UNIVERSIDADE FEDERAL DE JUIZ DE FORA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

Isabela Queirós Castro

Uma Aplicação de Métodos de Imputação no Estudo de Fatores Associados ao Baixo Peso ao Nascer

Juiz de Fora
2014

Isabela Queirós Castro

Uma Aplicação de Métodos de Imputação no Estudo de Fatores Associados ao Baixo Peso ao Nascer

Trabalho de Conclusão de Curso apresentado ao curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para obtenção do diploma de bacharel em Estatística.

Orientador: Professor Marcel de Toledo Vieira

Co-orientador: Professor Luiz Cláudio Ribeiro

Juiz de Fora

2014

Ficha catalográfica elaborada através do Programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Castro, Isabela Queirós.
Uma Aplicação de Métodos de Imputação no Estudo de Fatores Associados ao Baixo Peso ao Nascer / Isabela Queirós Castro. -- 2014.
78 p. : il.

Orientador: Marcel de Toledo Vieira
Coorientador: Luiz Cláudio Ribeiro
Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas, 2014.

1. Imputação de dados. 2. Regressão Logística Binária. 3. Baixo peso ao nascer. I. Vieira, Marcel de Toledo, orient. II. Ribeiro, Luiz Cláudio, coorient. III. Título.

Isabela Queirós Castro

Uma Aplicação de Métodos de Imputação no Estudo de Fatores Associados ao Baixo Peso ao Nascer

Trabalho de Conclusão de Curso apresentado ao curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para obtenção do diploma de bacharel em Estatística.

Aprovada em 16 de julho de 2014.

BANCA EXAMINADORA

_____.

Ph. D. em Estatística. Marcel de Toledo Vieira

Universidade Federal de Juiz de Fora

_____.

Doutor em Engenharia Elétrica. Alfredo Chaoubah

Universidade Federal de Juiz de Fora

_____.

Mestre em Estudos Populacionais e Pesquisas Sociais. Augusto Carvalho Souza

Universidade Federal de Juiz de Fora

AGRADECIMENTOS

Primeiramente, gostaria de agradecer a minha família pelo suporte nessa longa jornada que foi o curso de Estatística. Ao meu irmão Guilherme, que no seu papel de irmão mais velho me deu conselhos valiosos. E a minha querida mãe que vivenciou essa experiência junto comigo, nos momentos bons e ruins, sempre me incentivando e nunca me deixando desistir, aguentando os estresses e frustações e compartilhando as alegrias e conquistas.

Agradeço também aos companheiros de curso, em especial aos amigos Marcos, Cadu, Jarbas, Giovanna, Lidiana, Carol e Jéssica. Compartilhamos as dificuldades, mas nos apoiamos e estudamos juntos para conseguir essa vitória. Mais especial ainda ao Jarbas, que passou a ser mais que um amigo. Seu carinho e seu amor comigo foram essenciais e me deram força para conseguir realizar esse sonho.

Aos professores do curso de Estatística por seus ensinamentos e conhecimentos compartilhados, em especial ao professor e orientador Marcel, pela dedicação e orientação, sempre aberto aos alunos e disposto a nos ajudar. Em especial também ao professor e orientador Luiz Cláudio, que me acompanhou desde o início, me apoiando e incentivando e proporcionando minha primeira experiência profissional.

E por fim o mais importante. Agradeço ao meu pai, razão da minha escolha pela Estatística. Sempre me apoiou e incentivou nos estudos, muitas vezes até mais empolgado que eu com a faculdade. Foi curto o tempo que curtimos esse sonho juntos, mas sei que, onde estiver, ele está orgulhoso por minha conquista.

Agradeço a todos que de alguma forma contribuíram com a realização desse sonho.

RESUMO

A ocorrência de não resposta em grandes bases de dados é um problema comum na grande maioria dos estudos, transversais ou longitudinais. Este fenômeno tem como consequência a presença de dados faltantes que, dependendo da sua proporção, pode causar impactos na estimação dos parâmetros de interesse do estudo e levar a conclusões erradas. Existem técnicas apropriadas para tratar esses dados faltantes, como por exemplo, a imputação. Proposta por Rubin (1987), a imputação consiste na substituição de dados faltantes por valores plausíveis, preenchendo a base de dados e possibilitando que o pesquisador utilize os métodos estatísticos tradicionais para dados completos. Nosso objetivo é estudar diferentes métodos de imputação e sobretudo o efeito de ignorar a existência de dados faltantes na análise de uma base de dados reais, provenientes da Pesquisa Nacional de Demografia e Saúde (PNDS) de 2006. Além disso, um objetivo secundário é avaliar fatores associados ao baixo peso ao nascer, a partir dessa base de dados. O baixo peso é um importante indicador de saúde, além de um fator de risco para a morbi-mortalidade infantil. A Organização Mundial de Saúde define como baixo peso ao nascer, pesos inferiores à 2,500 Kg. Foram propostos modelos de Regressão Logística, considerando variáveis maternas, gestacionais e do recém-nascido como preditoras e o modelos das bases imputadas e incompletas foram comparados. Nossos resultados confirmam a importância da consideração adequada do problema da não resposta. Esperamos que o estudo desperte nos pesquisadores a atenção para a necessidade de realizar um tratamento adequado aos dados faltantes, antes das fases de análise e conclusão.

Palavras-chave: Imputação de dados, Regressão Logística Binária, Baixo peso ao nascer.

ABSTRACT

Non-response in large databases is a common problem in most, cross-sectional and longitudinal studies. This phenomenon results in the presence of missing data which, depending on their proportion and nature, may affect the estimation of the parameters of interest in the study and lead to wrong conclusions. There are proper techniques to handle missing data, such as imputation. Proposed by Rubin (1987), imputation is a methodology based upon the substitution of plausible values for missing data, populating the database and allowing the researcher to use the traditional statistical methods for complete data. Our aim is to study different methods of imputation and to evaluate the effect of ignoring the existence of missing data in the analysis of a database of real data from the Brazilian 2006 National Demographic and Health Survey (PNDS). Moreover, a secondary objective is to evaluate factors associated with low birth weight, from this database. Low birth weight is an important indicator of health, and a risk factor for infant morbidity and mortality. The World Health Organization defines as low birth weight, less than the 2,500 kg weights. Logistic Regression models were proposed, considering maternal, gestational and newborn variables as predictors and models to both imputed and incomplete bases were compared. Our results confirm the importance of proper consideration of the problem of non-response. We hope that the study will assist researchers to be more aware of the need for adequate treatment to missing data before the phases of analysis and conclusion.

Keywords: Imputation of data, Binary Logistic Regression, Low birthweight.

Sumário

Capítulo 1

Introdução

A ocorrência de não resposta é um problema recorrente em procedimentos de coleta de dados. Este fenômeno tem como consequência a presença de dados faltantes (*missing data*) tanto em estudos transversais quanto longitudinais.

Existem dois tipos de não resposta: não resposta a um item ou não resposta de uma unidade amostral. A primeira ocorre quando apenas uma ou algumas questões de um questionário não são respondidas. Já a segunda, ocorre quando um questionário inteiro não é respondido por uma pessoa que foi selecionada para compor a amostra.

A prevalência de dados faltantes é um grande desafio em estudos epidemiológicos, onde muitas vezes o objetivo é estudar fatores que contribuem para a ausência ou presença de determinada doença. A perda dessas informações pode causar problemas na análise dos dados (Nunes, Klück, & Fachel, 2009). Sendo assim, é importante estabelecer estratégias para a prevenção da ocorrência de dados faltantes, na fase de planejamento da pesquisa, ou posterior tratamento do problema durante as análises através da adoção de técnicas estatísticas adequadas.

No quadro 1 enumeramos alguns exemplos de situações que podem estar relacionadas à ocorrência de não resposta, bem como os tipos de estudos em que ocorrem.

Quadro 1: Situações para ocorrência de não resposta

Situação	Tipo de Não Resposta	Tipo de Estudo
Perda (óbito) de paciente antes do término da pesquisa	Unidade	Longitudinal
Desistência de um indivíduo em participar da pesquisa	Unidade ou Item	Transversal e Longitudinal
Falta de cooperação	Unidade	Transversal e Longitudinal
Incapacidade em responder	Unidade	Transversal e Longitudinal
Perda de documento	Unidade	Transversal e Longitudinal
Questões mal formuladas ou mal compreendidas	Item	Transversal e Longitudinal
Recusa em prestar informações para questões sensíveis	Item	Transversal e Longitudinal
Erros de transcrição ou digitação	Unidade ou Item	Transversal e Longitudinal
Efeito do entrevistador	Item	Transversal e Longitudinal

Em muitas vezes, é comum que a análise estatística seja feita levando-se em consideração apenas os casos com respostas completas e excluindo aqueles onde houve alguma não resposta. Porém, ignorar o problema da não resposta não é a melhor solução, uma vez que as estimativas obtidas em tais análises podem estar viesadas e consequentemente levar a conclusões errôneas.

Os métodos de análises estatísticas e pacotes computacionais são na sua maioria voltados para tratar dados completos. Assim, uma pequena parcela de dados faltantes na amostra pode gerar viés e ineficiência nas estimativas, e por isso, não podemos desconsidera-los nas análises (Nunes, Klück, & Fachel, 2010).

Para contornar o problema de não-resposta, desde os anos 1970 (Rubin, 1976; e Little e Rubin, 1983) surgiram técnicas estatísticas que envolvem a imputação dos dados faltantes, ou seja, substitui-los por valores plausíveis. Uma vez imputados, podemos utilizar os métodos de análises para dados completos.

As primeiras técnicas de imputação desenvolvidas eram mais simples em que os dados faltantes são substituídos pela média, pela mediana, por interpolação ou regressão linear (Nunes, Klück, & Fachel, 2009). São as chamadas imputações únicas, onde o valor ausente é substituído apenas uma vez e o banco de dados fica completo para análise com as técnicas tradicionais. Embora sejam ainda muitos utilizados, devido à facilidade de aplicação, existem algumas desvantagens na imputação única como a subestimação da variabilidade da variável imputada e a dificuldade de levar em consideração a variabilidade que possa existir entre diferentes imputações (Nunes, Klück, & Fachel, 2010). Idealmente, seria recomendável considerar na análise a incerteza adicional associada ao processo de imputação, uma vez que o valor imputado não é o verdadeiro valor da variável.

Para contornar as dificuldades da imputação única, Donald Rubin desenvolveu técnicas de imputação múltipla na década de 80 (Rubin D. B., Multiple imputation for nonresponse in surveys, 1987). A imputação múltipla consiste em imputar as não respostas *m* vezes e realizar *m* análises para dados completos. Os resultados obtidos em cada uma das *m* bases completas são combinados de forma simples para a análise final.

O objetivo dessa monografia é o estudo de métodos de imputação de dados faltantes e avaliar os efeitos de ignorarmos a presença dos mesmos em análises de dados epidemiológicos. Com esta finalidade será conduzida uma aplicação a dados reais provenientes da Pesquisa Nacional de Demografia e Saúde – PNDS – 2006 (Brasil, Pesquisa Nacional de Demografia e Saúde da Criança e da Mulher, 2009) e os resultados obtidos em

análises com aplicação desses métodos serão comparados com os resultados obtidos quando não há imputação.

Mais especificamente, nesta monografia são estudados os métodos de imputação única: (*i*) imputação pela média; (*ii*) imputação pelo vizinho mais próximo; e (*iii*) imputação por regressão, e os métodos de imputação múltipla: (*i*) método de Regressão Linear Bayesiana; (*ii*) método da média preditiva; e (*iii*) método de Monte Carlo (MCMC), fazendo-se ressalva de que apenas os métodos de imputação múltipla estarão sendo aplicados.

Um objetivo secundário da monografia é estudar os fatores associados ao baixo peso ao nascer a partir dos dados da PNDS - 2006, que fornece diversas informações acerca da saúde da mulher e da criança. O baixo peso ao nascer (BPN) é um importante indicador de saúde, uma vez que está diretamente relacionado à morbi-mortalidade infantil, além de influenciar na probabilidade de sobrevivência ao período neonatal e pós-neonatal (Uchimura, Pelissari, & Uchimura, 2008). De acordo com o Ministério da Saúde, o BPN

> "é um preditor da sobrevivência infantil; quanto menor o peso ao nascer, maior a possibilidade de morte precoce." (DATASUS)

Ainda de acordo com o Ministério da Saúde,

> "proporções elevadas de nascidos vivos de baixo peso estão associadas, em geral, a baixos níveis de desenvolvimento socioeconômico e de assistência materno-infantil." (DATASUS)

A prevalência de BPN no mundo chega a 15,5% (Maia & Souza, 2010), sendo que no Brasil esse número vem aumentando, sendo de 7,75% em 1996 e chegando a 8,53% em 2011, havendo diferença entre as regiões (DATASUS).

Além de ser fator de risco para mortalidade infantil, crianças nascidas com baixo peso são mais propensas a desenvolverem doenças na vida adulta, como obesidade, síndrome metabólica, bem como doenças crônicas (diabetes e hipertensão) e coronarianas (Franciotti, Mayer, & Cancelier, 2010).

A Organização Mundial de Saúde estabelece como baixo peso ao nascer, pesos inferiores a 2,500 Kg. Neste trabalho os pesos serão classificados em: peso adequado (p>=2,500 Kg) e baixo peso ao nascer (entre 1,500 e 2,499 kg).

Serão consideradas na aplicação as seguintes variáveis: variáveis maternas (idade, paridade, histórico de aborto); variáveis referentes à gestação e ao parto (idade gestacional, tipo de parto, pré-natal adequado, tabagismo) e variáveis referentes ao recém-nascido (sexo e raça/cor). As possíveis associações serão verificadas com a aplicação do Teste Qui-quadrado e serão empregados Modelos de Regressão Logística Binomial.

No capítulo 2 são discutidos métodos de imputação, bem como algumas características referentes a não resposta que ajudam na escolha do melhor método para tratá-las. No capítulo 3 temos a aplicação das técnicas de imputação abordadas a uma base de dados e a um problema real: o baixo peso ao nascer e seus fatores associados. No capítulo 4 temos as comparações dos resultados obtidos na análise a partir dos diferentes métodos de imputação propostos. E por fim, no capítulo 5, temos as conclusões e as considerações finais acerca desta monografia.

Capítulo 2

Neste capítulo serão explorados alguns conceitos com o objetivo de facilitar a compreensão do fenômeno da não resposta. Veremos o que são dados faltantes, quais são os padrões e os mecanismos geradores de não respostas, como podemos evitar que elas ocorram, e caso ocorram, algumas das formas adequadas de tratá-las.

2.1 - Definição

Uma base de dados é dita completa quando todas as variáveis que a compõe se encontram preenchidas, ou seja, os seus dados foram devidamente coletados durante o trabalho de campo da pesquisa. Quando ocorre a não resposta em uma ou mais variáveis, para alguma unidade selecionada para a amostra, dizemos que a base de dados está incompleta, ou seja, que ela contém dados faltantes. Com frequência a ocorrência de dados faltantes dificulta o uso de métodos estatísticos tradicionais para análise dos dados e, quando os mesmos são adotados, existe o risco dos estimadores se tornarem tendenciosos por conta das diferenças existentes entre os valores da variável de interesse para os respondentes e para os não respondentes. Tal viés dificilmente é eliminado, pois na prática, não conhecemos o real motivo que gerou a não resposta (Rubin, 1987 *in* Nunes, 2007).

A fim de se realizar um tratamento adequado dos dados, é importante que o pesquisador defina algumas características dos dados faltantes, como os mecanismos que os geraram, se existe algum padrão entre eles e a proporção de dados faltantes na base (Veroneze, 2011).

2.2 - Tipos/Padrão de não respostas

Inicialmente é importante fazer uma distinção entre os conceitos de “padrão de não resposta” e de “mecanismos geradores de não resposta”, apesar de serem tratados da mesma maneira em alguns artigos. O primeiro se refere à configuração dos dados observados e não observados em uma base de dados, enquanto o segundo descreve uma possível relação entre as variáveis medidas (variável de interesse e variáveis auxiliares) e a probabilidade de ocorrência de não resposta (Enders, 2010).

Quando temos uma base de dados incompleta, é importante verificar como ocorrem as não respostas e onde estão localizados, ou seja, qual é o seu padrão. Para melhor

entendimento, apresentamos na Figura 1 quatro padrões de não resposta discutidos na literatura.

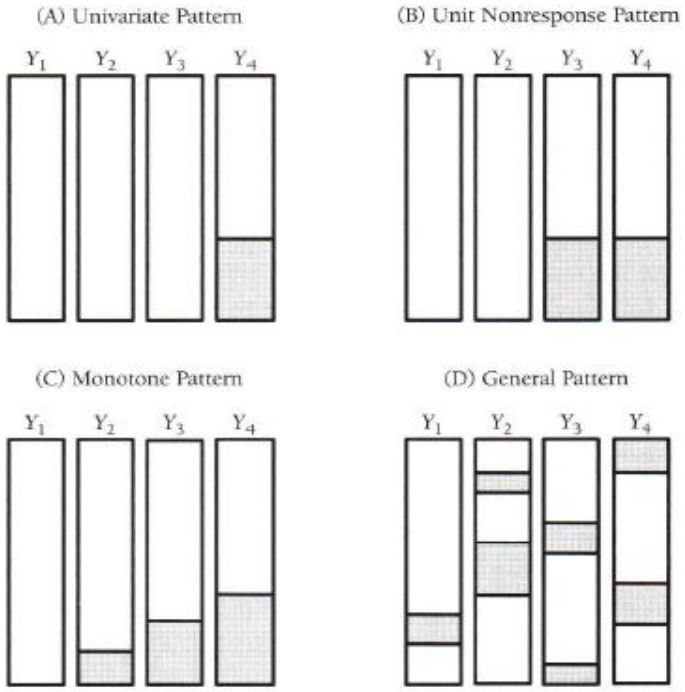


Figura 1: Padrões de não resposta

Abaixo, informações adicionais sobre cada um dos padrões de não resposta ilustrados acima são apresentadas (Enders, 2010).

A - Padrão Univariado (*Univariate Pattern*): Relativamente raro na maioria das aplicações práticas, podendo ocorrer em alguns tipos de estudos experimentais. Neste caso, os dados faltantes ocorrem apenas em uma variável isolada (Y_4 na Figura 1, por exemplo). Note que as variáveis Y_1 a Y_3 apresentam respostas completas.

B - Padrão Não Resposta Unidades (*Unit Nonresponse Pattern*): Comum em pesquisas de opinião, censos, pesquisas domiciliares, a não resposta ocorre quando algumas unidades da amostra selecionada (ou da população em casos de censos) se recusam a responder um ou mais itens do questionário. Note, nesta situação que Y_1 e Y_2 são variáveis disponíveis para todos os casos e Y_3 e Y_4 são variáveis incompletas, ou seja, uma ou mais não resposta.

C - Padrão de Não Resposta Monotônico (*Monotone Missing Data Pattern*): Comum em estudos longitudinais onde o participante desiste ou não é mais encontrado (fenômeno conhecido também como *atrição* ou abandono). Na Figura 1, este padrão é ilustrado de forma semelhante a uma escada, uma vez que a proporção de não respostas aumenta a cada momento do tempo (se considerarmos o contexto longitudinal).

D - Padrão de Não Resposta Geral (*General Missing Data Pattern*): Padrão mais comumente encontrado. Neste caso, os dados faltantes ficam totalmente dispersos na matriz de dados. É aparentemente aleatório, mas na maioria das situações as não respostas são sistemáticas e, por isso, são consideradas não ignoráveis.

Na seção seguinte serão discutidos os diferentes mecanismos geradores de não resposta.

2.3 - Mecanismos geradores de não resposta

Antes de realizar os possíveis tratamentos aos dados faltantes, é importante definir quais são os mecanismos os geraram, ou seja, como a probabilidade de um respondente não informar o valor da variável de interesse (ou seja, probabilidade de ocorrência de não resposta) está relacionada com o próprio valor da variável de interesse que foi omitido e com os valores de outras variáveis levantadas na pesquisa. O conhecimento do mecanismo gerador da não resposta permite a escolha correta da técnica para melhor tratá-los.

Conforme a classificação de Little e Rubin (1987), abordamos três mecanismos geradores de não resposta: Mecanismo Completamente Aleatório (*Missing Completely at Random*) - MCAR, Mecanismo Aleatório (*Missing at Random*) - MAR e Mecanismo Não Aleatório (*Missing Not at Random*) – MNAR. Enders (2010) foi utilizado como principal referência para a elaboração desta Seção.

2.3.1 - Mecanismo Completamente Aleatório (MCAR)

O mecanismo que gera a não resposta é completamente aleatório (MCAR) quando a probabilidade de não resposta em uma variável Y não é em função de outra variável X que foi medida nem do valor da variável Y para unidade onde ocorreu não resposta. Podemos dizer que os dados que foram observados em Y são uma amostra aleatória simples dos dados que seriam analisados caso não houvesse não resposta em Y.

2.3.2 - Mecanismo Aleatório (MAR)

Dizemos que o mecanismo que gerou a não resposta em uma variável Y onde ocorreu a não resposta é aleatório (MAR), quando a probabilidade de não resposta é função de uma variável X completa, ou seja, uma variável onde não ocorreram não respostas. Porém esta probabilidade não depende do valor da variável Y em si. O mecanismo MAR implica que existe uma relação sistemática entre uma ou mais variáveis mensuradas e a probabilidade de dados ausentes na variável Y.

Um problema prático do mecanismo MAR é que não temos como confirmar que a probabilidade de não resposta em Y é somente em função de outra variável medida. Como podem existir dificuldades para testarmos o mecanismo MAR, isto pode ser considerado um problema grave uma vez que estimadores de máxima verossimilhança e métodos de imputação múltipla trabalham com este pressuposto.

2.3.3 - Mecanismo Não Aleatório (MNAR)

Quando a probabilidade de não resposta em uma variável Y é definida em função dos valores da variável Y em si, mesmo quando controlada por outras variáveis medidas, então, dizemos que o mecanismo que gerou as não respostas em Y é não aleatório (MNAR).

2.3.4 – Distribuição de Dados Faltantes

Considerando a Teoria dos Dados Faltantes de Rubin (1976) podemos dividir uma base de dados A em dois componentes: os dados observados (A_{com}) e os dados faltantes (A_{falt}). Definimos R como a variável indicadora da observação ou não de uma variável Y, tal que:

$$R = \begin{cases} 1, & quando\ a\ variável\ \text{é}\ observada \\ 0, & quando\ ocorre\ \text{não}\ resposta \end{cases}$$

Note que *R* é uma variável aleatória que possui uma distribuição de probabilidade. A probabilidade de não resposta pode ou não estar relacionada com outras variáveis presentes na base de dados. A forma como a relação entre *R* e a base de dados ocorre é o que diferencia os mecanismos geradores de não resposta.

A definição dos mecanismos de não resposta envolve diferentes distribuições de probabilidade para a variável indicadora *R*, que descrevem diferentes relações entre *R* e a base de dados. A distribuição de probabilidade para o mecanismo MNAR, por exemplo, inclui

todas as possíveis relações entre os dados faltantes e a base de dados. Essa distribuição pode ser escrita na forma:

$$p(R|A_{com}, A_{falt}, \Phi) \quad (1)$$

Onde,

p(.) é uma distribuição de probabilidade qualquer;

Φ é um parâmetro que descreve a relação entre R e a base de dados.

Interpretando a equação (1) temos que a probabilidade de R assumir 0 ou 1 depende tanto de A_{com} quanto de A_{falt}, ou seja, a probabilidade de ocorrência de dados faltantes em Y pode depender de outras variáveis e também dos valores de Y.

Temos um mecanismo MAR quando a probabilidade de não resposta em uma variável Y está relacionada à outra variável X que foi medida e que está sendo considerada pelo analista no modelo que está sendo ajustado. Entretanto, tal probabilidade é independente dos valores faltantes da própria variável Y, ou seja, R depende de A_{com} mas não depende de A_{falt}. Neste caso, a distribuição de probabilidade pode ser escrita como:

$$p(R|A_{com}, \Phi) \quad (2)$$

A interpretação da equação (2) é que a probabilidade de dados faltantes depende dos dados observados, a partir de algum parâmetro que relaciona A_{com} e R.

E por último, o mecanismo MCAR, que diz que a probabilidade de não resposta é totalmente independente da base de dados, ou seja, R não depende nem dos valores observados A_{com} nem dos faltantes A_{falt}. Podemos escrever a distribuição como:

$$p(R|\Phi) \quad (3)$$

Interpretando a equação (3) temos que, a probabilidade de R assumir 0 ou 1 é independente da base de dados, dependendo apenas de um parâmetro que descreve a probabilidade de não resposta Φ.

Na figura 2, temos um resumo gráfico dos mecanismos geradores de não resposta, retirado de Schafer e Graham (2002). Temos quatro tipos de variáveis: X representa as variáveis completas, Y é uma variável com dados faltantes, Z são variáveis que não foram medidas, ou seja, não observadas, e podem estar relacionadas a probabilidade de não resposta e R é a variável indicadora dos dados faltantes.

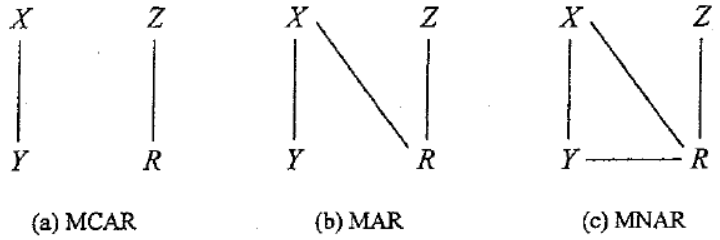


Figura 2: Mecanismos geradores de não resposta

Podemos adicionalmente classificar os mecanismos de não resposta como ignoráveis ou não ignoráveis (Allison, 2001). Os mecanismos MAR e MCAR são considerados mecanismos ignoráveis, pois não é necessário considerar no modelo o mecanismo que gerou a não resposta. Já o mecanismo MNAR é considerado não ignorável, pois o mecanismo que gerou a não resposta precisa ser considerado no modelo, para que sejam calculadas estimativas da variável de interesse.

Antes de descrever os métodos de imputação única, na próxima seção são discutidas brevemente algumas medidas de prevenção que devem ser tomadas na fase de planejamento e execução da pesquisa. Tais medidas podem contribuir para que o número de dados faltantes seja o menor possível.

2.4 - Métodos para prevenção de não resposta

É possível planejar uma pesquisa de forma que a ocorrência de não resposta seja menor. Quando planejamos um levantamento amostral é importante considerar os seguintes fatores, que ajudam a evitar a não resposta (Vieira, 2012):

- conteúdo da pesquisa: cuidado ao definir o conteúdo da pesquisa e em caso de assuntos sensíveis, como uso de drogas ou informações financeiras por exemplo, abordar tais itens com cautela para evitar alta taxa de não resposta;
- período adequado para a realização da pesquisa: evitar realizar pesquisas em época de férias ou feriados, ou mesmo durante grandes eventos como a copa do mundo por exemplo;
- treinamento apropriado dos entrevistadores: o treinamento adequado dos entrevistadores pode diminuir a taxa de não respostas em uma pesquisa. É importante que o pesquisador disponibilize um manual aos entrevistadores para auxiliá-los e preveni-los de possíveis situações que podem gerar não resposta;

- planejamento cuidadoso na fase de elaboração do questionário: evitar questionários longos e pouco objetivos que demandem muito tempo do respondente para não desestimular o entrevistado;
- não exagerar ao insistir com o respondente: se insistirmos muito com um respondente para obter uma informação, podemos levá-lo a se recusar de continuar participando da pesquisa. Além disso, devemos evitar abordar o mesmo participante para diferentes pesquisas em um curto período de tempo;
- fazer uma apresentação adequada da pesquisa antes de iniciá-la: é fundamental no início da pesquisa fazer uma apresentação formal, a partir de cartas ou mesmo um contato por telefone par informar ao respondente sobre a visita para a pesquisa. Além disso, o entrevistador deve enfatizar a importância da participação do respondente, bem como a utilidade e relevância da pesquisa e lembrá-los da confiabilidade de suas informações;
- fornecimento de incentivo aos participantes: ao fornecer incentivos como brindes e sorteios, por exemplo, estimulamos os respondentes a contribuírem positivamente com a pesquisa, evitando altas taxas de não resposta;
- adoção de procedimentos de reabordagem (*follow up*): em caso de uma primeira tentativa sem sucesso ao encontrar o respondente, é importante tentar novamente em horários alternativos, por exemplo.

Mesmo com todo o cuidado no planejamento da pesquisa e na aplicação dos questionários, é muito provável que ainda assim existam dados faltantes. Porém, existem métodos estatísticos adequados para tratá-los, conforme segue na próxima seção.

2.5 - Métodos para tratamento de não resposta

Uma vez investigados os mecanismos geradores de não resposta e o seu padrão, é possível utilizar métodos para o tratamento de dados faltantes em nossa base de dados. Iremos apresentar alguns métodos de imputação única como a imputação pela média, imputação pelo vizinho mais próximo (*Hot-deck*) e imputação por regressão, e também métodos de imputação múltipla como a regressão linear Bayesiana, método da média preditiva e o método de Monte Carlo (MCMC).

O termo imputação única se dá pelo fato de que cada valor faltante será substituído uma única vez, ao contrário da imputação múltipla que permite a geração de várias cópias da

base de dados com dados faltantes, e em cada uma dela, substitui os valores faltantes por um valor diferente.

A vantagem da imputação é o fato de que as bases de dados passam a ser completas e a imputação única é vantajosa, porque, ao contrário das abordagens da eliminação que descartam os casos que apresentam dados faltantes, ela faz uso dos mesmos. Além disso, em geral, a imputação única fornece estimativas pontuais consistentes (Baracho, 2003). Porém, como os dados são imputados (não sendo os mesmos valores reais observados), o erro padrão amostral dos estimadores tendem a aumentar, uma vez que adicionamos ruídos às estimativas dos parâmetros.

A principal desvantagem associada à Imputação Única é a subestimação dos erros padrão das estimativas, resultando em intervalos de confiança muito estreitos ou a valores muito altos para as estatísticas dos testes, uma vez que os dados faltantes são substituídos apenas uma vez, desconsiderando outros valores que poderiam ter sido imputados, ou seja, não considera a variabilidade entre imputações (Robert Groves *et al*., 2009). A subestimação do erro padrão pode acarretar em erros do Tipo I, ou seja, na obtenção de resultados estatisticamente significativos sem evidências suficientes (Veroneze, 2011).

2.5.1 – Imputação pela média

A ideia da imputação a partir da média é substituir os dados faltantes pela média dos valores que foram observados na variável de interesse. Esta técnica de imputação é bem antiga e os metodologistas a atribuem a Wilks (1932).

Quando tratamos de dados do tipo quantitativos, os dados faltantes são substituídos pela média da variável, podendo ser a média geral (média de todos os valores observados) ou a média de um grupo que seja semelhante ao caso onde houve não resposta, identificado por outras variáveis presentes no banco de dados.

Devemos lembrar que, apesar da imputação pela média não alterar a média amostral, ela altera outras características da distribuição.

2.5.2 – Imputação pelo vizinho mais próximo (*Hot-deck)*:

A imputação pelo vizinho mais próximo utiliza informações de variáveis auxiliares para substituir o dado faltante. Esse procedimento foi desenvolvido por estatísticos do *Census Bureau of Statistics* para lidar com dados faltantes em base de dados públicas, sendo muito utilizada em pesquisas no mundo todo (Enders, 2010).

A ideia básica desse método é imputar os dados faltantes com a informação de outros casos. É realizado um sorteio aleatório dos dados que foram observados na variável de interesse para substituir cada valor faltante.

O sorteio é condicionado à informação dada em uma variável auxiliar completa, ou seja, foi observada para todos os casos. Assim, temos uma sub-amostra formada pelos casos que são similares em relação a variável auxiliar.

2.5.3 – Imputação por regressão

Esse método substitui os valores faltantes por valores preditos a partir de um modelo de regressão que tem como covariáveis, outras variáveis completas presentes na base de dados. Assim como a imputação pela média, a imputação por regressão é um método antigo (Buck, 1960).

A ideia da imputação por regressão é usar a informações de variáveis completas para preencher os valores da variável de interesse onde houve não resposta. Como as variáveis são correlacionadas, faz sentindo tomar emprestadas as informações das variáveis que foram observadas.

Primeiro estimamos as equações de regressão que predizem as variáveis com dados faltantes, que podem ser geradas a partir de uma análise de dados completos. Depois, geramos os valores preditos para os dados faltantes, a partir das regressões.

As regressões utilizadas para imputação podem ser simples ou múltiplas, utilizando uma ou mais variáveis completas presentes na base de dados.

Podemos generalizar a equação de regressão, considerando uma ou mais variáveis com dados faltantes, da seguinte forma:

$$\widehat{y_i} = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j \quad (3)$$

Onde:

$\widehat{y_i}$ é a estimativa da variável com dados faltantes;

β_0 é o intercepto do modelo;

β_j é o j-ésimo coeficiente do modelo;

x_j é a j-ésima variável auxiliar;

j = 1,2,...,n.

2.5.4 – Imputação Múltipla (IM)

Nos últimos vinte anos os métodos para imputação de dados tem sido muito estudados, e surgiram alternativas metodológicas capazes de amenizar as desvantagens encontradas na imputação única. Tais métodos têm como base a Imputação Múltipla (IM), proposta por Rubin ainda nos anos 1970.

Os métodos de Imputação Única tendem a subestimar a variabilidade dos estimadores, uma vez que consideram os valores imputados como valores observados (Baracho, 2003). Esse problema pode ser corrigido a partir da IM, pois este método permite a consideração de medidas da incerteza associadas à imputação no cálculo da variância das estimativas produzidas (Rubin, 1987; *in* Nunes, 2007).

A IM foi proposta por Rubin em 1978 e vem sendo muito recorrente em estudos sobre imputação de dados, sendo mais utilizada recentemente, devido aos avanços computacionais. O quadro abaixo resume os principais *software*/aplicativos utilizados para imputação de dados, citados na literatura:

Quadro 2: Aplicativos mais usados para imputação de dados

Aplicativos	**Página na WEB**	**Imputação Única**	**Imputação Múltipla**
Grátis			
Amelia	http://gking.harvard.edu/amelia/		x
CAT	http://www.stat.psu.edu/~jls/misoftwa.html#aut	x	
EMCOV	http://methcenter.psu.edu/downloads/EMCOV.html	x	
NORM	http://www.stat.psu.edu/~jls/misoftwa.html#aut	x	x
MICE	Free with R, commercial with S-Plus http://www.multiple-imputation.com		
			x
MIXED	Free with R, commercial with S-Plus http://www.stat.psu.edu/~jls/misoftwa.html#aut		
		x	x
MX	http://www.vcu.edu/mx/	x	
PAN	Free with R, commercial with S-Plus http://www.stat.psu.edu/~jls/misoftwa.html#aut		
		x	x
Comercial			
EQS	http://www.mvsoft.com/	x	
HLM	http://ssicentra.com/hlm/index.html	x	x
Mplus	http://www.statmodel.com	x	x
SAS	http://www.sas.com	x	x
SOLAS	http://www.statsol.ie/solas/imputationtechniques.htm	x	x
S-Plus	http://www.stat.psu.edu/~jls/misoftwa.html#aut	x	x
SPSS	http://www.spss.com	x	
Stata	http://www.stata.com	x	

Fonte: Nunes (2007).

A IM consiste em imputar *m* valores (*m* > 1) para cada dado faltante. Assim, obtemos *m* bases de dados completas que são analisadas pelos procedimentos tradicionais, e os resultados, que podem ser diferentes, são combinados a partir de aritmética simples para obtermos uma estimativa pontual para o parâmetro de interesse a partir da média das imputações e seu erro padrão a partir da variância das imputações, que refletem a incerteza dos dados faltantes (Schafer & Graham, 2002).

Podemos citar três vantagens importantes atribuídas a IM (Rubin, 1987; *in* Nunes, 2007): (*i*) aumento na eficiência da estimação, uma vez que as imputações forem feitas de forma aleatória procurando representar a distribuição dos dados; (*ii*) quando são feitas *m* imputações tendo como base um mesmo modelo para não resposta, inferências válidas, que refletem a variabilidade adicional devida aos dados faltantes, são obtidas através da combinação simples de processos de estimação aplicados aos dados completos; e (*iii*) gerando imputações múltiplas sob diferentes modelos podemos estudar a sensibilidade das inferências em relação a vários modelos de não resposta. Além disso, o modelo utilizado para a imputação não necessariamente precisa ser o mesmo utilizado para a análise, fazendo com que a IM seja ainda mais atrativa.

Temos também três desafios relacionados à aplicação da IM: (*i*) é mais trabalhosa para obter os valores imputados; (*ii*) necessita mais espaço para armazenar as bases de dados imputadas; e (*iii*) demanda mais trabalho para analisar as bases de dados completas. No entanto, estes desafios vêm se tornando menos relevantes com o passar do tempo devido aos avanços da tecnologia da informação.

Determinamos *m* a partir da proporção de dados faltantes na base de dados. A eficiência de uma estimativa baseada em *m* imputações em relação a uma baseada em infinitas imputações é $(1 + {}^{\lambda}/_{m})^{-1}$, onde λ é a proporção de dados faltantes (Rubin, 1987; *in* Schafer & Graham, 2002 e Baracho, 2003).

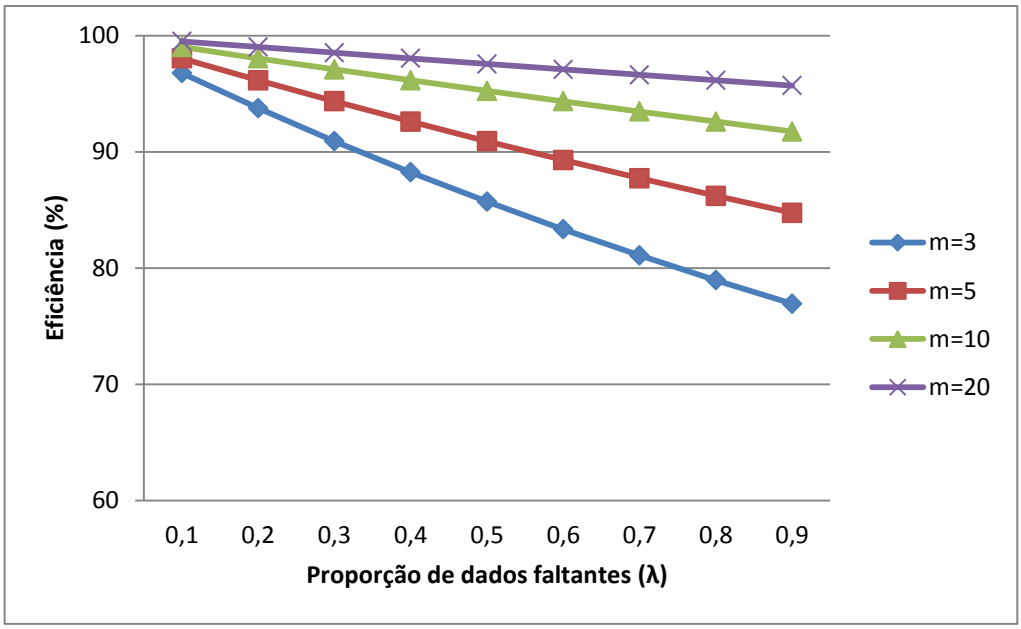
Por exemplo, se temos λ=0,5 (50% de dados faltantes) e *m*=10 imputações, temos uma eficiência de $(1 + {}^{0,5}/_{10})^{-1} = 0{,}95$. Sendo assim, imputações adicionais fariam pouca diferença em termos de aumentos da eficiência das estimativas em relação ao aumento do esforço computacional necessário (Schafer & Graham, 2002).

Na década de 1980, usavam-se valores pequenos para *m*, devido a dificuldades práticas em lidar com tantas bases de dados. A literatura indica *m* entre 3 e 10 imputações, porém, tornou-se usual *m* = 5, devido a experiências de pesquisadores, que verificaram que

um número pequeno de imputações é suficiente para que as conclusões sejam estatisticamente eficientes. No entanto, com a evolução computacional, hoje em dia, é possível realizar imputações com *m* maiores, sem que isso afete a análise ou demande muito tempo.

No Gráfico 1, temos a relação da eficiência da IM em função da proporção de dados faltantes λ, para melhor visualização.

Gráfico 1: Eficiência x Proporção de dados faltantes (λ)



Portanto, resumidamente a IM pode ser definida em três passos, conforme representada na Figura 3:

1. Imputação: são obtidos *m* > 1 banco de dados completos, a partir das *m* imputações;
2. Análise: cada uma das *m* bases de dados são analisadas, separadamente, pelos métodos tradicionais para bases de dados completas;
3. Combinação: os *m* resultados obtidos são combinados de forma simples, conforme as Regras de Rubin, descritas a seguir.

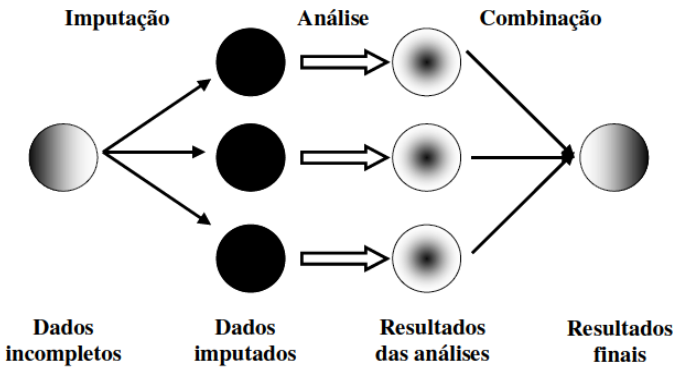


Figura 3: Representação da Imputação Múltipla

Fonte: www.multiple-imputation.com *in* (Nunes, Métodos de imputação de dados aplicados na área da saúde, 2007)

As Regras de Rubin referem-se à combinação dos resultados obtidos nas diferentes análises e podem ser usadas, independente do método de IM.

Para cada análise das *m* bases de dados completas, obtemos uma estimativa para um parâmetro escalar de interesse Q, ou seja, $Q_j,\ j = 1,2,\dots,m$. Sejam $\hat{Q}$ e $\sqrt{U}$ as estimativas de Q e o desvio padrão, respectivamente, caso não houvesse dados faltantes (Schafer & Graham, 2002).

A estimativa combinada será a média das estimativas individuais, conforme a equação abaixo:

$$\bar{Q} = \frac{1}{m}\sum_{j=1}^{m}\widehat{Q_j} \quad (4)$$

A variância combinada tem duas partes: variância intra-imputações (5) e variância entre-imputações (6).

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m} U_j \quad (5) \qquad B = \frac{1}{m-1}\sum_{j=1}^{m}(\widehat{Q_j} - \bar{Q})^2 \quad (6)$$

Portanto, a variância total será a variância combinada *T:*

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B \quad (7)$$

O primeiro passo da IM – Imputação – é o mais importante, pois as técnicas de imputação utilizadas devem considerar a relação entre os dados faltantes e os observados, mecanismos de não resposta e o padrão dos dados faltantes. Neste passo é estabelecido um algoritmo iterativo que abrange dois passos: passo de imputação (passo-I) e passo a *posteriori* (passo-P), segundo Enders (2010). O passo-I usa uma estimativa do vetor de médias e da matriz de covariância para construir uma equação de regressão que irá predizer as variáveis com dados faltantes baseando-se nas variáveis observadas e adiciona resíduos aleatórios aos escores preditos.

O modelo de regressão segue uma estrutura $Y_i \sim N(X_i\beta, \sigma^2)$, que especifica $f(Y_i|X_i, \theta)$, $\theta = (\beta, log\ \sigma)$, onde $\beta = (\beta_0, \beta_1, \dots, \beta_k)$ e σ é um escalar. Assume-se uma *priori* não informativa para θ e $n>k$, n o total de casos da base de dados.

Pelo resultado 5.3 de Rubin (1987; *in* Nunes, 2007):

- a *posteriori*, θ depende apenas do valores observados de Y;
- a *posteriori*, β, dado σ^2, segue uma distribuição normal multivariada $NMV(\hat{\beta}_*, \sigma^2 V)$, tal que;

$$\hat{\beta}_* = V\left[\sum_{i=1}^{n} X_i^t Y_i\right], V = (X_i^t X_i)^{-1} \quad (8)$$

- a *posteriori*, σ^2 é $\hat{\sigma}_*^2/(n-k)$ dividido por uma variável aleatória χ^2_{n-k}, tal que:

$$\hat{\sigma}_*^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_*)^2}{(n-k)} \quad (9)$$

Portanto, se tivermos uma distribuição a *posteriori* de θ para uma distribuição conhecida, podemos estimar os parâmetros utilizados na fase de imputação.

A equação da imputação é (Enders, 2010):

$$Y_i = (\hat{\beta}_0 + \hat{\beta}_k X_{ki}) + z_i \quad (10)$$

Onde:

Y_i é o valor da variável imputada para o caso *i;*

$(\hat{\beta}_0 + \hat{\beta}_k)$ são os coeficientes da regressão que geram o valor da variável *Y* para o indivíduo *i*;

X_{ki} é o valor da *k*-ésima variável completa para o caso *i*;

z_i é o resíduo aleatório de uma distribuição normal, com média 0 e variância igual a variância residual da regressão de Y_i.

Numa visão Bayesiana, criamos uma distribuição preditiva a *posteriori*, ou seja, a distribuição dos valores faltantes condicional aos valores observados. O objetivo da fase da Imputação é gerar *m* bases de dados, cada uma preenchida com estimativas únicas dos valores faltantes. Para isso, são necessárias diferentes estimativas dos coeficientes da regressão em cada passo-I e o propósito do passo-P é gerar estimativas alternativas para o vetor de médias e matriz de covariância (Enders, 2010).

O passo-P inicia usando os dados preenchidos a partir do passo-I anterior para estimar o vetor de médias e a matriz de covariância, e depois o algoritmo gera novos valores para os parâmetros adicionando um resíduo aleatório a cada elemento de $\widehat{\boldsymbol{\mu}}$ e $\widehat{\boldsymbol{\Sigma}}$.

Adicionar termos residuais aos elementos do vetor de médias e da matriz de covariância produz valores para os parâmetros que diferem aleatoriamente daqueles produzidos pelos coeficientes de regressão no passo-I predecessor. Atualizando as previsões no próximo passo-I, temos novos coeficientes de regressão e diferentes imputações, que são consideradas no próximo passo-P. Assim, o algoritmo gera outras estimativas de parâmetros plausíveis. Repetir este procedimento de duas etapas várias vezes cria várias cópias dos dados, cada uma com estimativas únicas dos valores faltantes (Enders, 2010).

Nas próximas seções são abordados os seguintes métodos de IM: Método de Regressão Linear Bayesiana, o Método da Média Preditiva e o Método MCMC (*Markov Chain Monte Carlo*).

2.5.4.1 – Método da Regressão Linear Bayesiana

Nessa seção veremos como os princípios da estimação Bayesiana se aplicam a fase da imputação. A análise Bayesiana abrange três passos (Enders, 2010):

a) Especificar uma distribuição a *priori* para o parâmetro de interesse;

A distribuição a *priori* descreve suposições sobre a probabilidade relativa de diferentes valores dos parâmetros, antes da coleta dos dados. A abordagem Bayesiana permite que o pesquisador incorpore seus conhecimentos à análise, utilizando-se da meta-análise para formular uma distribuição a *priori*. Especificar uma distribuição a *priori* requer três tipos de informação: (*i*) a localização ou média da distribuição; (*ii*) a dispersão ou desvio-padrão da distribuição; e (*iii*) o número de observações hipotéticas associadas à *priori.*

É usual a adoção de *priores* não informativas, o que representa a falta de conhecimento do pesquisador quanto aos parâmetros da população. Assim, a distribuição *posteriori* gerada é definida unicamente pelos dados.

b) Usar uma função de verossimilhança;

Coletados os dados, a função de verossimilhança é usada para resumir as evidências sobre os diferentes valores dos parâmetros. Substituindo os dados e algum valor para o parâmetro em uma função densidade de probabilidade, temos a probabilidade relativa, ou verossimilhança, dos dados, dado o valor do parâmetro específico. Quando repetimos este processo para diferentes valores dos parâmetros, obtemos a função de verossimilhança que descreve a probabilidade relativa dos dados a partir de diferentes valores de parâmetros.

c) Definir a distribuição a *posteriori* do parâmetro de interesse.

A distribuição a *posteriori* combina informações da distribuição a *priori* e a verossimilhança para gerar uma distribuição que descreve a probabilidade relativa de diferentes valores de parâmetros. A ideia básica é ponderar cada ponto da função de verossimilhança levando em consideração uma medida de magnitude das suposições da *priori*. Cabe ressaltar que o formato da distribuição a *posteriori* é um aspecto importante da análise Bayesiana.

Definição 2.1 **O Teorema de Bayes**

O Teorema de Bayes é o mecanismo matemático que da suporte à análise Bayesiana e é fundamental para definir o formato de uma distribuição a *posteriori*. Sua equação descreve uma relação entre duas probabilidades condicionais e é definido por:

$$p(B|A) = \frac{p(B)p(A|B)}{p(A)} \quad (11)$$

Onde:

A e B são dois eventos aleatórios;

$p(B|A)$ é a probabilidade condicional de observar o evento *B*, dado que o evento *A* ocorreu ;

$p(A|B)$ é a probabilidade condicional de *A* dado *B*;

$p(B)$ é a probabilidade do evento *B*;

$p(A)$ é a probabilidade marginal do evento *A*.

De acordo com Enders (2010), se substituirmos os eventos A e B pelos dados amostrais Y e pelo parâmetro de interesse θ, respectivamente, na equação (11) notamos a relação entre o teorema de Bayes e a Estatística, como segue na equação (12).

$$p(\theta|Y) = \frac{p(\theta)p(Y|\theta)}{p(Y)} \quad (12)$$

Onde:

$p(\theta)$ é a distribuição a *priori* do parâmetro de interesse;

$p(Y|\theta)$ é a verossimilhança, ou seja, a probabilidade condicional dos dados, dado algum valor assumido para θ;

$p(Y)$ é a distribuição marginal dos dados;

$p(Y|\theta)$ é a distribuição a *posteriori*, ou seja, a probabilidade condicional do parâmetro, dado os dados.

Em palavras, o Teorema de Bayes pode ser entendido por:

$$Posteriori = \frac{Priori \text{ x } Verossimilhança}{Fator\ escalar} \quad (13)$$

O fator escalar nada mais é do que uma constante que faz com que a área sob a distribuição a *posteriori* some um. A divisão por uma constante não altera o formato da mesma e, portanto, podemos simplificar a expressão (13) conforme segue abaixo.

$$Posteriori \propto Prioxi \text{ x } Verossimilhaça \quad (14)$$

A expressão (14) implica que a distribuição a *posteriori* é proporcional a distribuição a *priori* vezes a verossimilhança. Essa é a ideia básica da estimação Bayesiana.

□

Observação 2.1 **A estimação Bayesiana e a imputação múltipla**

A imputação múltipla gera *m* bases de dados incompletas e cada cópia é preenchida com diferentes estimativas para os dados faltantes. Como explicado mais acima, o processo de imputação se dá por um algoritmo iterativo de duas fases: o passo-I e o passo-P.

O passo-I usa equações de regressão para predizer as variáveis incompletas a partir das observadas e adiciona resíduos aleatórios aos escores preditos. O passo-P seleciona aleatoriamente um novo vetor de média e uma nova matriz de covariância, ou parâmetros simulados, de suas respectivas distribuições *posteriori*.

Podemos representar o passo-I, conforme a equação (15) (Enders, 2010).

$$Y^*_t \sim p\left(Y_{falt} \middle| Y_{obs}, \theta^*_{t-1}\right) \quad (15)$$

Onde:

Y_t^* é o valor imputado no t-ésimo passo-I;

Y_{falt} são os dados faltantes na base de dados;

Y_{obs} são os dados observados na base de dados;

θ^*_{t-1} denota o vetor de médias e a matriz de covariância do passo-P anterior, ou seja, os valores dos parâmetros que geraram a equação de regressão de imputação.

Interpretando a equação (15) temos que os valores imputados em um passo-I particular são uma realização de uma distribuição de possíveis valores substituídos que depende dos dados observados e dos parâmetros estimados atuais.

O passo-P usa os dados preenchidos do passo-I anterior para calcular as médias amostrais $\widehat{\boldsymbol{\mu}}$ e a soma dos quadrados e a matriz de produtos cruzados amostral $\widehat{\boldsymbol{\Lambda}}$, ou a matriz que define a dispersão da distribuição. A partir dessas medidas, a *posteriori* da matriz de covariância é $p(\boldsymbol{\Sigma}|\widehat{\boldsymbol{\mu}}, Y) \sim W^{-1}(N-1, \widehat{\boldsymbol{\Lambda}})$ (16), onde $\sim W^{-1}$ é a inversa da distribuição Wishart e *N-1* são os graus de liberdade.

Definido o formato da distribuição a *posteriori*, o algoritmo de *data augmentation* utiliza a simulação computacional de Monte Carlo para gerar a nova matriz de covariância da *posteriori*, ou covariância simulada $\boldsymbol{\Sigma}^*$. O algoritmo usa um processo semelhante para criar um novo conjunto de médias. A média amostral e a matriz de covariância simuladas definem a distribuição a *posteriori* do vetor de médias como $p(\widehat{\boldsymbol{\mu}}|Y, \boldsymbol{\Sigma}) \sim NMV(\widehat{\boldsymbol{\mu}}, N^{-1}\boldsymbol{\Sigma}^*)$ (17), onde $\widehat{\boldsymbol{\mu}}$ é o vetor de médias amostrais e $\boldsymbol{\Sigma}^*$ é a matriz de covariância simulada. Então, a simulação de Monte Carlo gera um novo conjunto de médias da distribuição da equação (16), denotada $\boldsymbol{\mu}^*$.

Depois de obtidos os novos parâmetros da distribuição a *posteriori*, o próximo passo-I usa os valores de parâmetros atualizados para construir novas equações de regressão que são diferentes daquelas no passo-I anterior, que gera novos valores para os dados faltantes. As novas imputações são utilizadas no próximo passo-P, onde o algoritmo produz novas estimativas plausíveis para os parâmetros. Repetindo o processo diversas vezes obtemos várias cópias dos dados, cada uma com estimativas únicas dos dados faltantes (Enders, 2010).

O passo-P pode ser representado pela seguinte equação (Enders, 2010):

$$\boldsymbol{\theta}^*_t \sim p(\boldsymbol{\theta}|Y_{obs}, Y^*_t) \quad (18)$$

Onde:

$\boldsymbol{\theta}^*{}_t$ são o valores dos parâmetros simulados a partir do passo-P ($\boldsymbol{\mu}^*$ e $\boldsymbol{\Sigma}^*$);

Y_{obs} são os dados observados;

$Y^*{}_t$ contém os valores imputados a partir do passo-I anterior.

A interpretação da equação (18) nos diz que os valores dos parâmetros simulados a partir do passo-P são realizações de uma distribuição que depende dos dados observados e dos valores substituídos no passo-I anterior. Simplificando, o passo-P consiste em utilizar os dados preenchidos para estimar o vetor de médias e a matriz de covariância e gerar novos valores plausíveis para os parâmetros, adicionando um resíduo aleatório para cada elemento em $\widehat{\boldsymbol{\mu}}$ e $\widehat{\boldsymbol{\Sigma}}$.

2.5.4.2 – Métodos da Média Preditiva

O método da média preditiva foi proposto por Little (1988 *in* Allisson, 2001) e pode ser considerado com um método do tipo *hot-deck*, uma vez que são utilizados na imputação apenas valores observados.

O método tem início com a regressão de *Y*, variável com dados faltantes que será imputada, em função das demais variáveis completas observadas. Essa regressão é usada para gerar valores preditos, tanto para as variáveis com dados faltantes, quanto para as completas. Então, para cada caso em que houve dado faltante, encontramos um conjunto de casos com dados completos, que tenha os valores preditos de *Y* próximos ao valor predito para o caso de dados em falta. Deste conjunto de casos, selecionamos aleatoriamente um caso que será o doador para o caso com dado faltante (Allisson, 2001).

É importante definir um ponto de corte de proximidade que delimite o número de possíveis doadores de informação. Se o grupo de doadores for pequeno, há maior variabilidade amostral nas estimativas. Por outro lado, um grande número de doadores pode levar a um possível viés, pois muitos doadores podem ser diferentes dos destinatários.

Também é importante estarmos atentos ao fato de que os coeficientes de regressão são apenas estimativas dos coeficientes verdadeiros. Tais estimativas se dão extraindo aleatoriamente da sua distribuição a *posteriori*, um novo conjunto de parâmetros de regressão, antes de se calcular os valores preditos para cada conjunto de dados imputados, conforme os passos abaixo (Allison, 2001):

- ajustar um modelo para *Y* tendo *X* como um vetor de co-variáveis para os n_1 casos sem dados faltantes em *Y*, produzindo coeficientes de regressão β (um vetor *k* x 1) e variância residual estimada s^2;
- considerar uma realização da distribuição a *posteriori* da variância residual (assumindo uma *piori* não informativa). Isto se dá pelo cálculo $(n_1 - k)s^2 / X^2$, onde X^2 representa uma realização de uma distribuição Qui-quadrado, com $(n_1 - k)$ graus de liberdade e $s^2{}_{[1]}$ a primeira observação desta realização;
- considerar uma realização da distribuição a *posteriori* dos coeficientes de regressão, por meio de uma distribuição normal multivariada com média β e matriz de covariância $s^2{}_{[1]}(\boldsymbol{X'X})^{-1}$, onde $\boldsymbol{X}$ é uma matriz (n_1 x *k*) dos valores de *X*. Informações práticas sobre estes cálculos podem ser encontradas em Schafer (1997).

Para cada novo conjunto de parâmetros de regressão, os valores preditos são gerados para todos os casos. Assim, para cada caso em *Y* com dados faltantes, temos um grupo de doadores baseados nos valores preditos e escolhemos aleatoriamente um dos valores observados de *Y*, a partir do conjunto de doadores.

2.5.4.3 – Métodos MCMC (*Markov Chain Monte Carlo)*

O método da Cadeia de Markov de Monte Carlo tem o objetivo de simular distribuições multivariadas, que tenham como limite uma cadeia de Markov estacionária com a distribuição que queremos encontrar (Nunes, 2007).

A figura 4 a seguir mostra como o método MCMC é utilizado para a imputação, conforme o esquema proposto por Chantala e Suchindran (2005 *in* Nunes, 2007).

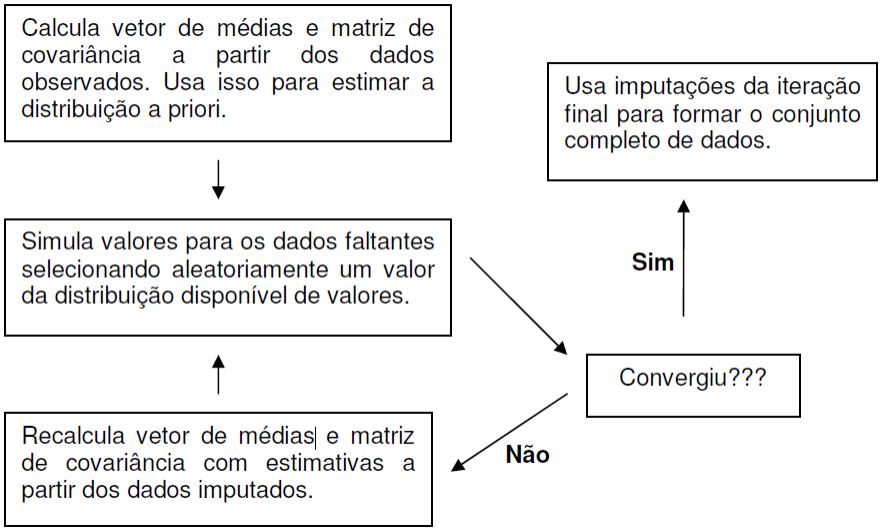


Figura 4: Esquema do MCMC para imputação
Fonte: (Nunes, Métodos de imputação de dados aplicados na área da saúde, 2007)

2.6 – Regressão Logística

Como já mencionado, um dos objetivos dessa monografia é verificar se existe diferença nos resultados encontrados nas análises das bases de dados imputadas e incompletas. Essa diferença é investigada a partir do estudo do baixo peso ao nascer na base de dados da PNDS 2006.

A Regressão Logística considera modelos capazes de predizer os valores assumidos por uma variável categórica, no presente caso binária, a partir de outras variáveis explicativas, contínuas e/ou categóricas. Em outras palavras, a regressão logística estima as chances de um evento de interesse ocorrer, em funções das variáveis explicativas.

O Modelo de Regressão Logística e Regressão Linear se diferem em função da variável resposta. Em modelos de regressão linear simples e múltipla a resposta Y é uma variável aleatória contínua, com distribuição supostamente normal. Mas temos situações em que a resposta é uma variável dicotômica, geralmente categorizada com o valor 1 para representar "sucesso" (quando a característica de interesse é observada) e 0 para representar "fracasso" (Pagano, 2006).

A proporção de ocorrência de sucessos p é a média da variável aleatória dicotômica Y, ou seja, p = P(Y=1) = P (sucesso). Podemos estimar a probabilidade p, associada a uma resposta dicotômica, para diferentes valores de uma variável explicativa, a partir da Regressão Logística, de forma que:

$$p = \alpha + \beta x \qquad (19)$$

Onde:

x representa uma variável explicativa;

α é o intercepto da linha;

β é a inclinação da linha.

No entanto, não podemos usar o modelo como na equação (19), uma vez que *p* é uma probabilidade, ou seja, $0 \leq p \leq 1$, e $\alpha + \beta x$ pode gerar valores não pertencentes a esse intervalo. Mas podemos ajustar o modelo (19) a partir da função logística, garantindo que as estimativas produzidas respeitem esta restrição. Assim, temos o modelo abaixo (Pagano, 2006):

$$p = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} \qquad (20)$$

Se um evento ocorre com probabilidade *p*, a chance a seu favor é de *p*/(1-*p*) para um. Portanto, se a probabilidade de sucesso é (20), a chance favorável o sucesso é:

$$\frac{p}{1-p} = \frac{e^{\alpha+\beta x}/(1 + e^{\alpha+\beta x})}{1/(1 + e^{\alpha+\beta x})} = e^{\alpha+\beta x} \qquad (21)$$

Aplicando o logaritmo natural de cada lado da equação (21), temos:

$$ln\left(\frac{p}{1-p}\right) = ln\left(e^{\alpha+\beta x}\right) = e^{\alpha+\beta x}$$

Portanto, quando modelamos a probabilidade *p* com a função logística estamos ajustando um modelo de regressão onde a resposta contínua *y* é substituída pelo logaritmo da chance de sucesso de uma variável aleatória dicotômica. Não existe uma relação linear entre *p* e *x*, mas sim entre ln(*p*/1-*p*) e *x*, ou seja:

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\alpha} + \hat{\beta} x \qquad (22)$$

Como no modelo linear, $\hat{\alpha}$ e $\hat{\beta}$ são estimadores dos coeficientes da população. Porém, ao invés de ajustarmos o modelo logístico pelo método de mínimos quadrados, que assume que a variável resposta é contínua e normalmente distribuída, utilizamos a estimação de máxima verossimilhança, que usa informações de uma amostra para estimar os parâmetros mais prováveis de terem produzido os dados observados (Pagano, 2006).

Analogamente, temos a Regressão Logística Múltipla, em que a probabilidade *p* pode ser modelada em função de duas ou mais variáveis explicativas, discretas ou contínuas, conforme a equação (23):

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_i x_i, \qquad i = 1,2,\ldots,n \quad (23)$$

A interpretação da regressão logística se dá a partir da razão das probabilidades de sucesso e fracasso, pela função de *odds ratio* (OR).

Se substituirmos dois valores diferentes x_1 e x_2 da variável explicativa na equação (22), temos:

$$OR = \frac{e^{\alpha+\beta x_2}}{e^{\alpha+\beta x_1}} \quad (24)$$

Aplicando o logaritmo natural à equação (24), obtemos:

$$\begin{aligned} ln(OR) &= \ln\left(\frac{e^{\alpha+\beta x_2}}{e^{\alpha+\beta x_1}}\right) \\ &= \ln\left(e^{\alpha+\beta x_2}\right) - \ln\left(e^{\alpha+\beta x_1}\right) \\ &= \alpha + \beta x_2 - (\alpha + \beta x_1) \\ &= \beta(x_2 - x_1) \\ ln(OR) &= e^{\beta} = \beta \end{aligned}$$

A OR nos permite fazer comparações entre grupos distintos, em relação a característica de interesse. Por exemplo, no presente estudo sobre BPN, podemos analisar a prevalência de baixo peso em função do sexo a partir da OR.

O objetivo da regressão logística é propor um modelo capaz de prever *y* em função de uma ou mais variáveis explicativas, queremos então que esse modelo seja o mais parcimonioso possível. A significância dos coeficientes para a inclusão ou não no modelo pode ser testada por técnicas diferentes, como seguem descritas abaixo (Agresti, 2013). Testamos as hipóteses H_0: $\beta_i = 0$ *vs*. H_1: $\beta_i \neq 0$. O teste de Wald é utilizado para testar a significância estatística de cada coeficiente β_i no modelo.

Temos que a estatística de teste, denotada por *W*, é obtida comparando a estimativa de máxima verossimilhança do parâmetro $\hat{\beta}$ com a estimativa do seu erro padrão é $W_i = \frac{\hat{\beta}_i}{\widehat{dp}(\hat{\beta}_i)}$, que, sob H_0, segue uma distribuição normal padrão. O p-valor é definido por $P(|Z|>|W_i|)$, onde Z é a variável aleatória da normal padrão.

O teste da Razão de Verossimilhança, por sua vez, usa a razão entre a verossimilhança do modelo simples (L_0), com apenas uma variável explicativa e a verossimilhança do modelo saturado (L_1), com mais de uma variável explicativa. Ou seja, comparamos os valores

observados com os valores preditos pelos modelos simples e saturado (L_1), a partir da função de verossimilhança. A estatística deste teste é dada por:

$$G = -2\,ln\frac{L_0}{L_1} = -2\,ln(L_0 - L_1) \quad (25)$$

Sob H_0, G segue uma distribuição Qui-quadrada com 1 grau de liberdade. O p-valor é definido por $P(X_1^2 > G)$.

Outro teste frequentemente adotado neste contexto é o teste *Score*, que possui estatística dada por:

$$ST = \frac{\sum_i^n x_i(y_i - \bar{y})}{(\bar{y}(1-\bar{y})\sum_i^n (x_i - \bar{x})^2)^{1/2}} \quad (26)$$

Onde $\bar{y} = \hat{p}$, proporção de sucessos na amostra. Neste caso, o p-valor é definido por P(|Z|>|ST|), onde Z é a variável aleatória da normal padrão.

De forma resumida, podemos afirmar que a regressão logística pode ser utilizada para (Faculty & Staff, 2014):

- prever uma variável dependente categórica com base em variáveis independentes contínuas e / ou categóricas;
- determinar a magnitude do efeito das variáveis independentes sobre a variável dependente;
- classificar a importância relativa das variáveis independentes;
- avaliar os efeitos de interações;
- compreender o impacto de variáveis de controle de co-variáveis, geralmente explicado a partir da *odds ratio.*

Capítulo 3

A Pesquisa Nacional de Demografia e Saúde (PNDS)

3.1 - Introdução

Nesta monografia foram utilizados dados provenientes da Pesquisa Nacional de Demografia e Saúde da Criança e da Mulher (PNDS 2006), que está inserida na 5ª fase do projeto internacional MEASURE DHS (*Demographicand Health Survey*), pesquisa em escala global que teve apoio da Agência dos Estados Unidos para o Desenvolvimento (USAID). As pesquisas DHS têm como objetivo

> "*prover dados e análises para um amplo conjunto de indicadores de planejamento, monitoramento e avaliação de impacto nas áreas de população, saúde e nutrição de mulheres e crianças nos países em desenvolvimento* (Brasil, Pesquisa Nacional de Demografia e Saúde da Criança e da Mulher, 2009)."

Portanto, seus resultados permitem comparações a nível nacional e internacional e fornecem subsídios para avaliar os avanços ocorridos na saúde da mulher e da criança no Brasil. A pesquisa de 2006 foi a terceira edição realizada, sendo que a primeira ocorreu em 1986 (Pesquisa Nacional sobre Saúde Materno-Infantil e Planejamento Familiar – PNSMIPF) e a segunda em 1996 (Pesquisa Nacional sobre Demografia e Saúde - PNDS).

As principais motivações da pesquisa são: (*i*) estudar a população feminina em idade fértil e crianças com idade inferior a cinco anos, no que diz respeito a fatores demográficos, socioeconômicos e culturais; (*ii*) identificar padrões conjugais, de parentescos e reprodutivos; (*iii*) identificar perfis de morbi-mortalidade a infância bem como de amamentação; (*iv*) avaliar o estado nutricional, a segurança ou insegurança alimentar e o teor de iodo disponível no ambiente domiciliar; além de (*v*) avaliar o acesso a serviços de saúde e medicamentos (Brasil, Pesquisa Nacional de Demografia e Saúde da Criança e da Mulher, 2009).

A PNDS 2006 pode ser considerada um estudo de coorte transversal, permitindo identificar a situação atual da população quanto às informações disponíveis na pesquisa, tanto demográficas e socioeconômicas quanto àquelas relacionadas a saúde. Além disso, também permite a recuperação de informações retrospectivas e coleta de dados prospectivos.

Os dados coletados pela PNDS 2006 são armazenados em seis arquivos, sendo um para cada um dos seguintes temas: mulheres, gravidezes, filhos, domicílios, além de dois sobre medicamentos. Foram coletados dados de 15.575 mulheres entre 15 e 49 anos, totalizando 27.477 registros de filhos na história de nascimentos e 6.833 registros de gravidezes na história de gravidezes e perdas. A pesquisa também reuniu informações de 14.617 domicílios.

3.2 - Desenho Amostral da PNDS 2006

A PNDS 2006 é uma pesquisa domiciliar, por amostragem probabilística complexa, com representatividade nacional no Brasil. As unidades amostrais foram selecionadas em dois estágios dentro de cada estrato: setores censitários (unidades primárias de amostragem) e unidades domiciliares (unidades secundárias de amostragem). Sendo assim, o desenho amostra desta pesquisa pode ser classificado como amostragem estratificada e conglomerada simples em dois estágios.

Antes da ida da PNDS à campo, foi realizado um estudo piloto com 324 domicílios em 5 municípios do País, no período de 10 de junho a 06 de julho de 2006. A pesquisa completa, por sua vez, foi realizada em 14.617 domicílios de 674 municípios, no período de 3 de novembro de 2006 a 3 de maio de 2007. Os domicílios pertencentes à população alvo da pesquisa eram particulares permanentes, de acordo com a definição do Instituto Brasileiro de Geografia e Estatística, selecionados em dez estratos amostrais independentes, resultantes da combinação das cinco macrorregiões geográficas brasileiras e das áreas urbanas e rurais.

A coleta de dados foi realizada a partir de entrevistas pessoais nos domicílios selecionados, classificados em elegíveis (domicílios em que residia ao menos uma mulher em idade fértil – entre 15 e 49 anos) e não elegíveis (não residia nenhuma mulher em idade fértil). O questionário de domicílios foi aplicado em todos os domicílios, porém, naqueles do tipo não elegível, só foram perguntadas as questões até a informação de educação do responsável pelo domicílio, ficando sem respostas as demais perguntas do questionário. Nos domicílios elegíveis, além do questionário dos domicílios completo, foi aplicado também o questionário das mulheres.

Foram realizadas mensurações antropométricas, como altura e peso das mulheres e crianças e circunferência da cintura das mulheres. Além disso, foram coletadas amostras de sangue para dosagens de vitamina A e Hemoglobina. E também, foi medido o teor de iodo disponível no sal consumido nos domicílios.

3.3- Problema Motivador: Estudo do Baixo Peso ao Nascer

Como já mencionado na introdução, um dos objetivos dessa monografia é aplicar os métodos de imputação estudados a uma base de dados real, e realizar comparações entre ajustes de modelos de regressão logística para ocorrência de baixo peso ao nascer (BPN). As variáveis consideradas como explicativas neste modelo estão disponíveis nas bases de dados da PNDS (2006) e são descritas na próxima seção.

O BPN é um importante indicador de saúde pública que também pode ser utilizado como *proxy* para o estudo do desenvolvimento socioeconômico de uma população, uma vez que está diretamente relacionado à morbimortalidade infantil. A Organização Mundial de Saúde (OMS) considera como baixo peso ao nascer, recém-nascidos com peso inferior a 2,500 Kg, e aqueles que nascem com peso igual ou superior a 2,500 Kg são considerados com peso adequado.

No Brasil, a prevalência de baixo peso em nascidos vivos vem aumentando, sendo de 7,75% em 1996 e 8,53% em 2011, havendo diferenças entre as regiões: em 2011 a proporção de BPN foi de 7,34% na região Norte e 9,28% na região Sudeste, conforme dados do Ministério da Saúde (DATASUS). A ocorrência do BPN pode estar relacionada à: (*i*) fatores socioeconômicos e demográficos, como escolaridade materna, estado conjugal, cor da pele; (*ii*) fatores relacionados à mulher como variáveis antropométricas (peso, altura, ganho de peso gestacional), reprodutivas (número de gestações, histórico de aborto, intervalo entre as gestações) e comportamentais (tabagismo, consumo de álcool); e (*iii*) fatores relacionados à gestação (cuidado no pré-natal) ((Paula, Salvador, Barbosa, & Cotta, 2011), (Melo, Kassar, Lira, Coutinho, Eickmann, & Lima, 2013), (Rojas, Francisco, Siqueira, & Carminatti, 2012)).

3.4 - Variáveis Consideradas no Modelo

A PNDS (2006) é composta por diferentes bases de dados e variáveis, que fornecem variadas informações sobre as mulheres, tanto de aspectos biológicos e socioeconômicos quanto culturais.

Para a construção da base de dados utilizada nessa monografia foram considerados os bancos de dados de Domicílio, Mulheres e Filhos, que foram ligados através da variável identificadora da mulher no domicílio. A população alvo considerada nesta monografia inclui apenas filhos nascido-vivos entre os anos 2001 e 2006, resultando em um subconjunto de 6.068 crianças com idade de até cinco anos.

A base de dados final consistiu na seleção das variáveis relevantes para o estudo do baixo peso ao nascer consideradas no modelo, tendo como base resultados publicados na literatura (Araújo & Sant'Ana (2003), Sclowitz (2007), Araújo D.M. (2012), Benício, Monteiro, Souza, Castilho & Lamonica (1985), Minamisava, Barbosa, Malagoni & Andraus (2004), Caçola & Bobbio (2010), Bisceski et. al. (2012)).

No quadro 3 temos a relação das variáveis que foram selecionadas e em seguida, são consideradas algumas justificativas para inclusão de cada variável no estudo do baixo peso ao nascer.

Quadro 3: Variáveis consideradas na base de dados

Nome	Variável original	Variável recodificada
IDADE_M	Idade da mãe ao ter o filho	1 Até 20 anos 2 De 20 a 34 anos 3 35 anos ou mais
PESO	Peso ao nascer da criança	0 Baixo Peso ao Nascer (<2,5Kg) 1 Peso adequado (≥ 2,5Kg)
EST_CONJUGAL	Estado conjugal da mulher 1 Casada 2 Em união estável 3 Viúva 4 Separa/disquitada/divorciada 5 Não está em união	1 Casada/Em união estável 2 Sem união/Separada/Viúva
FUMANTE	Mulher é fumante 0 Sim 1 Não	0 Sim 1 Não
COR	Classificação da cor da mulher 1 Branca 2 Preta 3 Parda 4 Amarela 5 Indígena	1 Branca 2 Não branca
ABORTO	Mulher já abortou? 1 Sim 2 Não	1 Sim 2 Não
PARTOS	Número de partos que a mulher teve	1 Primeiro filho 2 2 ou 3 filhos 3 Mais de 3 filhos
CESAREA	O parto foi cesária 1 Sim 2 Não	1 Sim 2 Não
INTERNADO	Depois do parto, a criança ficou internada no hospital 1 Sim 2 Não	1 Sim 2 Não

Variáveis consideradas na base de dados (Continuação)

Nome	Variável original	Variável recodificada
MOT_INTERNADO	Motival pelo qual a criança ficou internada 1 Ganhar peso 2 Banho de luz 3 Teve infecção 4 Nasceu antes do tempo	1 Ganhar peso 2 Banho de luz 3 Teve infecção 4 Nasceu antes do tempo
SEXO	Sexo da criança 1 Masculino 2 Feminino	1 Masculino 2 Feminino
INTERVALO	Intervalo em meses entre os partos (anterior)	1 Primeiro filho 2 Intervalo curto (≤18 meses) 3 Intervalo longo(>18meses)
PLAN_SAUDE	Mulher tem convênio ou plano de saúde 1 Sim 2 Não	1 Sim 2 Não
QUANT_CONS	Quantas consultas de pré-natal a mulher realizou durante a gravidez	1 Menos de 6 consultas 2 6 ou mais consultas
MESES_CONS	Meses de gravidez quando realizou a primeira consulta pré-natal	1 Até 3 meses 2 Mais de 3 meses
INJEÇÃO	Durante a gravidez, tomou injeção para prevenir o bebê contra tétano 1 Sim 2 Não	1 Sim 2 Não
PRE_NATAL	Mulher realizou um pré-natal adequado*	1 Adequado 2 Inadequado
REGIÃO	Macro região administrativa 1 Norte 2 Nordeste 3 Sudeste 4 Sul 5 Centro-oeste	1 Norte e Nordeste 2 Demais regiões
SITUAÇÃO	Situação do domicílio 1 Urbano 2 Rural	1 Urbano 2 Rural
MESES_ULT	Quantos meses de gravidez a mulher tinha quando realizou a última consulta pré-natal	1 Pré-termo (<9 meses) 2 A termo (≥ 9 meses)
ÁCIDO_FOL	Durante a gravidez, tomou ácido fólico 1 Sim 2 Não	1 Sim 2 Não
QTO_ÁCIDO_FOL	Durante quanto tempo tomou ácido fólico	1 Primeiro trimestre 2 Segundo ou terceiro trimestre

***Foram considerados como pré-natal adequado, àqueles realizados com no máximo 3 meses de gravidez, se foram realizadas um mínimo de 6 consultas pré-natal e se tomou injeção para prevenção do tétano**

- Idade da mãe ao ter o filho: filhos de mães adolescentes têm maior risco de nascer com baixo peso, além de maior risco de morbi-mortalidade infantil. A imaturidade do sistema genital e o ganho de peso inadequado das mães adolescentes, bem como a marginalidade social e pobreza e o estilo de vida são fatores que podem justificar esse

risco. Já as mães com idade superior a 35 anos, têm seu sistema reprodutivo menos eficiente.

- Estado conjugal: mulheres que vivem em união estável dividem as responsabilidades e cuidados da gestação com seus parceiros. Mulheres solteiras ou que vivem uma relação instável com seus parceiros, não tem suporte familiar e deixam de se preocupar com a gestação, prejudicando o crescimento e desenvolvimento do feto.
- Tabagismo: estudos indicam quanto maior a proporção de mães tabagistas, maior a ocorrência de baixo peso ao nascer e também menor a duração da gestação. Além disso, a interrupção do tabagismo durante a gestação indica um aumento no peso ao nascer, ao passo que filhos de mãe tabagistas no terceiro trimestre tiveram uma redução no peso ao nascer, para cada cigarro fumado por dia (Bernstein, 2005, *in* Sclowitz, 2007).
- Cor/raça da mãe: a cor/raça, em geral, reflete a condição socioeconômica da mãe e a vulnerabilidade a determinadas doenças pode ser fruto da existência de desigualdades sociais e falta de acesso a saúde.
- Histórico de aborto: mulheres que já abortaram em outras gestações estão mais suscetíveis à complicações em futuras gestações, ou até mesmo, outro aborto.
- Número de partos: o número elevado de filhos pode ser relacionado com idades maternas maiores e/ou intervalos curtos entre os nascimentos, fatores de risco para o baixo peso ao nascer e morbi-mortalidade infantil.
- Intervalo entre os partos: intervalo curto entre os nascimentos, ou seja, intervalos inferiores a 18 meses, podem comprometer a saúde materna, uma vez que a mãe não teria recuperado todos os nutrientes necessários para passar para o feto.
- Parto foi cesárea: o aumento de intervenções médicas, como cesarianas e induções de partos, aumenta o número de partos prematuros, que está diretamente relacionado com o baixo peso ao nascer.
- Criança permaneceu internada após o parto e o motivo da internação: esta variável foi utilizada como preditora para a imputação da variável “peso ao nascer”.
- Sexo da criança: a prevalência da morbimortalidade é maior em crianças do sexo masculino e o baixo peso é fator de risco para a morbimortalidade. Além disso, o sexo feminino amadurece o pulmão mais rápido que o masculino, o que protege contra complicações respiratórias.

- Mãe tem plano de saúde: a saúde do recém-nascido está estritamente relacionada ao comportamento da gestante durante o período gestacional. Além de um fator que ligado a condição econômica, está relacionado com a assistência ao pré-natal e acesso aos serviços de saúde.
- Quantas consultas pré-natais a mulher realizou durante a gravidez, Quantos meses de gravidez na primeira consulta pré-natal e Injeção contra tétano: estas variáveis foram consideradas para classificação do Pré-natal como adequado ou não, conforme abaixo.
- Pré-natal foi adequado: o pré-natal é de suma importância para garantir a saúde materno-infantil, podendo ser utilizando para avaliação da qualidade de assistência e acesso aos serviços de saúde. O preconizado pelo Ministério da Saúde, para que um pré-natal seja considerado adequado, é que a primeira consulta pré-natal seja realizada no máximo ao terceiro mês de gravidez e que sejam realizadas um mínimo de seis consultas pré-natais. Além disso, a mãe deve receber vacinação antitetânica. O acompanhamento da mãe durante a gravidez é importante para determinação de possíveis problemas que possam comprometer a saúde materno-fetal, como por exemplo, a falta de nutrientes necessários e o não crescimento do bebê.
- Macro região administrativa: as regiões Norte e Nordeste do Brasil se encontram em diferentes estágios de transição demográfica e de desenvolvimento socioeconômico quando comparadas as demais regiões brasileiras. Além desses, outros fatores como parturição elevada, piores condições de saneamento básico e níveis de escolaridade mais baixos, podem justificar as diferenças existentes nas proporções de baixo peso ao nascer.
- Situação do domicilio: zonas rurais são menos favorecidas em termos de acesso aos serviços de saúde, podendo impossibilitar diagnósticos precoces e o tratamento de condições que acarretam no baixo peso aos nascer.
- Meses de gravidez na última consulta pré-natal: como não temos a variável “idade gestacional”, utilizamos esta informação para verificar se a criança nasceu prematura ou não. A prematuridade está diretamente relacionada ao baixo peso ao nascer, uma vez que o período de gestação foi menor e o bebê não cresceu como esperado, nem se desenvolveu completamente.
- Tomou ácido fólico durante a gravidez e por quanto tempo tomou: o ácido fólico é receitado pelo médico a todas as gestantes no primeiro trimestre de gravidez, para

ajudar no desenvolvimento e dar nutrientes ao bebê, além disso, é fundamental para o fechamento do tubo neural do bebê. A deficiência de nutrientes durante a gestação pode desencadear um menor desenvolvimento do feto quanto ao seu tamanho, bem como, prejudicar a formação dos órgãos. O fato de tomar ácido fólico por mais de três meses (em especial no último trimestre), pode ser um indício de que a mãe não está fornecendo os nutrientes necessários para o crescimento e desenvolvimento da criança, podendo explicar o baixo peso ao nascer.

No capítulo 4 temos a aplicação dos métodos de imputação estudados nessa monografia à base de dados da PNDS 2006, bem como as comparações entre os resultados obtidos a partir das regressões das bases imputadas e da base incompleta.

Capítulo 4

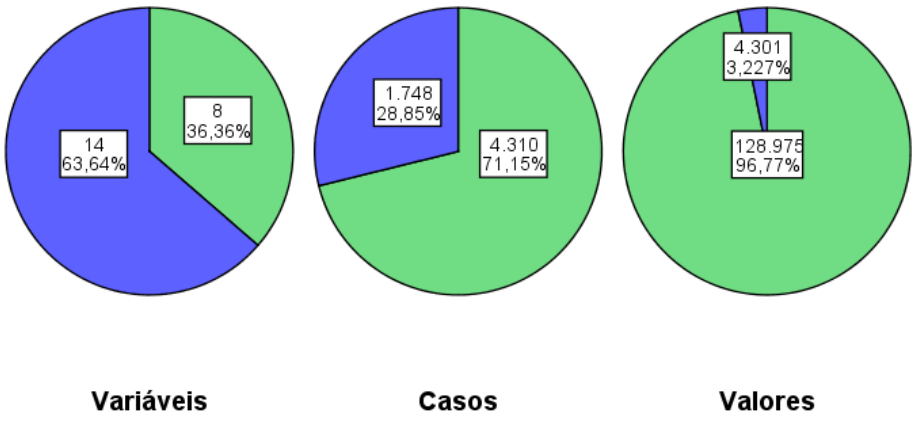
Aplicação dos Métodos de Imputação

Este capítulo é dedicado à aplicação dos métodos de imputação, análise e comparação dos modelos para o BPN, propostos a partir das bases de dados imputadas e dados faltantes. Entretanto, antes da imputação, precisamos definir algumas características dos dados faltantes, isto é, analisar o padrão e o mecanismo que geraram as não respostas na nossa base de dados. Além disso, precisamos decidir qual é a técnica de imputação mais adequada para tratá-las: única ou múltipla.

A escolha entre imputação única ou múltipla depende da proporção de dados faltantes (λ) nas variáveis. Se $\lambda \leq 0{,}05$ podemos utilizar a imputação única ou mesmo ignorar os casos com não resposta e analisar apenas os dados completos. Se λ for entre 0,05 e 0,15, podemos utilizar imputação única ou múltipla. E se $\lambda \geq 0{,}15$ devemos utilizar a imputação múltipla (Harrell, 2001 *in* Nunes, 2007).

O Gráfico 2 trás a distribuição de λ entre as variáveis, os casos e a matriz de dados como um todo (valores). Em verde temos a proporção de dados completos e em azul a proporção de dados incompletos.

Gráfico 2: Resumo geral dos dados faltantes



Pelo Gráfico 2, temos que a proporção de variáveis com dados faltantes é de 63,64%, ou seja, ocorreu não resposta em 14 das 22 variáveis selecionadas para o estudo do BPN. Em

relação aos casos, a proporção de dados faltantes é de 28,85%, ou seja, das 6058 crianças estudadas, falta ao menos uma informação para 1748 delas. Finalmente, a proporção de dados faltantes em todas as células da matriz de dados (22 variáveis x 6058 casos) é de 3,23%, ou seja, temos 4301 valores em branco.

Na Tabela 1 estão relacionadas as variáveis com maior proporção de dados faltantes.

Tabela 1: Resumo das variáveis com dados faltantes

Variáveis	Porcentagem
Durante grav. tomou ácido fólico por quanto tempo	10,8%
Pré-natal adequado	9,3%
Peso ao nascer	9,1%
Quantas consultas de pré-natal fez durante a gravidez	8,7%
Durante gravidez tomou ácido fólico	8,6%
Meses gravidez quando fez a 1a. consulta pré-natal	5,0%
Meses de gravidez fez a última consulta pré-natal	5,0%
Motivo bebê permaneceu internado	4,4%
Bebê recebeu alta após o parto	4,3%
Tomou injeção para prevenir o bebê contra tétano	3,6%
Classificação de cor da mãe	1,1%
O parto do bebê foi cesárea	0,6%
Intervalo interpartal anterior	0,40%
Tem convênio ou plano de saúde	0,10%

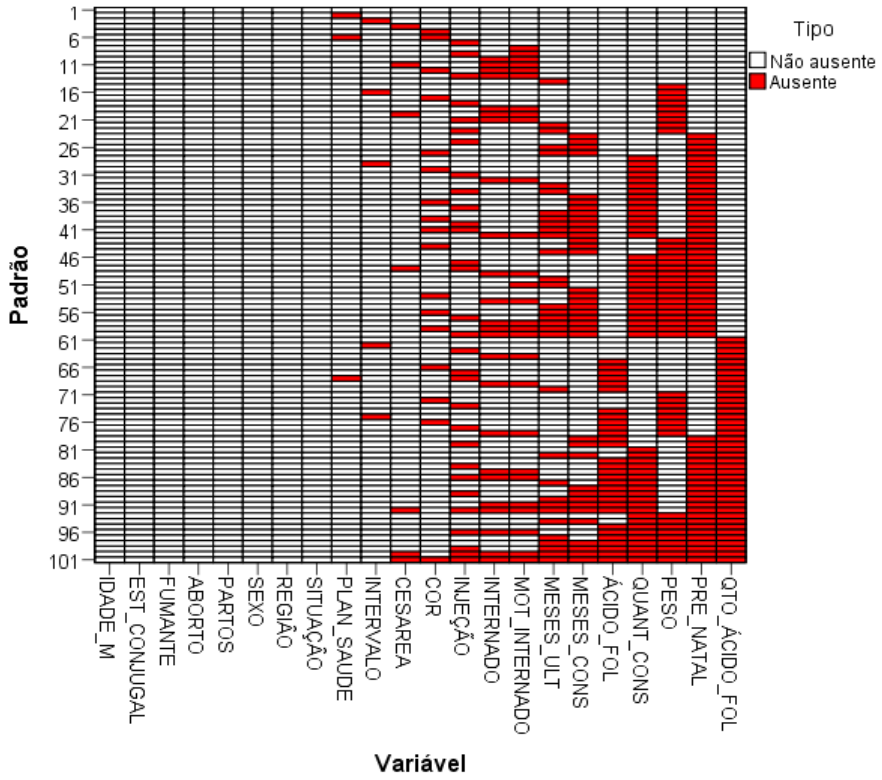
Apesar de termos apenas 3,23% de dados faltantes entre as células da matriz de dados, a variável de interesse do estudo “Peso ao nascer” apresenta uma proporção de 9,1% de dados faltantes. Portanto, decidimos pela a imputação múltipla.

O próximo passo é analisar os padrões de não resposta, que estão apresentados no Gráfico 3. As células vermelhas representam os dados faltantes e as brancas os dados observados, e cada linha do gráfico representa um padrão e mostra um grupo de casos com o mesmo padrão. Os padrões ou grupos de casos são apresentados baseados na localização dos dados faltantes, ou seja, em cada variável. No eixo x estão representadas as variáveis, ordenadas pela proporção de dados faltantes conforme a Tabela 1. As primeiras variáveis do gráfico não apresentam dados faltantes e estão todas em branco. A primeira célula vermelha aparece na variável “Plano de saúde” que detém a menor proporção de dados faltantes, enquanto a variável “Quanto tempo tomou ácido fólico” detém a maior.

No primeiro padrão temos os casos sem observações faltantes. O segundo mostra os casos com dados faltantes apenas na variável com menor λ, neste caso, “Plano de saúde”. E assim sucessivamente, até o último padrão em que todas as variáveis, com exceção das duas primeiras, apresentam dados faltantes.

Poderíamos identificar a presença de monotonicidade dos padrões, isso é, aumento ou diminuição sistemática em uma sequência, se todas as células brancas e vermelhas estivessem agrupadas. Quando observarmos aglomerados ou "ilhas" das células, como é o caso do padrão observado no Gráfico 3, tais resultados sugerem que não temos monotonicidade.

Gráfico 3: Padrões de dados faltantes



Análises a partir do gráfico acima não nos permitem concluir que o mecanismo de não resposta seja do tipo MNAR. Desta forma, tal observação corrobora nossa decisão pela adoção do método de IM por MCMC, trabalhando com o pressuposto de que o mecanismo gerador de não resposta foi do tipo MAR e que o padrão de não resposta é não monotônico.

4.1 – Estatísticas Descritivas da Não Resposta nas Variáveis Estudadas

Foi realizada uma análise descritiva das variáveis selecionadas para o estudo do BPN. As variáveis quantitativas estão descritas pelas medidas resumo média, desvio padrão, mínimo e máximo, e as variáveis categóricas, pelas frequências relativa e absoluta, considerando os percentuais dos casos válidos. As estatísticas descritivas dessa seção referem-se à base de dados incompleta e estão resumidas nas Tabelas 2 e 3.

Tabela 2: Estatísticas descritivas para as variáveis categóricas

Variáveis categóricas		n (%)
REGIÃO	Norte	1324 (21,9%)
	Nordeste	1194 (19,7%)
	Sudeste	1154 (19,0%)
	Sul	1143 (18,9%)
	Centro-Oeste	1243 (20,5%)
SITUAÇÃO	Urbano	3925 (64,8%)
	Rural	2133 (35,2%)
COR	Branca	2044 (34,1%)
	Preta	608 (10,2%)
	Parda	3017 (50,4%)
	Amarela	166 (2,8%)
	Indígena	151 (2,5%)
EST_CONJUGAL	Casada	2222 (36,7%)
	Em união estável	2887 (47,7%)
	Viúva	36 (0,6%)
	Separada, Divorciada	607 (10,0%)
	Não está em união	306 (5,1%)
FUMANTE	Sim	965 (15,9%)
	Não	5093 (84,1%)
PLAN_SAUDE	Sim	945 (15,6%)
	Não	5106 (84,4%)
ABORTO	Sim	1348 (22,3%)
	Não	4710 (77,7%)
PRE-NATAL	Adequado	3016 (54,9%)
	Inadequado	2477 (45,1%)
INJEÇÃO	Sim	4185 (71,6%)
	Não	1656 (28,4%)
ÁCIDO_FOL	Sim	1814 (32,8%)
	Não	3724 (67,2%)
SEXO	Masculino	3153 (52,0%)
	Feminino	2905 (48,0%)
CESÁREA	Sim	2387 (39,6%)
	Não	3635 (60,4%)
INTERNADO	Não	5502 (94,9%)
	Sim	242 (4,2%)
	Morreu antes da alta	52 (0,9%)
MOT_INTERNADO	Ganhar peso	66 (27,6%)
	Banho de luz	43 (18,0%)
	Teve infecção	32 (13,4%)
	Nasceu antes do tempo	40 (16,7%)
	Outro	58 (24,3%)

A amostra de 6058 crianças se distribui de forma semelhante pelas regiões, mas não pela situação do domicílio, sendo 64,8% deles urbanos. Quanto à cor/raça, 34,1% das mães se declararam brancas, além disso, mais da metade delas alegaram ser casadas ou estar em uma união estável (84,4%). O tabagismo foi presente em 15,9% da amostra. Apenas 15,6% das mães possuem plano ou convênio de saúde e 22,3% já sofreram aborto. A proporção de adequação do pré-natal foi de 54,9% na amostra e 71,6% das mães receberam a vacina para prevenir o bebê contra tétano. Apenas 32,8% das mães tomaram ácido fólico durante a

gravidez. Em relação a realização ao sexo dos recém-nascidos, 52% foram do sexo masculino e 48% feminino. Quanto ao parto, a proporção de cesáreas foi de 39,6%. Após o parto, apenas 4,2% dos recém-nascidos permaneceram internados e 0,9% vieram a óbito antes de receber alta. Dos 242 recém-nascidos que permaneceram internados, 27,6% foi para ganhar peso e 16,7% porque nasceram antes do tempo.

Tabela 3: Estatísticas descritivas para as variáveis quantitativas

Variáveis quantitativas	Média ± desv. padrão (Mín - Máx)
PESO	3,24 ± 0,55 (0,5 - 5,5)
IDADE_M	25,22 ± 6,28 (12 - 46)
PARTOS	2,62 ± 1,9 (1 - 15)
INVERVALO	35,28 ± 36,06 (10 - 266)
QUANT_CONS	7,58 ± 3,23 (1 - 32)
MESES_CONS	2,33 ± 1,5 (0 - 9)
MESES_ULT	8,58 ± 0,78 (1 - 10)
QTO_ÁCIDO_FOL	1,12 ± 2,22 (0 - 9)

A média dos pesos dos recém-nascidos foi de 3,24 Kg, com desvio padrão de 0,55 Kg variando de 0,5 Kg a 5,5 Kg. As idades das mães ao ter o filho variaram de 12 a 46 anos, com média de 25,22 anos e desvio padrão de 6,28 anos. Quanto ao número de partos, ou número de filhos, a média foi de 2,62 filhos e o desvio padrão foi 1,9 filhos, com mínimo de 1 filho e máximo de 15 filhos. A média dos intervalos entre os filhos foi de 35,28 meses, com desvio padrão de 36,06 meses, variando entre 10 e 266 meses. O número médio de consultas pré-natais foi 7,58 com desvio padrão 3,23, com mínimo de 1 consulta e máximo 32. A média de meses de gravidez na primeira consulta pré-natal foi 2,33 meses e desvio padrão 1,5, variando de 0 a 9 meses. Já a média de meses de gravidez na última consulta pré-natal foi 8,58 meses e desvio padrão 0,78, com mínimo de 1 mês e máximo de 10. Por fim, a média de tempo tomando ácido fólico durante a gravidez foi 1,12 meses, desvio padrão 2,22 meses, variando de 0 a 9 meses.

4.2 – Aplicação do Método de Imputação

No processo de imputação, 17 variáveis foram selecionadas para serem imputadas e entrarem como preditoras no modelo de imputação. Destas, oito estavam completas e não precisaram ser imputadas. As demais variáveis, entraram no modelo de imputação apenas como preditoras no modelo e são as seguintes: “Bebê recebeu alta depois do parto”, “Motivo pelo qual o bebê permaneceu internado”, “Tomou ácido fólico durante a gravidez”, “Por quanto tempo tomou ácido fólico” e “Meses de gravidez na última consulta pré-natal”.

A fim de se evitar que valores não realistas fossem imputados, estipulou-se limites baseados nas estatísticas descritivas de mínimo e máximo das variáveis quantitativas incompletas, conforme a Tabela 3 acima. Foram realizadas *m*=5 imputações, com 10 iterações para cada, pelo método de imputação múltipla MCMC. As estatísticas descritivas do processo de imputação se encontram no Anexo 1, sendo apresentadas em três blocos da seguinte maneira: no primeiro estão incluídas as estatísticas dos dados incompletos, no segundo as estatísticas referentes apenas aos casos que foram imputados e no terceiro temos as estatísticas das novas bases de dados imputadas.

Analisando as variáveis quantitativas observamos que os casos imputados apresentam médias um pouco diferentes e desvios padrão menores que as medidas dos dados incompletos. No entanto, quando comparamos as médias e desvios padrão das cinco diferentes bases imputadas com os dados incompletos, as medidas são muito semelhantes.

As frequências relativas e absolutas das variáveis categóricas acompanham o mesmo comportamento: quantidades diferentes entre os dados originais e os casos imputados e semelhantes entre as cinco bases completas e os dados originais.

Podemos visualizar estes comportamentos a partir da Tabela 4, que apresenta as estatísticas descritivas da imputação da variável de interesse deste estudo, "Peso ao nascer".

Tabela 4: Estatísticas descritivas da imputação da variável Peso

Dados		**N**	**Média**	**Desvio padrão**	**Mínimo**	**Máximo**
Dados originais		5504	3,24	28,78	0,50	5,50
Valores imputados	1	554	3,17	27,04	0,66	5,49
	2	554	3,19	26,59	0,51	5,43
	3	554	3,21	26,65	0,52	5,43
	4	554	3,21	28,35	0,83	5,47
	5	554	3,21	25,99	0,59	5,33
Dados preenchidos após a imputação	1	6058	3,23	28,64	0,50	5,50
	2	6058	3,24	28,59	0,50	5,50
	3	6058	3,24	28,59	0,50	5,50
	4	6058	3,24	28,74	0,50	5,50
	5	6058	3,24	28,54	0,50	5,50

A Tabela 5 mostra a distribuição da variável resposta "Peso ao nascer" classificada, isto é, "Baixo peso ao nascer" (<2,500Kg) e "Peso adequado" (peso≥2,500Kg), após o processo de imputação.

Tabela 5: Distribuição da variável peso ao nascer após a imputação

Número de imputação		N	%
Dados originais	Peso adequado	5101	84,2
	Baixo peso aos nascer	403	6,7
	Dados faltantes	554	9,1
	Total	6058	100,0
1	Peso adequado	5533	91,3
	Baixo peso aos nascer	525	8,7
2	Peso adequado	5537	91,4
	Baixo peso aos nascer	521	8,6
3	Peso adequado	5533	91,3
	Baixo peso aos nascer	525	8,7
4	Peso adequado	5525	91,2
	Baixo peso aos nascer	533	8,8
5	Peso adequado	5550	91,6
	Baixo peso aos nascer	508	8,4
Agrupado	Peso adequado	5536	91,4
	Baixo peso aos nascer	522	8,6
	Total	6058	100,0

Os dados originais sugerem que 6,7% das crianças nasceram com peso inferior a 2,500 Kg, que 84,2% nasceram com peso igual ou superior a 2,500 Kg e apresentou uma proporção de 9,1% de dados faltantes. Analisando a Tabela 4 vemos proporções diferentes. Dos 554 casos com dados faltantes, 21,5% foram imputados como “Baixo peso” e 78,5% como “Peso adequado”, e a proporção de crianças com peso inferior a 2,500 Kg aumentou para 8,6%. A mesma analogia pode ser feita para as demais variáveis que foram imputadas, a partir das tabelas das variáveis categóricas após a imputação, que se encontram no Anexo 2.

Na Tabela 5 temos as associações entre as variáveis selecionadas para o estudo e a variável resposta, verificadas a partir do Teste Qui-quadrado. Foi considerado um nível de significância de 10%. Os resultados referem-se aos dados originais incompletos.

Tabela 5: Teste Qui-quadrado para associação das variáveis

Fatores Associados		Peso ao nascer		P-valor
		Peso adequado	Baixo peso	
Idade da mãe	Até 20 anos	91,3%	8,7%	0,059
	De 21 a 35 anos	93,2%	6,8%	
	Mais de 35 anos	92,9%	7,1%	
Estado conjugal	Em união/Casada	93,0%	7,0%	0,012
	Não está em união/Separada/Viúva	90,7%	9,3%	
Tabagismo	Sim	89,4%	10,6%	<0,001
	Não	93,2%	6,8%	
Cor/raça	Branca	92,3%	7,7%	0,231
	Não branca	92,8%	7,2%	
Aborto	Sim	92,6%	7,4%	0,502
	Não	92,7%	7,3%	

Tabela 5: Teste Qui-quadrado para associação das variáveis (Continuação)

Fatores Associados		Peso ao nascer		P-valor
		Peso adequado	Baixo peso	
Cesárea	Sim	92,1%	7,9%	0,095
	Não	93,1%	6,9%	
Sexo	Masculino	93,2%	6,8%	0,082
	Feminino	92,2%	7,8%	
Intervalo interpartal	Primeiro filho	91,9%	8,1%	<0,001
	Intervalo curto	88,6%	11,4%	
	Intervalo longo	94,0%	6,0%	
Plano de saúde	Sim	93,1%	6,9%	0,317
	Não	92,6%	7,4%	
Número de consultas pré-natal	Menos de 6 consultas	89,5%	10,5%	<0,001
	6 ou mais consultas	93,8%	6,2%	
Meses gravidez no 1º pré-natal	Até 3 meses	93,1%	6,9%	0,028
	4 meses ou mais	91,3%	8,7%	
Injeção antitetânica	Sim	93,4%	6,6%	0,001
	Não	90,8%	9,2%	
Pré-natal	Adequado	94,7%	5,3%	<0,001
	Inadequado	90,7%	9,3%	
Região	Norte e Nordeste	93,7%	6,3%	0,007
	Demais regiões	91,9%	8,1%	
Situação de domicílio	Urbano	92,6%	7,4%	0,421
	Rural	92,8%	7,2%	
Número de filhos	Primeiro filho	91,3%	8,7%	0,022
	2 ou 3 filhos	93,5%	6,5%	
	Mais de 3 filhos	92,5%	7,5%	

A partir da base de dados imputada, foi ajustado um modelo de regressão logística binária, para explicar o baixo peso ao nascer, em função das variáveis que foram significantes no teste de associação de Qui-quadrado, apresentado na Tabela 5. Ao todo, tivemos sete modelos: o modelo baseado nos dados incompletos antes da imputação, um modelo para cada base de dados imputada e o 'modelo agrupado', fruto da combinação das diferentes estimativas das m=5 bases de dados imputadas. A tabelas com os resultados de cada um dos modelos ajustados para cada uma das cinco bases de dados imputadas se encontram no Anexo 3 e são discutidas na próxima sessão.

4.3 – Comparação dos Resultados e Discussão

As análises nessa seção referem-se às comparações entre o modelo com os dados originais incompletos, Modelo 1, e o 'modelo agrupado' das estimativas combinadas das bases imputadas, Modelo 2.

O procedimento de seleção de modelos adotado foi o de *backward* para cada base de dados imputados e no modelo das estimativas combinadas foi adotado o procedimento *enter*. Neste segundo passo, foram candidatas para o modelo todas as variáveis que foram

significativas nos modelos das bases imputadas, considerando um nível de significância de 5%. Em todas as covariáveis categóricas, consideramos como categorias de referência sempre o primeiro valor. Os resultados estão apresentados na Tabela 6.

O principal resultado que observamos é o fato de que as variáveis explicativas significativas nos modelos não são as mesmas. As variáveis que foram significantes no Modelo 1 foram: "Fumante", "Intervalo entre os partos", "Número de consultas pré-natal", "Pré-natal" e "Região". No Modelo 2, as variáveis significativas foram: "Fumante", "Número de consultas pré-natal", "Pré-natal", "Região", "Sexo" e "Número de filhos".

Quando comparamos as variáveis que foram significantes nos dois modelos, observamos que as razões de chance (RC) são diferentes. Na variável "Pré-natal", por exemplo, no Modelo 1 a chance de uma criança nascer com baixo peso é 64% maior para mães que realizaram um pré-natal inadequado, quando comparada com as mães que realizaram um pré-natal adequado, controlando pelas demais variáveis. Já no Modelo 2 (estimativas agrupadas), essa chance é de 56%.

Tabela 6: Modelo de Regressão Logística dos dados originais e agrupados

Variáveis		**Modelo 1**		**Modelo 2**	
		RC (Erro Padrão)	**IC (95%)**	**RC (Erro Padrão)**	**IC (95%)**
Fumante	Sim	-	-	-	-
	Não	0,651 (0,143)*	(0,492 - 0,862)	0,645 (0,134)*	(0,492 - 0,845)
Intervalo interpartal	Primeiro filho	-	-	Não significante	
	Intervalo curto	1,083 (0,723)	(0,723 - 1,623)		
	Inervalo longo	0,635 (0,502)**	(0,502 - 0,803)		
Número de consultas pré-natal	< 6 consultas	-	-	-	-
	≥ 6 consultas	0,727 (0,152)*	(0,540 - 0,978)	0,703 (0,170)*	(0,491 - 1,006)
Pré-natal	Adequado	-	-	-	-
	Inadequado	1,644 (0,138)**	(1,253 - 2,155)	1,563 (0,124)**	(1,222 - 1,999)
Região	Norte/Nordeste	-	-	-	-
	Demais regiões	1,463 (0,121)*	(1,154 - 1,855)	1,393 (0,117)*	(1,100 - 1,763)
Sexo	Masculino	Não significante		-	-
	Feminino			1,217 (0,104)*	(0,990 - 1,495)
Número de filhos	Primeiro filho	Não significante		-	-
	2 ou 3 filhos			0,748 (0,111)*	(0,602 - 0,929)
	Mais de 3 filhos			0,931 (0,147)	(0,695 - 1,247)

* p-valor ≤ 0,005
** p-valor < 0,001

Quando analisamos os erros padrão dos dois modelos observamos valores diferentes. Com exceção da variável "Número de consultas pré-natal", todas as variáveis apresentaram menor erro padrão no modelo das estimativas combinadas. Comparando os intervalos de

confiança para as variáveis significantes nos dois modelos, o resultado também foi diferente, e novamente, com exceção da variável "Número de consultas pré-natal", todas as demais apresentaram intervalos de mais estreitos para o Modelo 2.

4.3 – Interpretação e Discussão do Modelo Agrupado

Uma vez que o objetivo principal dessa monografia é o estudo dos métodos de imputação, não serão discutidos em detalhes todos os modelos de BPN, bem como seus resultados epidemiológicos e implicações. Selecionamos o modelo das estimativas agrupadas para interpretar o BPN, conforme apresentado na Tabela 7.

Tabela 7: Modelo Logístico Binