018	2,082032	2,362220
	0,021731	0,022109	0,023106	0,030199	0,019951	0,021636	0,015516
4	2,420279	2,665213	2,282355	2,480449	2,525032	2,082023	2,362247
	0,021730	0,022103	0,023106	0,030198	0,019949	0,021638	0,015511
5	2,420268	2,665240	2,282358	2,480444	2,525046	2,082014	2,362273
	0,021729	0,022097	0,023106	0,030198	0,019947	0,021640	0,015506

Tabela 11 – Estimativa média $\hat{\beta}^*$ das médias de alturas, medidas associadas a sua variabilidade (B,$\bar{W}$ e T) e Teste t-Student, com v_{obs} graus de liberdade, para comparação da média original nas localidades, com 15% de dados faltantes

	Localidades						
Estimativas	1	2	3	4	5	6	7
$\hat{\beta}^*$	$2,420289$	$2,665185$	$2,282353$	$2,480453$	$2,525018$	$2,082032$	$2,362220$
B	$2,7 \times 10^{-10}$	$1,9 \times 10^{-9}$	$1,2 \times 10^{-11}$	$5,2 \times 10^{-11}$	$4,9 \times 10^{-10}$	$2,0 \times 10^{-10}$	$1,7 \times 10^{-9}$
$\bar{W}$	$0,000472$	$0,000489$	$0,000534$	$0,000912$	$0,000398$	$0,000468$	$0,000241$
T	$0,000472$	$0,000489$	$0,000534$	$0,000912$	$0,000398$	$0,000468$	$0,000241$
r	$6,8 \times 10^{-7}$	$4,7 \times 10^{-6}$	$2,9 \times 10^{-8}$	$6,8 \times 10^{-8}$	$1,4 \times 10^{-6}$	$5,1 \times 10^{-7}$	$8,7 \times 10^{-7}$
λ	$6,8 \times 10^{-7}$	$4,7 \times 10^{-7}$	$2,9 \times 10^{-8}$	$6,8 \times 10^{-8}$	$1,4 \times 10^{-6}$	$5,1 \times 10^{-7}$	$8,7 \times 10^{-6}$
ER	$1,00$	$0,999999$	$1,00$	$1,00$	$1,00$	$1,00$	$0,999999$
v_{obs}	$17,273$	$17,273$	$17,273$	$17,273$	$17,273$	$17,273$	$17,273$
$\bar{Y}$	$2,4240$	$2,6625$	$2,2840$	$2,4840$	$2,5155$	$2,0805$	$2,3285$
t_{calc}	$-0,17$	$0,12$	$-0,07$	$-0,12$	$0,48$	$0,07$	$2,17$
$valor-p$	$0,8664$	$0,9047$	$0,9440$	$0,9079$	$0,6393$	$0,9444$	$0,0439$

Pelo teste t-Student, com 5% de significância, na localidade 7, o valor-p foi de 0,0439, isto é, $p < 0,05$, indicando uma diferença da média estimada com a média original. Porém, o intervalo de confiança engloba, perto do seu limite inferior, o verdadeiro valor da média. Para as outras localidades não houve diferença significativa entre as médias estimadas ($\hat{\beta}^*$) com os dados originais.

A variabilidade (T) das estimativas ($\hat{\beta}^*$), o aumento relativo na variância, devido à ausência de dados (r) e a precisão da estimativa se nenhum dado estivesse ausente (λ), continuaram valores próximos de zero, indicando boas estimativas. As eficiências relativas (ER), isto é, o número de imputações utilizadas foi 100% e $0,999999$ de uma estimativa feita com infinitas imputações.

O banco de dados completado, com 15% de *missing*, considerando a média das alturas dos 5 bancos gerados pela imputação, são expostos na Tabela 12.

Tabela 12 – Matriz de dados completada pela imputação com 15% de valores ausentes.

	Localidade						
Cultivar	1	2	3	5	5	6	7
1	2,42	2,85	2,35	2,55	2,55	2,17	2,3
2	2,65	2,83	2,35	2,75	2,72	2,22	2,43
3	2,46	2,63	2,2	2,62	2,58	2,04	2,48
4	2,45	2,73	2,32	2,63	2,55	1,98	2,37
5	2,37	2,72	2,35	2,43	2,55	2,12	2,28
6	2,3	2,53	2,32	2,4	2,43	2	2,27
7	2,45	2,66	2,15	2,47	2,45	2,17	2,4
8	2,35	2,66	2,32	2,45	2,38	2,1	2,32
9	2,5	2,53	2,42	2,55	2,55	2,2	2,32
10	2,45	2,68	2,35	2,55	2,6	2,13	2,38
11	2,44	2,8	2,37	2,58	2,53	2,23	2,45
12	2,53	2,65	2,4	2,51	2,58	2,2	2,47
13	2,45	2,6	2,25	2,43	2,52	2,05	2,4
14	2,4	2,52	2,18	2,42	2,52	2,05	2,35
15	2,18	2,55	2,25	2,42	2,37	2,02	2,34
16	2,4	2,66	2,23	2,47	2,47	2	2,36
17	2,43	2,68	2,25	2,43	2,7	2,08	2,23
18	2,43	2,77	2,3	2,25	2,48	2,05	2,43
19	2,47	2,69	2,32	2,57	2,5	1,93	2,37
20	2,28	2,56	1,97	2,13	2,47	1,9	2,29
Média	2,42	2,66	2,28	2,48	2,52	2,08	2,36

Com a retirada aleatória de 30% do conjunto de dados originais, foram excluídos 42 valores, distribuídos pelas 7 localidades, Tabela 13 . As maiores quantidades de ausências continuaram nas localidade 1 e 7.

Nas mesmas posições de retirada dos dados com 5% de ausência, os valores imputados continuaram variando para mais ou para menos em relação aos originais, com medidas de variância ainda próximas de zero.

Tabela 13 – Valores originais (VO), média ($\bar{M}$) e variância (Var) das imputações, para as alturas, na retirada aleatória de 30% dos valores da Tabela 4. Continua.

Posição retirada	VO	Imputação (m)					$\bar{m}$	Var
		1	2	3	4	5		
$(1;1)$	$2,43$	$2,4286$	$2,4285$	$2,4283$	$2,4282$	$2,4281$	$2,4283$	$3,76 \times 10^{-8}$
$(3;1)$	$2,48$	$2,4714$	$2,4714$	$2,4714$	$2,4714$	$2,4714$	$2,4714$	$1,06 \times 10^{-10}$
$(4;1)$	$2,50$	$2,4526$	$2,4526$	$2,4525$	$2,4524$	$2,4524$	$2,4525$	$7,32 \times 10^{-9}$
$(8;1)$	$2,35$	$2,3764$	$2,3764$	$2,3765$	$2,3765$	$2,3766$	$2,3765$	$4,61 \times 10^{-9}$
$(9;1)$	$2,50$	$2,4234$	$2,4233$	$2,4232$	$2,4231$	$2,423$	$2,4232$	$3,15 \times 10^{-8}$
$(11;1)$	$2,55$	$2,4663$	$2,4661$	$2,4659$	$2,4657$	$2,4654$	$2,4659$	$1,08 \times 10^{-7}$
$(14;1)$	$2,28$	$2,3768$	$2,3769$	$2,3771$	$2,3772$	$2,3773$	$2,3771$	$5,09 \times 10^{-8}$
$(1;2)$	$2,85$	$2,6693$	$2,6692$	$2,6691$	$2,669$	$2,6689$	$2,6691$	$2,61 \times 10^{-8}$
$(4;2)$	$2,73$	$2,6669$	$2,6669$	$2,6668$	$2,6668$	$2,6668$	$2,6668$	$4,81 \times 10^{-9}$
$(7;2)$	$2,67$	$2,6231$	$2,6231$	$2,6231$	$2,6231$	$2,6231$	$2,6231$	$7,92 \times 10^{-10}$
$(8;2)$	$2,67$	$2,6329$	$2,633$	$2,633$	$2,633$	$2,633$	$2,633$	$3,04 \times 10^{-11}$
$(11;2)$	$2,80$	$2,6805$	$2,6803$	$2,6801$	$2,68$	$2,6798$	$2,6801$	$6,44 \times 10^{-8}$
$(12;2)$	$2,65$	$2,6677$	$2,6676$	$2,6675$	$2,6674$	$2,6673$	$2,6675$	$3,51 \times 10^{-8}$
$(16;2)$	$2,68$	$2,6269$	$2,6269$	$2,627$	$2,627$	$2,6271$	$2,627$	$5,77 \times 10^{-9}$
$(17;2)$	$2,68$	$2,6263$	$2,6263$	$2,6263$	$2,6264$	$2,6264$	$2,6263$	$2,75 \times 10^{-9}$
$(19;2)$	$2,78$	$2,667$	$2,667$	$2,667$	$2,6669$	$2,6669$	$2,667$	$1,66 \times 10^{-9}$
$(20;2)$	$2,38$	$2,4966$	$2,497$	$2,4974$	$2,4978$	$2,4982$	$2,4974$	$4,04 \times 10^{-7}$
$(10;3)$	$2,35$	$2,3133$	$2,3132$	$2,3131$	$2,3131$	$2,313$	$2,3131$	$1,44 \times 10^{-8}$
$(18;3)$	$2,33$	$2,2951$	$2,2951$	$2,295$	$2,295$	$2,295$	$2,295$	$1,18 \times 10^{-9}$
$(2;4)$	$2,75$	$2,6051$	$2,6047$	$2,6043$	$2,6039$	$2,6035$	$2,6043$	$4,04 \times 10^{-7}$
$(5;4)$	$2,43$	$2,4608$	$2,4607$	$2,4606$	$2,4605$	$2,4604$	$2,4606$	$2,75 \times 10^{-8}$
$(12;4)$	$2,55$	$2,5196$	$2,5194$	$2,5192$	$2,519$	$2,5188$	$2,5192$	$1,07 \times 10^{-7}$
$(16;4)$	$2,50$	$2,4528$	$2,4529$	$2,453$	$2,4531$	$2,4532$	$2,453$	$2,86 \times 10^{-8}$
$(17;4)$	$2,43$	$2,492$	$2,4919$	$2,4918$	$2,4917$	$2,4916$	$2,4918$	$1,73 \times 10^{-8}$
$(5;5)$	$2,55$	$2,5383$	$2,5383$	$2,5383$	$2,5382$	$2,5382$	$2,5383$	$1,35 \times 10^{-9}$
$(8;5)$	$2,38$	$2,5053$	$2,5053$	$2,5054$	$2,5055$	$2,5055$	$2,5054$	$1,20 \times 10^{-8}$
$(18;5)$	$2,48$	$2,5282$	$2,5282$	$2,5282$	$2,5283$	$2,5283$	$2,5282$	$5,16 \times 10^{-10}$
$(19;5)$	$2,50$	$2,543$	$2,5431$	$2,5431$	$2,5431$	$2,5431$	$2,5431$	$1,63 \times 10^{-9}$
$(20;5)$	$2,28$	$2,4408$	$2,4411$	$2,4414$	$2,4417$	$2,4421$	$2,4414$	$2,49 \times 10^{-7}$
$(3;6)$	$1,93$	$1,9877$	$1,9875$	$1,9873$	$1,9871$	$1,987$	$1,9873$	$8,37 \times 10^{-8}$
$(5;6)$	$2,20$	$2,1534$	$2,1535$	$2,1535$	$2,1535$	$2,1536$	$2,1535$	$3,84 \times 10^{-9}$
$(10;6)$	$2,13$	$2,1149$	$2,1149$	$2,1148$	$2,1147$	$2,1146$	$2,1148$	$1,38 \times 10^{-8}$
$(2;7)$	$2,43$	$2,3987$	$2,3985$	$2,3984$	$2,3983$	$2,3981$	$2,3984$	$4,94 \times 10^{-8}$

Continua na próxima página...

Tabela 13 – Conclusão

Posição retirada	VO	Imputação (m) 1	2	3	4	5	$\bar{m}$	Var
$(3;7)$	$2,48$	$2,3674$	$2,3674$	$2,3674$	$2,3675$	$2,3675$	$2,3674$	$1,05 \times 10^{-9}$
$(7;7)$	$2,40$	$2,3401$	$2,3401$	$2,3402$	$2,3402$	$2,3402$	$2,3402$	$5,06 \times 10^{-9}$
$(9;7)$	$2,32$	$2,3526$	$2,3526$	$2,3525$	$2,3524$	$2,3524$	$2,3525$	$1,20 \times 10^{-8}$
$(10;7)$	$2,38$	$2,3585$	$2,3585$	$2,3584$	$2,3584$	$2,3583$	$2,3584$	$7,79 \times 10^{-9}$
$(11;7)$	$2,45$	$2,3762$	$2,3761$	$2,376$	$2,3759$	$2,3757$	$2,376$	$2,99 \times 10^{-8}$
$(14;7)$	$2,27$	$2,3122$	$2,3123$	$2,3124$	$2,3125$	$2,3126$	$2,3124$	$2,42 \times 10^{-8}$
$(15;7)$	$2,13$	$2,2944$	$2,2945$	$2,2946$	$2,2947$	$2,2948$	$2,2946$	$2,51 \times 10^{-8}$
$(16;7)$	$2,27$	$2,3374$	$2,3375$	$2,3375$	$2,3376$	$2,3376$	$2,3375$	$3,85 \times 10^{-9}$
$(20;7)$	$2,00$	$2,259$	$2,2593$	$2,2596$	$2,2599$	$2,2602$	$2,2596$	$2,14 \times 10^{-7}$

Os resultados da segunda etapa da IM, *Proc UNIVARIATE* do *SAS*, para cada uma das localidades completadas pelas 5 imputações, estão apresentados na Tabela 14.

Tabela 14 – Médias e erros padrão de alturas das localidades completadas pelas imputações, com 30% de dados faltantes.

Imputações	Localidade (média/erro padrão) 1	2	3	4	5	6	7
1	2,419275	2,635865	2,280419	2,477510	2,533780	2,080302	2,341833
	0,021435	0,018850	0,022892	0,027422	0,018526	0,022162	0,013412
2	2,419260	2,635866	2,280414	2,477475	2,533801	2,080291	2,341844
	0,021432	0,018840	0,022891	0,027416	0,018522	0,022164	0,013403
3	2,419244	2,635866	2,280409	2,477441	2,533821	2,080280	2,341855
	0,021430	0,018830	0,022891	0,027410	0,018518	0,022166	0,013394
4	2,419229	2,635867	2,280404	2,477406	2,533841	2,080270	2,341866
	0,021427	0,018820	0,022891	0,027404	0,018513	0,022168	0,013385
5	2,419214	2,635868	2,280399	2,477372	2,533861	2,080259	2,341877
	0,021425	0,018810	0,022890	0,027398	0,018509	0,022170	0,013377

As estimativas das localidades combinadas por meio da *Proc MIANALYZE* do *SAS* (Anexo C), estão apresentados na Tabela 15.

As estimativas das médias das localidades ($\hat{\beta}^*$), para 30% de ausência, continuaram com uma pequena variabilidade, isto é, com T próximos de zero. A falta de dados continuou

exercendo pouca influência em $\hat{\beta}^*$, pois, tanto λ quanto r continuaram com valores muito pequenos.

Pelo teste t-Student, não significativo para todas as localidades, tem-se uma semelhança entre cada média de alturas dos ambientes para os dados originais (MO) e sua respectiva estimativa média ($\hat{\beta}^*$) nos dados completados.

Tabela 15 – Estimativa média $\hat{\beta}^*$ das médias de alturas, medidas associadas a sua variabilidade (B,$\bar{W}$ e T) e Teste t-Student, com v_{obs} graus de liberdade, para comparação da média original nas localidades, com 30% de dados faltantes

	Localidades						
Estimativas	1	2	3	4	5	6	7
$\hat{\beta}^*$	$2,419244$	$2,635866$	$2,280409$	$2,477441$	$2,533821$	$2,080281$	$2,341855$
B	$5,8 \times 10^{-10}$	$2,05 \times 10^{-12}$	$5,9 \times 10^{-12}$	$2,9 \times 10^{-9}$	$1,0 \times 10^{-9}$	$2,9 \times 10^{-10}$	$3,0 \times 10^{-10}$
$\bar{W}$	$0,000459$	$0,000355$	$0,000524$	$0,000751$	$0,000343$	$0,000491$	$0,000179$
T	0.000459	0.000355	0.000524	$0,000751$	$0,000343$	$0,000491$	$0,000179$
r	$1,52 \times 10^{-6}$	6.94×10^{-9}	$1,36 \times 10^{-7}$	$4,75 \times 10^{-6}$	$3,52 \times 10^{-6}$	$7,26 \times 10^{-7}$	$2,04 \times 10^{-6}$
λ	$1,51 \times 10^{-6}$	6.94×10^{-9}	$1,36 \times 10^{-7}$	$4,75 \times 10^{-6}$	$3,52 \times 10^{-6}$	$7,26 \times 10^{-7}$	$2,04 \times 10^{-6}$
ER	$1,00$	$1,00$	$1,00$	$0,999999$	$0,999999$	$1,00$	$1,00$
v_{obs}	$17,273$	$17,273$	$17,273$	$17,273$	$17,273$	$17,273$	$17,273$
$\bar{Y}$	$2,4240$	$2,6625$	$2,2840$	$2,4840$	$2,5155$	$2,0805$	$2,3285$
t_{calc}	$-0,22$	$-1,41$	$-0,16$	$-0,24$	$0,99$	$-0,01$	$1,00$
$valor-p$	$0,8270$	$0,1750$	$0,8772$	$0,8137$	$0,3361$	$0,9922$	$0,3325$

O banco de dados completado, com 30% de *missing*, considerando a média das alturas dos 5 bancos gerados pela imputação, são expostos na Tabela 16.

Tabela 16 – Matriz de dados completada pela imputação com 30% de valores ausentes.

	Localidade						
Cultivar	1	2	3	5	5	6	7
1	2,43	2,67	2,35	2,55	2,55	2,17	2,3
2	2,65	2,83	2,35	2,6	2,72	2,22	2,4
3	2,47	2,63	2,2	2,62	2,58	1,99	2,37
4	2,45	2,67	2,32	2,63	2,55	1,98	2,37
5	2,37	2,72	2,35	2,46	2,54	2,15	2,28
6	2,3	2,53	2,32	2,4	2,43	2	2,27
7	2,45	2,62	2,15	2,47	2,45	2,17	2,34
8	2,38	2,63	2,32	2,45	2,51	2,1	2,32
9	2,42	2,53	2,42	2,55	2,55	2,2	2,35
10	2,45	2,68	2,31	2,55	2,6	2,11	2,36
11	2,47	2,68	2,37	2,58	2,53	2,23	2,38
12	2,53	2,67	2,4	2,52	2,58	2,2	2,47
13	2,45	2,6	2,25	2,43	2,52	2,05	2,4
14	2,38	2,52	2,18	2,42	2,52	2,05	2,31
15	2,18	2,55	2,25	2,42	2,37	2,02	2,29
16	2,4	2,63	2,23	2,45	2,47	2	2,34
17	2,43	2,63	2,25	2,49	2,7	2,08	2,23
18	2,43	2,77	2,3	2,25	2,53	2,05	2,43
19	2,47	2,67	2,32	2,57	2,54	1,93	2,37
20	2,28	2,5	1,97	2,13	2,44	1,9	2,26
Média	2,42	2,64	2,28	2,48	2,53	2,08	2,34

Na Tabela 17, são apresentadas as medidas globais de acurácia ou exatidão T_{acc} , para cada porcentagem de valores ausentes.

Tabela 17 – Medida geral da acurácia do método IMLD, com 5%, 15% e 30% de dados faltantes

Ausência	Acurácia geral		
	V_E	VQM	T_{acc}
5%	$2,043414 \times 10^{-8}$	$0,002000418$	$0,002000439$
15%	$8,017942 \times 10^{-8}$	$0,01629397$	$0,01629405$
30%	5.057484×10^{-8}	$0,009562117$	$0,009562168$

Composta pela variância entre imputações V_E e pelo viés quadrático médio VQM, valores pequenos para a T_{acc} indicam diferenças extremamente pequenas entre os valores originais e os valores imputados. É evidente também, que mesmo com o aumento da quantidade de $missing$, o método IMLD continuou muito preciso.

Além da precisão, o método de IMLD, não possui restrição quanto ao padrão e mecanismo de ausência dos dados faltantes, é livre de suposições sobre a distribuição ou estrutura dos dados e ainda utiliza a maior quantidade de informação disponível no banco de dados.

CONSIDERAÇÕES FINAIS

Este estudo apresentou metodologias para imputação múltipla de dados, especificamente o algoritmo MICE e o método IMLD, que se mostraram boas alternativas para o tratamento de $missing$. Os ajustes do modelo logístico com valores obtidos por imputação múltipla via algoritmo MICE , em geral, foram melhores do que as estimativas mantendo os valores perdidos. Com a IMLD, a baixa variabilidade da matriz dos valores de altura das cultivares em relação à média dos valores imputados, indicando uma boa precisão do processo de imputação.

Os resultados deste trabalho ainda não podem ser generalizados, pois foram obtidos para situações particulares, usando o banco de dados como uma ferramenta para estudar e divulgar as técnicas de imputação. Itens como tamanho amostral, tipo de variável e condição da relação entre as variáveis envolvidas devem ser estudados de modo mais específico. Portanto, situações que fogem dos cenários estudados, ficam como sugestão para estudos futuros, assim como o estudo de outros métodos e metodologias para a imputação.

Existe uma gama de métodos para lidar com dados omissos, tanto sob a perspectiva da imputação simples, como pela metodologia de imputação múltipla. No entanto, o emprego da imputação deve ser feito com cautela, pois requer uma avaliação de qual método é adequado, bem como a validação de seus pressupostos.

Para contribuir com a expansão da área de imputação de dados e consequentemente com melhorias nas análises estatísticas, é essencial a produção de mais trabalhos metodológicos que salientem a importância do tratamento de dados faltantes, avaliando a eficiência dos métodos, apontando os melhores e em quais situações devem ser empregados, de modo a facilitar a aplicação dos mesmos.

TRABALHOS FUTUROS

Para que o desempenho dos métodos de imputação apresentados neste trabalho seja melhor avaliado considera-se importante:

- Em estudos sobre o algoritmo MICE incluir mais variáveis com dados faltantes e verificar o comportamento desta técnica em muitas proporções distintas de *missing data*;
- Avaliar o método IMLD quando utilizado em dados com variáveis resposta de distribuição discreta e estudar o efeito de diferentes distribuições de probabilidade sobre a precisão do método.
- Efetuar avaliações do método IMLD em bases dados que apresentem casos *outliers* ou dados desbalanceados.

Referências

ACOCK, A. C. Working with missing values. *Journal of Marriage and Family*, Wiley Online Library, v. 67, n. 4, p. 1012–1028, 2005. 26

ARCINIEGAS-ALARCÓN, S.; DIAS, C. T. d. S. Imputação de dados em experimentos com interação genótipo por ambiente: uma aplicação a dados de algodão. *Revista Brasileira de Biometria*, v. 27, n. 1, p. 125–138, 2009. 45

ARCINIEGAS-ALARCÓN, S.; DIAS, C. T. d. S. Análise ammi com dados imputados em experimentos de interação genótipo x ambiente de algodão. *Pesquisa Agropecuária Brasileira*, v. 44, n. 11, p. 1391–1397, 2010. 45

ARCINIEGAS-ALARCÓN, S.; DIAS, C. T. d. S.; GARCÍA-PEÑA, M. Imputação múltipla livre de distribuição em tabelas incompletas de dupla entrada. *Pesquisa Agropecuária Brasileira*, v. 49, n. 9, p. 683–691, 2014. 46

ARCINIEGAS-ALARCÓN, S. et al. Imputing missing values in multi-environment trials using the singular value decomposition: An empirical comparison. *Communications in Biometry and Crop Science*, v. 9, n. 2, p. 54–70, 2014. 46

BARACHO, S. Tratamento de dados ausentes em estudos longitudinais. *Minas Gerais: Universidade Federal de Minas Gerais*, 2003. 29

BEAUJEAN, A. A.; BEAUJEAN, M. A. A. Package 'bayloredpsych'. 2012. 24

BERGAMO, G. C. *Imputaçao múltipla livre de distribuiçao utilizando a decomposiçao por valor singular em matriz de interaçao*. Tese (Doutorado) — Escola Superior de Agricultura "Luiz de Queiroz, 2007. 17, 18, 22, 42, 44, 45

BERGAMO, G. C.; DIAS, C. T. d. S.; KRZANOWSKI, W. J. Distribution-free multiple imputation in an interaction matrix through singular value decomposition. *Scientia Agricola*, SciELO Brasil, v. 65, n. 4, p. 422–427, 2008. 45

BRAND, J. *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*. [S.l.]: Erasmus MC: University Medical Center Rotterdam, 1999. 26

BROWN, M. B.; DIXON, W. J. *BMDP statistical software*. [S.l.]: Univ of California Press, 1983. 24

BUHI, E. R.; GOODSON, P.; NEILANDS, T. B. Out of sight, not out of mind: strategies for handling missing data. *American journal of health behavior*, PNG Publications, v. 32, n. 1, p. 83–92, 2008. 26

BUUREN, S. V. *Flexible imputation of missing data*. [S.l.]: CRC press, 2012. 17, 18, 20, 21, 27, 31

BUUREN, S. V. et al. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, v. 18, n. 6, p. 681–694, 1999. 29

BUUREN, S. V.; GROOTHUIS-OUDSHOORN, K. Mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, American Statistical Association, v. 45, n. 3, 2011. 18, 31

BUUREN, S. V.; OUDSHOORN, C. Multivariate imputation by chained equations. *MICE V1. 0 user's manual. Leiden: TNO Preventie en Gezondheid*, 2000. 18, 31

BUUREN, S. V.; OUDSHOORN, K. Flexible multivariate imputation by mice. *Leiden, The Netherlands: TNO Prevention Center*, 1999. 35

CALINSKI, T. et al. Em and als algorithms applied to estimation of missing data in series of variety trials. *Biuletyn Oceny Odmian*, v. 24, p. 25, 1992. 45

CAMARGOS, V. P. et al. Imputação múltipla e análise de casos completos em modelos de regressão logística: uma avaliação prática do impacto das perdas em covariáveis. *Cad Saúde Pública*, SciELO Brasil, v. 27, p. 2299–313, 2011. 37

CASCAES, A. M. et al. Prematuridade e fatores associados no estado de santa catarina, brasil, no ano de 2005: análise dos dados do sistema de informações sobre nascidos vivos. *Cad Saúde Pública*, SciELO Brasil, v. 24, n. 5, p. 1024–32, 2008. 34

COLLINS, L. M.; SCHAFER, J. L.; KAM, C.-M. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, American Psychological Association, v. 6, n. 4, p. 330, 2001. 23

DENIS, J.; BARIL, C. Sophisticated models with numerous missing values: the multiplicative interaction model as an example. *Biuletyn Oceny Odmian*, v. 24, n. 25, p. 33–45, 1992. 45

DIAS, C. T. d. S.; KRZANOWSKI, W. J. Model selection and cross validation in additive main effect and multiplicative interaction models. *Crop Science*, Crop Science Society of America, v. 43, n. 3, p. 865–873, 2003. 45

ENDERS, C. K. *Applied missing data analysis*. [S.l.]: Guilford Publications, 2010. 17, 18, 20, 26

GIGLIO, M. R. P. et al. Baixo peso ao nascer em coorte de recém-nascidos em goiânia-brasil no ano de 2000. *Rev Bras Ginecol Obstet*, SciELO Brasil, v. 27, n. 3, p. 130–6, 2005. 34

GONZALEZ, J. M.; ELTINGE, J. L. Multiple matrix sampling: A review. In: *Proceedings of the Section on Survey Research Methods, American Statistical Association*. [S.l.: s.n.], 2007. p. 3069–3075. 20

GOOD, I. J. Some applications of the singular decomposition of a matrix. *Technometrics*, Taylor & Francis Group, v. 11, n. 4, p. 823–831, 1969. 42

GRAHAM, J. W.; OLCHOWSKI, A. E.; GILREATH, T. D. How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science*, Springer, v. 8, n. 3, p. 206–213, 2007. 29

HILBE, J. M. *Logistic regression models*. [S.l.]: CRC Press, 2009. 34

HORTON, N. J.; KLEINMAN, K. P. Much ado about nothing:a comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, v. 61, n. 1, 2007. 18, 26

JAMSHIDIAN, M.; JALAL, S. Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika*, Springer, v. 75, n. 4, p. 649–674, 2010. 23

JAMSHIDIAN, M.; JALAL, S. J.; JANSEN, C. Missmech: an r package for testing homoscedasticity, multivariate normality, and missing completely at random (mcar). *Journal of Statistical Software*, JOURNAL STATISTICAL SOFTWARE, v. 56, n. 6, 2014. 23

JAMSHIDIAN, M.; MATA, M. Postmodeling sensitivity analysis to detect the effect of missing data mechanisms. *Multivariate Behavioral Research*, Taylor & Francis, v. 43, n. 3, p. 432–452, 2008. 23

JR, H. G. G.; ZOBEL, R. W. Imputing missing yield trial data. *Theoretical and Applied Genetics*, Springer, v. 79, n. 6, p. 753–761, 1990. 46

KRZANOWSKI, W. Missing value imputation in multivariate data using the singular value decomposition of a matrix. *Biometrical letters*, v. 25, n. 1-2, p. 31–39, 1988. 42, 44

LITTLE, R. J. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, Taylor & Francis, v. 83, n. 404, p. 1198–1202, 1988. 24

LITTLE, R. J. Regression with missing x's: a review. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 87, n. 420, p. 1227–1237, 1992. 23

LITTLE, R. J.; RUBIN, D. B. *Statistical analysis with missing data*. [S.l.]: John Wiley & Sons, 1987. 18, 26

LITTLE, R. J.; RUBIN, D. B. *Statistical analysis with missing data*. [S.l.]: John Wiley & Sons, 2014. 26

MOLENBERGHS, G.; KENWARD, M. *Missing data in clinical studies*. [S.l.]: John Wiley & Sons, 2007. 26

MOLENBERGHS, G.; VERBEKE, G. *Models for discrete longitudinal data*. [S.l.]: Springer Science & Business Media, 2006. 29

MYERS, W. R. Handling missing data in clinical trials: an overview. *Drug Information Journal*, SAGE Publications, v. 34, n. 2, p. 525–533, 2000. 26

NUNES, L. N.; KLÜCK, M. M.; FACHEL, J. M. G. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. *Cad. Saúde Pública*, SciELO Public Health, v. 25, n. 2, p. 268–278, 2009. 37

OUDSHOORN, K.; BUUREN, S. V.; RIJCKEVORSEL, J. V. *Flexible multiple imputation by chained equations of the AVO95 Survey. Leiden: TNO Prevention and Health*. [S.l.], 1999. 31

PEREIRA, E. A. Algumas propostas para imputação de dados faltantes em teoria de resposta ao item. 2014. 23

PIRDAWD, H. Q. *Multiple Imputation Method for Missing Data in (RCBD) Factorial Experiments*. Tese (Doutorado) — University of Sulaimani, 2007. 33

RAGHUNATHAN, T. E. What do we do with missing data? some options for analysis of incomplete data. *Annu. Rev. Public Health*, Annual Reviews, v. 25, p. 99–117, 2004. 17, 26

ROTH, P. L.; SWITZER, F. S.; SWITZER, D. M. Missing data in multiple item scales: A monte carlo analysis of missing data techniques. *Organizational Research Methods*, Sage Publications, v. 2, n. 3, p. 211–232, 1999. 17

RUBIN, D. B. Inference and missing data. *Biometrika*, Biometrika Trust, v. 63, n. 3, p. 581–592, 1976. 17, 18, 21

RUBIN, D. B. Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In: AMERICAN STATISTICAL ASSOCIATION. *Proceedings of the survey research methods section of the American Statistical Association*. [S.l.], 1978. v. 1, p. 20–34. 18, 29

RUBIN, D. B. *Multiple imputation for nonresponse in surveys*. [S.l.]: John Wiley & Sons, 1987. 18, 22, 27, 28, 29, 30

RUBIN, D. B.; LITTLE, R. J. Statistical analysis with missing data. *Hoboken, NJ: J Wiley & Sons*, 2002. 20

SCHAFER, J. L. *Analysis of incomplete multivariate data*. [S.l.]: CRC press, 1997. 26

SCHAFER, J. L.; GRAHAM, J. W. Missing data: our view of the state of the art. *Psychological methods*, American Psychological Association, v. 7, n. 2, p. 147, 2002. 17, 23, 26

SCHEFFER, J. Dealing with missing data. Massey University, 2002. 26

SHIOGA, S. P. et al. *Avaliação estadual de cultivares de milho segunda safra 2015*. [S.l.], 2015. 8, 9, 46

SILVA, M. J. C. da. *Imputação múltipla: comparação e eficiência em experimentos multiambientais*. Tese (Doutorado) — Escola Superior de Agricultura "Luiz de Queiroz, 2012. 25, 45

TSIKRIKTSIS, N. A review of techniques for treating missing data in om survey research. *Journal of Operations Management*, Elsevier, v. 24, n. 1, p. 53–62, 2005. 26

YAN, W. Biplot analysis of incomplete two-way data. *Crop Science*, The Crop Science Society of America, Inc., v. 53, n. 1, p. 48–57, 2013. 46

ZHUOFAN, W. *Proposta de um modelo de regressão binária com resposta contínua aplicado à análise dos dados do SINASC: identificação de fatores de risco para o baixo peso ao nascer*. Tese (Doutorado) — Universidade de São Paulo, 2011. 34

Anexos

ANEXO A

Programa no *software R*, para aplicação do método de Imputação Múltipla Livre de Distribuição (IMLD)

```
#-------------------------------------------------------------------------#
rm(list=ls())
#-------------------------------DADOS-------------------------------------#
y<-c(2.43, 2.85,2.35,2.55,2.55,2.17,2.30,
2.65,2.83,2.35,2.75,2.72,2.22,2.43,
2.48,2.63,2.20,2.62,2.58,1.93,2.48,
2.50,2.73,2.32,2.63,2.55,1.98,2.37,
2.37,2.72,2.35,2.43,2.55,2.20,2.28,
2.30,2.53,2.32,2.40,2.43,2.00,2.27,
2.45,2.67,2.15,2.47,2.45,2.17,2.40,
2.35,2.67,2.32,2.45,2.38,2.10,2.32,
2.50,2.53,2.42,2.55,2.55,2.20,2.32,
2.45,2.68,2.35,2.55,2.60,2.13,2.38,
2.55,2.80,2.37,2.58,2.53,2.23,2.45,
2.53,2.65,2.40,2.55,2.58,2.20,2.47,
2.45,2.60,2.25,2.43,2.52,2.05,2.40,
2.28,2.52,2.18,2.42,2.52,2.05,2.27,
2.18,2.55,2.25,2.42,2.37,2.02,2.13,
2.40,2.68,2.23,2.50,2.47,2.00,2.27,
2.43,2.68,2.25,2.43,2.70,2.08,2.23,
2.43,2.77,2.33,2.25,2.48,2.05,2.43,
2.47,2.78,2.32,2.57,2.50,1.93,2.37,
```

```
2.28,2.38,1.97,2.13,2.28,1.90,2.00)
Y<-matrix(y,ncol = 7,nrow = 20,byrow=T)

m_L<-apply(Y,2, mean)

#DEFININDO A QUANTIDADE DE VALORES AUSENTES#
L<-nrow(Y)
C<-ncol(Y)
PORC=5 #porcentagem de valores ausentes

NOM=(L*C)*(PORC/100)
#quantidade de valores ausentes
NOM<-ifelse(NOM<1,ceiling(NOM),floor(NOM))

#RETIRANDO VALORES DO BANCO ORIGINAL#
Y1<-Y

set.seed(91425)
s<-seq(L*C)
mis<-sample(s,NOM,replace = F)
Y1[mis]=NA

#IDENTIFICANDO LINHA E COLUNA DOS VALORES RETIRADOS#

for(i in 1:nrow(Y1))
{
for(j in 1:ncol(Y1))
{
if(is.na(Y1[i,j]))
{
cat("O elemento  ",i,j,"de Y é perdido","\n")
}
}
}

# SUBSTITUINDO OS VALORES AUSENTES PELA MÉDIA DA RESPECTIVA COLUNA#
Y2<-Y1
medias_col<- colMeans(Y2,na.rm=T)
for(i in 1:nrow(Y2))
```

```
{
for(j in 1:ncol(Y2))
{
if(is.na(Y2[i,j]))
{
Y2[i,j]=medias_col[j]
cat("o elemento",i,j,"foi substituído por", medias_col[j], "\n")
}
}
}

## PADRONIZAÇÃO#
media.desvio<- function(x, perdido) {
if (is.numeric(x)) {
c(media = mean(x, na.rm = perdido), desvio = sd(x, na.rm = perdido))
}
}

m_sd<-apply(X = Y2, MARGIN = 2, FUN = media.desvio, perdido = T)

YP<-matrix(nrow = nrow(Y), ncol = ncol(Y))
for(i in 1:nrow(Y2))
{
for(j in 1:ncol(Y2))
{
YP[i,j]<-(Y2[i,j]-m_sd[1,j])/(m_sd[2,j])
}
}

#DECOMPOSICão EM VALORES SINGULARES#
s<-svd(YP)
D<-diag(s$d)
U<-s$u
V<-s$v

A<-U%*%D%*%(t(V)) #verificando a decomposição

#DEFINE A SUBMATRIZ Y_i#
Y_i<-list(matrix(nrow =(nrow(Y)-1), ncol=ncol(Y)))
```

```
LF<-matrix(nrow =nrow(Y),ncol=ncol(Y))

for(i in 1:nrow(Y2))
{
for(j in 1:ncol(Y2))
{
if (is.na(Y1[i,j]))
{
LF[i,j]<-i
}
}
}

LF<-c(t(LF))
LF<-subset(LF,is.na(LF)==FALSE)

for (i in 1:length(LF))
{
Y_i[[i]]<-Y[-LF[i],]
}

#PADRONIZA A MATRIZ Y_i#
m_sd_i<- list(matrix(nrow =2, ncol=(nrow(Y)-1)))

media.desvio <- function(x, perdido) {
if (is.numeric(x)) {
c(media = mean(x, na.rm = perdido), desvio = sd(x, na.rm = perdido))
}
}

for (i in 1:length(Y_i))
{
m_sd_i[[i]]<-apply(X = Y_i[[i]], MARGIN = 2, FUN = media.desvio, perdido = T)
}

for (i in 1:length(Y_i))
{
for(j in 1:nrow(Y_i[[i]]))
{
```

```
for(k in 1:ncol(Y_i[[i]]))
{
Y_i[[i]][j,k]<-(Y_i[[i]][j,k]-m_sd_i[[i]][1,k])/(m_sd_i[[i]][2,k])
}
}
}

#DEFINE A SUBMATRIZ Y_j#
Y_j<-list(matrix(nrow =nrow(Y), ncol=(ncol(Y)-1)))
CF<-matrix(nrow =nrow(Y), ncol=(ncol(Y)))

for(i in 1:nrow(Y2))
{
for(j in 1:ncol(Y2))
{
if (is.na(Y1[i,j]))
{
CF[i,j]<-j
}
}
}

CF<-c(t(CF))
CF<-subset(CF,is.na(CF)==FALSE)

for (i in 1:length(CF))
{
Y_j[[i]]<-Y[,-CF[i]]
}

#PADRONIZA A MATRIZ Y_j#
m_sd_j<- list(matrix(nrow =2, ncol=nrow(Y)))

for (i in 1:length(Y_j))
{
m_sd_j[[i]]<-apply(X = Y_j[[i]], MARGIN = 2, FUN = media.desvio, perdido = T)
}

for (i in 1:length(Y_j))
```

```
{
for(j in 1:nrow(Y_j[[i]]))
{
for(k in 1:ncol(Y_j[[i]]))
{
Y_j[[i]][j,k]<-(Y_j[[i]][j,k]-m_sd_j[[i]][1,k])/(m_sd_j[[i]][2,k])
}
}
}

#DECOMPOSICão EM VALORES SINGULARES DA MATRIZ Y_i#
sa<-list()
DA<-list()
UA<-list()
VA<-list()

for (i in 1:length(Y_i))
{
sa[[i]]<- svd(Y_i[[i]])
DA[[i]] <- diag(sa[[i]]$d)
UA[[i]]=sa[[i]]$u
VA[[i]]=sa[[i]]$v
}

#DECOMPOSICão EM VALORES SINGULARES DA MATRIZ Y_j#
sb<-list()
DB<-list()
UB<-list()
VB<-list()

for (i in 1:length(Y_j))
{
sb[[i]]<- svd(Y_j[[i]])
DB[[i]] <- diag(sb[[i]]$d)
UB[[i]]=sb[[i]]$u
VB[[i]]=sb[[i]]$v
}

#ELIMINAR O ÚLTIMO ELEMENTO DE 'DA' E A ÚLTIMA COLUNA DE 'VA'#
```

```
DA_<-list(matrix(nrow =(nrow(DA[[1]])-1), ncol=(ncol(DA[[1]])-1)))

for (i in 1:length(DA))
{
DA_[[i]]<-DA[[i]][1:(nrow(DA[[i]])-1),1:(ncol(DA[[i]])-1)]
}

VA_<- list(matrix(nrow = nrow(VA[[1]]), ncol=(ncol(VA[[1]])-1)))

for (i in 1:length(VA))
{
VA_[[i]]<-t(VA[[i]][ ,1:(ncol(VA[[i]])-1)])
}

#CRIANDO AS IMPUTAÇÕES MÚLTIPLAS#
IMPMULT<-list(matrix(nrow = nrow(Y),ncol = ncol(Y)))

n<-c(8:12)
q<-c(1:NOM)
imp<-c(1:5)
p<-matrix(1:(length(q)*length(imp)),ncol = length(imp),nrow = length(q),byrow=T)

for (a in 1:length(q))
{
for (i in 1:length(imp))
{
IMPMULT[[(p[a,i])]]<-UB[[a]]%*%(DB[[a]]^(n[i]/20))%*%(DA_[[a]]^((20-n[i])/20))%*%
}
}

#VOLTANDO PARA A ESCALA ORIGINAL#
for (a in 1:length(q))
{
for (b in 1:length(imp))
{
for ( i in 1:nrow(Y))
{
for (j in 1:ncol(Y))
{
```

```
IMPMULT[[(p[a,b])]] [i,j]<-(m_sd[1,j] + (m_sd[2,j]*(IMPMULT[[(p[a,b])]][i,j])))
}
}
}
}

#CONSTRUINDO OS 5 BANCOS DE DADOS IMPUTADOS#
impmult1<-list()
impmult2<-list()
impmult3<-list()
impmult4<-list()
impmult5<-list()

#expoente 1/8
for (i in 1:length(p[,1]))
{
impmult1[[i]]<-IMPMULT[[(p[i,1])]]
}

impmult1<-apply(simplify2array(impmult1), 1:2, mean)

IMPMULT1<-matrix(nrow = nrow(Y), ncol = ncol(Y))

for (i in 1:nrow(Y))
{
for (j in 1:ncol(Y))
{
ifelse(is.na(Y1[i,j]),IMPMULT1[i,j]<-impmult1[i,j],IMPMULT1[i,j]<-Y[i,j])
}
}

#expoente 1/9
for (i in 1:length(p[,2]))
{
impmult2[[i]]<-IMPMULT[[(p[i,2])]]
}

impmult2<-apply(simplify2array(impmult2), 1:2, mean)
```

```
IMPMULT2<-matrix(nrow = nrow(Y), ncol = ncol(Y))

for (i in 1:nrow(Y))
{
for (j in 1:ncol(Y))
{
ifelse(is.na(Y1[i,j]),IMPMULT2[i,j]<-impmult2[i,j],IMPMULT2[i,j]<-Y[i,j])
}
}

#expoente 1/10
for (i in 1:length(p[,3]))
{
impmult3[[i]]<-IMPMULT[[(p[i,3])]]
}

impmult3<-apply(simplify2array(impmult3), 1:2, mean)

IMPMULT3<-matrix(nrow = nrow(Y), ncol = ncol(Y))

for (i in 1:nrow(Y))
{
for (j in 1:ncol(Y))
{
ifelse(is.na(Y1[i,j]),IMPMULT3[i,j]<-impmult3[i,j],IMPMULT3[i,j]<-Y[i,j])
}
}

#expoente 1/11
for (i in 1:length(p[,4]))
{
impmult4[[i]]<-IMPMULT[[(p[i,4])]]
}

impmult4<-apply(simplify2array(impmult4), 1:2, mean)

IMPMULT4<-matrix(nrow = nrow(Y), ncol = ncol(Y))

for (i in 1:nrow(Y))
```

```
{
for (j in 1:ncol(Y))
{
ifelse(is.na(Y1[i,j]),IMPMULT4[i,j]<-impmult4[i,j],IMPMULT4[i,j]<-Y[i,j])
}
}

#expoente 1/12
for (i in 1:length(p[,5]))
{
impmult5[[i]]<-IMPMULT[[(p[i,5])]]
}

impmult5<-apply(simplify2array(impmult5), 1:2, mean)

IMPMULT5<-matrix(nrow = nrow(Y), ncol = ncol(Y))

for (i in 1:nrow(Y))
{
for (j in 1:ncol(Y))
{
ifelse(is.na(Y1[i,j]),IMPMULT5[i,j]<-impmult5[i,j],IMPMULT5[i,j]<-Y[i,j])
}
}

#ARQUIVO COM OS 5 BANCOS COMPLETOS#

Imputação<-rep(c(1,2,3,4,5),each=nrow(Y))
dataset<-rbind(IMPMULT1,IMPMULT2,IMPMULT3,IMPMULT4,IMPMULT5)
dataset5<-cbind(Imputação,dataset)
colnames(dataset5) <- c("Imputacao","L1","L2","L3","L4","L5","L6","L7")

dataset5<-round(dataset5,10)
write.table(dataset5,file="C:\\Users\\mglb\\Desktop\\Imputação\\IMLD\\dados
\\imp5.csv",sep=",",dec=".",col.names=T, row.names=F)



#CONSTRUINDO O BANCO DE DADOS COM AS IMPUTAÇÕES#
IMPMEDIA<-matrix(nrow = nrow(Y),ncol = ncol(Y))
```

```
IMPMEDIA<-(IMPMULT1 + IMPMULT2 + IMPMULT3 + IMPMULT4 + IMPMULT5)/5
IMPMEDIA<-round(IMPMEDIA,2)

media_5<-apply(IMPMEDIA,2,mean)

write.table(IMPMEDIA,file="C:\\Users\\mglb\\Desktop\\Imputação\\IMLD
\\dados\\media_5.csv",sep=";",dec=",",col.names=T, row.names=F)
#-------------------------------------------------------------------------#



#-------------------------------------------------------------------------#
#-----------------------MEDIDAS DE ACURÁCIA-------------------------------#

#-----------------------ORGANIZAÇÃODOS DADOS------------------------------#
NAS<-sort(mis)
vo<-Y[NAS]
imput1<-IMPMULT1[NAS]
imput2<-IMPMULT2[NAS]
imput3<-IMPMULT3[NAS]
imput4<-IMPMULT4[NAS]
imput5<-IMPMULT5[NAS]

imput<-data.frame(imput1,imput2,imput3,imput4,imput5)

media<-apply(imput,1, mean)
sd<-apply(imput,1,sd)
Var<-(sd)^2

m_imput<-data.frame(vo,imput1,imput2,imput3,imput4,imput5,media)
m_imput<-round(m_imput,4)

m_imput<-data.frame(m_imput,Var)
attach(m_imput)

write.table(m_imput,file="C:\\Users\\mglb\\Desktop\\Imputação\\IMLD
\\dados\\vo_5.csv",sep=";",dec=",",col.names=T, row.names=F)

#--------------VARIÂNCIA ENTRE AS IMPUTAÇÕES(VE)--------------------------#
```

```
VE<-sum(Var)/length(vo)

#------------------VIÉS QUADRÁTICO MÉDIO (VQM)----------------------------#

m=5
na=length(vo)
total=((media-vo)^2)
vies=m*(sum(total))
VQM=vies/((m-1)*na)

#--------------------MEDIDA GERAL DE EXATIDÃO-----------------------------#

tacc=VE+VQM
```

ANEXO B

Programa no SAS, para calcular a média e o erro padrão de cada localidade nos cinco (M=5) conjuntos de dados completados

```
proc univariate data=imp5 noprint;
var L1 L2 L3 L4 L5 L6 L7;
output out=msd mean= L1 L2 L3 L4 L5 L6 L7
stderr= SL1 SL2 SL3 SL4 SL5 SL6 SL7;
by Imputacao;
run;
```

ANEXO C

Programa no SAS, para combinar as médias de alturas das localidades dos cinco conjuntos de dados completados

```
proc mianalyze data=msd edf=19 mu0= 2.4240 2.6625 2.2840 2.4840
2.5155 2.0805 2.3285;
modeleffects L1 L2 L3 L4 L5 L6 L7;
stderr SL1 SL2 SL3 SL4 SL5 SL6 SL7;
run;
```

ANEXO D

Artigo submetido à Revista Acta Scientiarum. Health Sciences, ISSN 1679-9291 (impresso) e ISSN 1807-8648 (on-line), publicada semestralmente pela Editora da Universidade Estadual de Maringá-Eduem.

Multiple Imputation in Logistic Regression Models: factors associated to the low weight at birth in the Parana State

Marina Gandolfi
Maringa's State University - marinagandolfi@hotmail.com

Sérgio Marcussi Gaspechak
Maringa's State University - smgaspechak@gmail.com

Eraldo Schunk Silva
Maringa's State University - eraldoschunk@gmail.com

Isolde Previdelli
Maringa's State University - isoldeprevidelli@gmail.com

Abstract

It is common in statistical analyzes the occurrence of incomplete databases. Generally, in these situations, it restricts the analysis to subjects with complete data, reducing the sample size that can produce biased estimates. The "filling" of the missing data can be done by multiple imputation method (MI). In a cross-sectional study of living newborns in the Paraná state, in 2012, it was obtained a sample of 3380 cases of Live Birth Information System (SINASC-PR), whose inclusion criterion was to select only records with the information complete. It was adjusted one logistic regression model for the outcome of low birth weight and this model was considered the gold standard. By simulation, to generate three sets of incomplete data, 5%, 10% and 20% missing data for the weight outcome. Models with missing data and imputed data were compared with the gold standard model. Even in a simplistic approach to multiple imputation, we can see through the estimates and their standard errors, a better adjustment of the models with imputation.

Keywords: Logistic regression; Low weight at birth; Missing data.

Resumo

É comum em análises estatísticas a ocorrência de bases de dados incompletas. Geralmente, nessas situações, restringe-se a análise aos sujeitos com dados completos, reduzindo o tamanho amostral que pode produzir estimativas tendenciosas. O "preenchimento" dos dados faltantes pode ser feito pelo método de imputação múltipla (IM). Em um estudo transversal de recém-nascidos vivos no estado de Paraná, no ano de 2012, foi obtida uma amostra de 3380 casos do Sistema de Informações sobre Nascidos Vivos (SINASC- PR), cujo critério de inclusão foi selecionar apenas registros com as informações completas. Foi ajustado um modelo de regressão logística para o desfecho baixo peso ao nascer e este modelo foi considerado padrão ouro. Por simulação, foram gerados três conjuntos de dados incompletos, com 5%, 10% e 20% de dados faltantes para o desfecho peso. Os modelos com dados faltantes e com os dados imputados foram comparados com o modelo padrão ouro. Mesmo em uma abordagem simplista da imputação múltipla, percebe-se, por meio das estimativas e seus respectivos erros-padrão, um melhor ajuste dos modelos com imputação.

Palavras-chave: Baixo peso ao nascer; Dado faltante; Regressão logística.

INTRODUCTION

It is common in scientific research the occurrence of missing values in databases. Present both in experimental and observational studies (Nunes, Klück & Fachel, 2009), (Camargos et al., 2011), incomplete databases may be generated in many circumstances: information not provided by the respondent, unavailable measures due to the death of some animals or damaged plants (Bergamo, 2007). Also, values are lost as a result of failures arising from the measurement step of the characteristics of interest (Schafer & Graham, 2002). The ideal condition in these cases would be to repeat the study to obtain new values and supply the missing data, but in practice, this is often impractical due to the limited financial resources or time.

There are several complications caused by missing data, for example, reduction of sample size, efficiency loss and trouble in handling and analyzing data and bias in estimates when the treatment of missing values is done improperly or their presence is ignored (Roth, Switzer & Switzer, 1999). There are specific analytical techniques in order that inference with "missing" becomes valid, however, despite the increasing methodological development in this area, it is common to find inadequate treatments for missing data.

The commonly used statistical software have as default a procedure called listwise exclusion or analysis of complete cases (ACC) (Buuren, 2012). The process eliminates all cases with one or more missing values in the variables and thus restricts the analysis to the cases completely observed (Enders, 2010). The great advantage of the full analysis process is convenience, however, biased estimates can be produced by inducing erroneous decision making (Raghunathan, 2004)

The first statistical techniques designed to restore missing data by verisimilar values to them emerged in the late 70s, called "imputation methods" (Rubin, 1976). Among these techniques is the fulfilment of missing data by the average or the variable's median, imputation by the nearest neighbor, hot deck imputation, imputation by linear regression and imputation by maximum likelihood. In these techniques, called simple imputation methods, the missing data is completed only once and then the full database is used for analysis. The fact that the missing data is imputed only once, makes the uncertainty associated with the procedure not aggregated to the estimates generated by the full database, bringing a major limitation to these methods (Enders, 2010).

Facing the need to control the bias associated with simple imputation, Rubin (1978) proposed a new method, the Multiple Imputation (MI), which is composed of three main steps:

imputation, analysis and clustering. Each missing value is replaced by a set of plausible values, representing the uncertainty over the amount to be attributed. Publications of articles (Rubin, 1976), (Rubin, 1978) and books (Rubin, 1987), (Little & Rubin, 1987), made by the precursor Donald Bruce Rubin, served as a support so that more researchers developed work with multiple imputation, however, only recently this technique has been used with more tenacity, due to the computational developments for its implementation.

Currently, multiple imputation is available in the main commercial or free statistical software, R, SAS, Stata. A good review of these implementations is presented by Horton and Kleinman (2007), comparing results and providing instructions with syntax for each software. However most of the implemented methods are parametric, and in these, usually there are strong assumptions over the distribution of data, most of the assumption is that the data follow a multivariate normal distribution, but in practice, data may be distorted, limiting the applicability of the imputation methods (Buuren, 2012).

Given the difficulty in meeting the required assumptions for parametric methods, we deal with procedures to performing multiple imputation which offer greater flexibility to the data distribution: the algorithm MICE (Roth, Switzer & Switzer, 1999), (Buuren, 2012) - Multivariate imputation algorithm by Chained Equations - a Markov chain Monte Carlo method (MCMC). Buuren (2012) presents several possibilities for single and multiple imputation with the MICE algorithm, through the "mice" software package R.

Constantly, review articles and tutorials have been published, supporting researchers in the treatment of missing data. Schafer and Graham (2002), Little and Rubin (2014) present a broad review of existing methods for dealing with missing values, indicating the conditions under which valid results are produced. Other review work are found: Myers (2000) and Tsikriktsis (2005). In Scheffer (2002), Acock (2005), Buhi, Goodson and Neilands (2008) are compared some of imputation techniques available in major statistical packages.

Due to frequent recurrence to the issue of missing data by researchers in the health area, some authors of the branch have approached the subject (Nunes, Klück & Fachel, 2009), (Camargos et al., 2011). This work is a cross-sectional study, which refers to living newborn residents born in Parana state, in 2012. A logistic model adjusted to complete data is considered as gold pattern, and then three scenarios are simulated with 5%, 10% and 20% missing data. Multiple imputation methodology is employed to such data. The models are adjusted with missing records, with the result of multiple imputation and compared with the results obtained in each setting with the gold pattern. It is intended to disclose the multiple imputation method, show-

ing that the researcher will gain by adopting this methodology, compared with the restriction of the analysis to complete cases, and thus it becomes commonly used in problems with missing observations.

METHODOLOGY

There are several concepts of missing data theory that should be taken into account when applying any method of imputation, as stated by Rubin and Little (2014) "an imputation without criteria can create more problems than they solve, distorting estimates, standard errors and hypothesis testing". Among the important concepts the mechanism of missing data is highlighted. The mechanism describes possible relationships between the measured variables and the probability of lacking data (Enders, 2010).

Missing Data Mechanisms

The occurrence of missing data in data bases normally obeys a mechanism that indicates the conditions of the missing data`s generation. The main terminology classification of mechanisms was created Rubin (1976), and defines three general theoretical mechanisms extensively used in the literature: Missing Completely at Random - MCAR; Missing at Random – MAR and Not Missing at Random - NMAR.

To explain each mechanism consider $\boldsymbol{Y}$ an array of data collected with m lines, which represent individuals, n columns, representing the variables, with $\boldsymbol{y_{ij}=(y_{i1},...,y_{in})}$, where $\boldsymbol{y_{ij}}$ is the value of the variable j for the object i. $\boldsymbol{Y}$ can be divided into two subsets $\boldsymbol{Y=\{Y_{obs},Y_{mis}\}}$, where $\boldsymbol{Y_{obs}}$ are the observed data and $\boldsymbol{Y_{mis}}$ are the missing data. Corresponding Y data matrix there is a missing data identifier, a $\boldsymbol{R}$ matrix of the same dimension of $\boldsymbol{Y}$, where $r_{ij}=1$, if y_{ij} is observed, and $r_{ij}=0$, otherwise (Buuren, 2012).

The distribution of **R** can depend on $\boldsymbol{Y=\{Y_{obs},Y_{mis}\}}$, by design or by random, and this relationship is described by the missing data model. The general expression of the missing date model is $Pr(R|Y_{obs},Y_{mis},\psi)$, and ψ contains the parameters of the missing data model.

The data are referred to as MCAR if $Pr(R=0|Y_{obs},Y_{mis},\psi)=Pr(R=0|\psi)$ meaning that the probability of an item showing absent responses does not depend on any of the quantities observed or not observed. This implies that the missing data occurrence probability is the same for all cases. The data are said to be MCAR if

$Pr(R=0|Y_{obs},Y_{mis},\psi)=Pr(R=0|Y_{obs},\psi)$, ie, the missing data depend only on observed information available for analysis and correlated with the variable that has lost values.. And, there is MNAR data if $Pr(R=0|Y_{obs},Y_{mis},\psi)$, in this situation the probability of missing also depends on the non observed information, i.e., the probability of having a given missing varies for reasons which are unknown (Roth, Switzer & Switzer, 1999)

The impact of each mechanism in the analysis produced by different methods have been mostly evaluated by simulation studies, as in Little (1992), Collins, Schafer and Kam (2001), Schafer and Graham (2002). For MCAR and MAR missing data mechanism many treatment methods have been applied, however for the pattern NMAR appropriate methods are still not defined.

Identifying the missing data mechanism is not a simple task. Several tests have been proposed to test MCAR **vs.** MAR**.** The LittleMCAR(x) function in the R software, part of the package *BaylorEdPsych*, Beaujean and Beaujean (2012), uses the "Little Test", proposed by Little (1988), to assess the completely random missing (MCAR) in multivariate data.

In Jamshidian Jalal and Jansen (2014) it is presented that the MissMech package in *R* software, which are implemented methods to test the hypothesis MCAR, proposed by Jamshidian and Jalal (2010). The main focus of the package is to test MCAR, but it performs other tasks, multivariate normality tests, tests for homoscedasticity and normality to complete data, to obtain maximum likelihood estimates of mean and covariance (including standard error), for incomplete data using the *EM* algorithm, among others.

Jamshidian and Mata (2008) and use the available data to examine the sensitivity of a particular model to the data mechanism missing, for cases in which the researcher does not have MCAR (or MAR). The authors provide a specific method to perform a sensitivity analysis "postmodeling" using a statistical test and graphics.

Method of Multiple Imputation

The imputation is the fulfilment of missing data with plausible values for later analysis of the complete data. It can be simple when only one value is put for each missing data, or multiple, when there is more than one value for each missing data.

Multiple imputation (Rubin, 1987), is a theme in great expansion in statistics and can be summarized in three main steps, imputation, analysis and pooling.

I. The analysis begins with observed data and incomplete data. Multiple imputation cre-

ates $m>1$ complete versions of data, replacing missing values for plausible data values. These plausible values are extracted from a distribution modeled specifically for each missing data input. The sets of imputed data are identical to the observed data entries, but differ in the imputed values. The magnitude of these differences reflects our uncertainty about the value to be attributed.

II. The second step is to estimate the parameters of interest for each set of imputed data, by applying pattern analytical methods for complete data. The results will be different because their input data are different, which is only a result of uncertainty about the value to be attributed.

III. In the last step the m results are grouped into a final point estimate plus the standard deviation, by rules of simple grouping, known as "Rubin Rules" (Rubin, 1987).

According to the "Rubin rules" grouping of estimates is given by, $\bar{Q}=\frac{1}{m}\sum_{i=1}^{m}Q_i$, in which Q_i is the estimate of i-th parameter considered, corresponding to the set (m) of imputed data.

The combination of the standard errors involves two sample sources of variation, one being the variance inside the imputations, defined as the arithmetic mean of m sample variances described as $\bar{U}=\frac{1}{m}\sum_{i=1}^{m}U_i$, on what U_i is the variance of the m-th set of imputed data. Another source of variation is the variance between imputations, which quantifies the variability of an estimate in all the m imputations, and is given by $B=\frac{1}{m-1}\sum_{i=1}^{m}(Q_i-\bar{Q})$. With the calculation of the combined estimates $(\bar{Q},\bar{U})$ and B, it is possible to get the total combined variance described by $T=\bar{U}+(1+\frac{1}{m})B$, being $(1+\frac{1}{m})$ the correction of infinite numbers of imputations.

A measurement to determine the relative increase in variance due to the missing units is also presented by Rubin (1987) $r=\frac{(1+m)^{-1}B}{\bar{U}}$ and **yet** a fraction of missing units approaching of $\lambda=\frac{r}{(1+r)}$ or $\lambda=\frac{|(r+2)/(v+3)|}{(1+r)}$.

Knowing a high fraction of lacking data, the number of data sets to be imputed should be given more attention.

The MICE algorithm

The MICE algorithm "Multivariate imputation by Chained Equations" (Buuren & Oudshoorn, 2000), (Buuren & Groothuis-Oudshoorn, 2011), begins with a draw from the observed data, and impute the incomplete data, variable by variable. An iteration is a loop passing around $\boldsymbol{Y}_j$. The number of iterations T is often below, 5 to 10. The MICE algorithm generates multiple imputations running process m times parallel.

Implemented in *R* software, by mice package, so that the user can specify an imputation method for each column of incomplete data. The method of imputation takes a full set of predictors, and returns a single imputation for each missing entry in the target incomplete column. Several imputations are created by repeated calls to the function (Buuren & Groothuis-Oudshoorn, 2011). One advantage of the variable by variable approach is that for each variable, a different imputation model can be used. Therefore, a data set may have both continuous variables such as categorical (Oudshoorn, Buuren & Rijckevorsel, 1999).

Sampling

The study sought to identify low birth weight risk factors at birth, using records available in the Parana State Live Birth Information System (SINASC- PR) in the year 2012. To manipulate the records it was necessary to unzip the files through the program TABWIN. Records of 153945 live births were obtained, these pulled out a random sample of 3380 complete records.

The sample size was calculated by means of a pilot sample of n = 3000, used to get an approximation of the low weight ratios at birth (LBW) and normal weight at birth (NBW), found the proportion of 10% for LBW, it became established a maximum tolerated error of 1% and 95% confidence level, obtaining n = 3381.

From the sample were created, by simulation, three banks of incomplete data, in which were excluded at random by the *R* software, 5%, 10% and 20% of the observations of the variable "Birth weight (BW)", for being the variable of interest. As the loss of data was by means of random sampling, the probability of *BW* variable data being missing is the same to all individuals, and independing of data, which characterizes the random non-response mechanism .

To compare the different adjustments, with missing and imputed data, a golden pattern model as the logistic model adjusted with the original sample was adopted, formed only by complete records. The variables included as predictors of the response variable, birth weight, are in accordance with some studies done in other brazilian regions (Zhuofan, 2011),

(Cascaes et al., 2008), (Giglio et al.,2005).

Model

The outcome was treated as binary, 1 for Low birth weight (<2500g) and 0 for Normal weight (≥ 2500 g), parameter adopted based on the definition of low birth weight of the World Health Organization.

The logistic regression model (Hilbe, 2009) was used, as it is appropriate to describe the relationship between a dichotomic variable (Y) and a set of predictor variables $\boldsymbol{X}=(X_1,X_2,...,X_p)'$.

We can present the response variable in the following way:

$$Y=\begin{cases}0 \text{ if BW is greater than or equal to 2500g.}\\ 1 \text{ if BW is lower than 2500g.}\end{cases}$$

Thus, Y following a *Bernoulli* distribution with parameter $E[Y]=\pi$. The logistic model in its usual form is given by

$$Y_i = E[Y_i]+\varepsilon_i$$

where

$$E[Y_i]=\pi_i=\frac{\exp(\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p)}{1+\exp(\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p)}.$$

An essential concept in the study of the logistic regression models is the *Odds Ratio*, defined as the ratio of the chance of an event occurring in one group and the chance of occurring in another group. Considering the vector of p variables $\boldsymbol{X}$, and the vector of unknown parameters $\boldsymbol{\beta}=(\beta_0,\beta_1,\beta_2,...,\beta_p)'$, the *OR* associated with the i-th variable, adjusted for other variables present in the vector $\boldsymbol{X}$, is estimated by e^{β_i}.

Of the variables, some of which were categorized according to the framework provided by SINASC, Chart 1.

Chart 1- Description of the variables used in the logistic regression model.

Variable	Category	Description
Weight	1	Low Weight at Birth (LWB): the newborn weighed less than 2500 g.
	0	Normal Weight at Birth (NWB): the newborn weighed 2500 g or more.

Age	(10 -15)	Mother at age between 10 and 15 years old
	(16 - 20)	Mother at age between 16 and 20 years old
	(21-29)	Mother at age between 21 and 29 years old
	(30 – 39)	Mother at age between 30 and 39 years old
	40 (+)	Mother with more than 40 years of age
Marital Status (M. st.)	Single	Mother with Marital Status of single
	Separated	Mother with Marital Status of separated
	Married	Mother with Marital Status of married
	Common-law marriage	Mother with Marital Status of common-law marriage
Primipara/More Children (Prim More Children)	Primipara	Primipara (1st child): mother who has no other living children and also no dead children,
	Not Primipara	Mother who already has living or dead children.
Pregnancy (Preg)	Single	Single type of pregnancy
	Double	Double type of pregnancy
	Triple (+)	Triple or more type of pregnancy
Gestation (Gest)	< 31	Less than 31 weeks of gestation
	32 a 36	From 32 to 36 weeks of gestation
	36 (+)	36 weeks of gestation or more
Education (Edu)	None	No concluded year of study
	1 to 3 years	From 1 to 3 years of completed studies
	4 to 7 years	From 4 to 7 years of completed studies
	8 to 11 years	From 8 to 11 years of completed studies
	12 (+)	12 or more years of completed studies
Consultations (Cons)	None	No prenatal consultations
	(01 - 03)	From 1 to 3 prenatal consultations
	(04 – 06)	From 4 to 6 prenatal consultations
	07 (+)	7 or more prenatal consultations

In Chart 2, the categories taken as reference (*baseline)* are indicated.

Chart 2 – Reference categories *(baseline).*

Variables	**Reference category**
Age	(20-29)
Marital State (M. St.)	Married
Education (Edu)	12 (+)
Gestation (Gest)	36 (+)
Prenatal Consultations (P. Cons.)	7 (+)
Pregnancy (Preg)	Single
Primipara/More Children (Prim More Children)	1st child

Exploratory analysis and univariate tests for the outcome variable were held. Posteriorly, the logistic regression model was set by maximum likelihood method, as well as waste analysis, the application *R*.

The final model obtained, for the outcome variable Weight at birth, included the following predictive variables: mother's age, Mother's marital state, Mother's education, Type of pregnancy (number of babies), Length/number of weeks at birth-Gestation, Number of Prenatal consultations and the mother being Primipara (first child).

The multiple imputation was held through the *Multivariate Imputation package by Chained Equations* (MICE) from *software R*. The chosen imputation method was *polyreg*, for being suitable for binary variables with two levels (Little & Rubin, 1987).

After the completion of multiple imputation, the collected data were analyzed by logistic regression, by the *gml* function available in the *start* package that comes with the basic installation of *R*. The *gml* function uses the maximum likelihood and Fisher Score methods for estimation of model parameters. Estimates of all the analysis of *m* complete data sets were combined into a single set of results according to the "Rubin grouping rules".

RESULTS AND DISCUSSIONS

The obtained results of the logistic model for the outcome variable "birth weight", with the sample of complete data for the simulation of 5% missing data and 5% imputation are exempleficados in Table 1, Table 2 to 10% and Table 3 to 20%.

Table 1- Estimates, standard errors and the p-value of the logistic model to complete, incomplete data with 5% missing data and 5% imputation.

Predictor Variables	**Estimates, standard error and the p-value of the logistic models adjusted**		
	Model Gold Pattern	Model with 5% missing data	Model with 5% imputation
Age (16-19)	-0,44392 0,21615 0,039996	-0,36208 0,22354 0,105282	-0,4804 0,2147 0,025249
Marital Status - Common-law marriage	0,16210 0,21681 0,454663	0,25188 0,22569 0,264409	0,1670 0,2181 0,443660
Education-None	-2,24377 0.67575 0,000899	-2,04573 0.75484 0,006725	-2,1720 0,6767 0,001330

Table 2- Estimates, standard errors and the p-value of the logistic model to complete, incomplete data with 10% missing data and 10% imputation.

Predictor Variables	**Estimates, standard error and the p-value of the logistic models adjusted**		
	Model Gold Pattern	Model with 10% missing data	Model with 10% imputation
Age 40 (+)	-0,61735 0,45328 0,173206	-0,527903 0,492033 0,283315	-0,67137 0,45299 0,138319
Education-None	-2,24377 0,67575 0,000899	-2,151827 0,757700 0,004512	-2,38725 0,67544 0,000409
Education (1-3)	0,03738 0,47565 0,937364	0,016299 0,543683 0,976084	0,05136 0,47822 0,914469
Primipara	0,49302 0,17413 0,004634	0,520596 0,185341 0,004972	0,55316 0,17564 0,001636

Table 3- Estimates, standard errors and the p-value of the logistic model to complete, incomplete data with 20% missing data and 20% imputation.

Predictor Variables	**Estimates, standard error and the p-value of the logistic models adjusted**		
	Model Gold Pattern	Model with 20%	Model with 20%

		missing data	imputation
Age (16-19)	-0,44392 0,21615 0,039996	-0,3582886 0,2452150 0,14398	-0,41401 0,21567 0,054898
Gestation <32	-4,02265 0,38468 <2e-16	-4,3316965 0,4612559 <2e-16	-4,08781 0,38799 <2e-16
Education (4 to 7)	-0,03195 0,25682 0,900998	0,0689991 0,2898631 0,81185	0,32590 0,25379 0,199092

The logistic model was adjusted in seven different situations, thus resulting in 133 estimated coefficients (excluding the seven intercepts). When comparing the settings of data missing and data with imputations to the original data, it is observed that in general, the estimated values and their standard errors are more similar to the reference model when imputation is held. In all three situations of data lack simulation, the logistic models using the imputation data showed, in most part, lower standard error, when compared with the reference estimates (full sample).

The predictor variable Age (16-19) in the first scenario, for example, was significant at 5% in the standard model and ceases to be in the model where there is presence of missing data, but again becomes significant when adjusted in the model with imputation. Still in the first scenario, we highlight the variable "Marital Status - Common-law marriage", to which both the estimate as the standard error and p-value of the model with imputation are very similar to the standard model, showing an improvement in relation to the model with missing data.

If thought in terms of the *Odds Ratio* for the variable "Marital Status - Common-law marriage" in the standard model, model with missing and model imputation, the estimated *OR* are 1,1759; 1,2864 and 1,1817, respectively. Because it is a measure widely used in health care, in studies that aim to identify risk factors for agents or pathogens that affect the health of the population, stands out the fact that, even with a small percentage of missing values (5%), in a considerably big sample, improvements with the application of the imputation method are noticeable.

For more examples of the effectiveness of used imputation method are cited the variables: Age 40 (+), Education-None, Education (1-3), Primipara, in the 10% scenario of missing and Age (16-19), Gestation <32, in the 20% scenario. The results corroborate with those pre-

sented by Nunes et al. (2009) and Camargos et al. (2011).

In some situations, such as the variable "Education (4 to 7)", in the scenario of 20% missing data, there is not a good fit with the imputation. As the variables were categorized, some were with a very low number of cases, which may explain the imprecision of estimates and errors standard high.

FINAL CONSIDERATIONS

Restricting the analysis to cases that have complete observations may cause important predictors fail to be identified. Furthermore, it can result in smaller sample sizes than planned and still generate less predictive models than the case with complete data, with standard errors greater in the estimators.

The multiple imputation method is now available in many conventional statistical software. Thus, it is recommended that when analyzing their data, researchers do not ignore the problem of missing data. Imputing missing data can increase considerably the reliability of the results. Furthermore, strategies to deal with missing data can increase the effective size of the data set, making the analysis more effective. Moreover, they are easy to implement.

REFERENCES

ACOCK, A. C. (2005). Working with missing values. *Journal of Marriage and Family*, Wiley Online Library, v. 67, n. 4, p. 1012-1028.

BEAUJEAN, A. A., & BEAUJEAN, M. A. A. (2012). *Package `bayloredpsych'*.

BERGAMO, G. C. (2007). *Imputação múltipla livre de distribuição utilizando a decomposição por valor singular em matriz de interação.* Doctoral thesis | Escola Superior de Agricultura \Luiz de Queiroz.

BUHI, E. R., GOODSON, P., & NEILANDS, T. B. (2008). Out of sight, not out of mind: strategies for handling missing data. *American journal of health behavior*, PNG Publications, v. 32, n. 1, p. 83-92.

BUUREN, S. V., & OUDSHOORN, C. (2000). Multivariate imputation by chained equations. MICE V1. 0 user's manual. *Leiden: TNO Preventie en Gezondheid*.

BUUREN, S. V., & GROOTHUIS-OUDSHOORN, K. (2011). Mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, American Statistical Association, v. 45, n. 3.

BUUREN, S. V. (2012). *Flexible imputation of missing data*. [S.l.]: CRC press.

CAMARGOS, V. P., CÉSAR, C. C., CAIAFFA, W. T., XAVIER, C. C., & PROIETTI, F. A. (2011). Imputação múltipla e análise de casos completos em modelos de regressão logística: uma avaliação prática do impacto das perdas em covariáveis. *Cad Saúde Pública*, 27: 2299-313.

CASCAES, A. M., GAUCHE, H., BARABARCHI, F. M., BORGES, C. M., & PERES, K. G. (2008). Prematuridade e fatores associados no Estado de Santa Catarina, Brasil, no ano de 2005: análise dos dados do Sistema de Informações sobre Nascidos Vivos. *Cadernos de Saúde Pública*, Rio de Janeiro, 24(5):1024-103.

COLLINS, L. M., SCHAFER, J. L., & KAM, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, American Psychological Association, v. 6, n. 4, p. 330.

ENDERS, C. K. (2010). Applied missing data analysis. [S.l.]: *Guilford Publications*.

GIGLIO, M. R. P., LAMOUNIER, J.A., MORAIS-NETO, O. L., & CÁSAR, C. C. (2005). Baixo peso ao nascer em coorte de recém-nascidos em Goiânia-Brasil no ano de 2000. *Revista Brasileira de Ginecologia e Obstetrícia*. 27(3): 130-136.

HILBE, J. M. (2009). *Logistic regression models*. [S.l.]: CRC Press.

HORTON, N. J., & KLEINMAN, K. P. (2007). Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, v. 61, n. 1.

342 JAMSHIDIAN, M., & MATA, M. (2008). Postmodeling sensitivity analysis to detect the
343 e_ect of missing data mechanisms. *Multivariate Behavioral Research*, Taylor & Francis, v.
344 43, n. 3, p. 432-452.
345
346 JAMSHIDIAN, M., & JALAL, S. (2010). Tests of homoscedasticity, normality, and missing
347 completely at random for incomplete multivariate data. *Psychometrika*, Springer, v. 75, n. 4,
348 p. 649-674.
349
350 JAMSHIDIAN, M., JALAL, S. J., & JANSEN, C. (2014). Missmech: an r package for testing
351 homoscedasticity, multivariate normality, and missing completely at random (mcar).
352 Journal of Statistical Software, *Journal of statistical software*, v. 56, n. 6.
353
354 LITTLE, R. J., & RUBIN, D. B. (1987). *Statistical analysis with missing data*. [S.l.]: John
355 Wiley & Sons.
356
357 LITTLE, R. J. (1988). A test of missing completely at random for multivariate data with miss-
358 ing values. *Journal of the American Statistical Association*, Taylor & Francis, v. 83, n. 404,
359 p.1198-1202.
360
361 LITTLE, R. J. (1992). Regression with missing x's: a review. *Journal of the American Statis-*
362 *tical Association*, Taylor & Francis Group, v. 87, n. 420, p. 1227-1237.
363
364 MYERS, W. R. (2000). Handling missing data in clinical trials: an overview. *Drug Infor-*
365 *mation Journal*, SAGE Publications, v. 34, n. 2, p. 525-533.
366
367 NUNES, L. N., KLÜCK, M. M., & FACHEL, J. M. G. (2009). Uso da imputação múltipla de
368 dados faltantes: uma simulação utilizando dados epidemiológicos Multiple imputations for
369 missing data: a simulation with epidemiological data.*Cad. Saúde Pública*, 25.2:268-278.
370
371 OUDSHOORN, K., BUUREN, S. V., & RIJCKEVORSEL, J. V. (1999). Flexible multiple
372 imputation by chained equations of the AVO95 Survey. *Leiden: TNO Prevention and Health*.
373 [S.l.].
374

RAGHUNATHAN, T. E. (2004). What do we do with missing data? some options for analysis of incomplete data. *Annu. Rev. Public Health*, Annual Reviews, v. 25, p. 99-117.

ROTH, P. L., SWITZER, F. S., & SWITZER, D. M. (1999). Missing data in multiple item scales: A Monte Carlo analysis of missing data techniques. *Organizational Research Methods*, Sage Publications, v. 2, n. 3.

RUBIN, D. B. (1976). Inference and missing data. Biometrika, *Biometrika Trust*, v. 63, n. 3, p. 581-592.

RUBIN, D. B. (1978). Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In: *American Statistical Association*. Proceedings of the survey research methods section of the American Statistical Association. [S.l.], v. 1, p. 20-34.

RUBIN, D. B. (1987). *Multiple imputation for nonresponse in surveys*. [S.l.]: John Wiley & Sons.

RUBIN, D. B., & LITTLE, R. J. (2014). *Statistical analysis with missing data*. Hoboken, NJ: J Wiley & Sons.

SCHEFFER, J. (2002). *Dealing with missing data*. Massey University.

SCHAFER, J. L., & GRAHAM, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, American Psychological Association, v. 7, n. 2, p. 147.

TSIKRIKTSIS, N. (2005). A review of techniques for treating missing data in om survey research. *Journal of Operations Management*, Elsevier, v. 24, n. 1, p. 53-62.

ZHUOFAN, W. (2011). *Proposta de um modelo de regressão binária com resposta contínua aplicado à análise dos dados do SINASC: identificação de fatores de risco para o baixo peso ao nascer*. 2011. 76 f. Masters dissertation (Community Health) – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto.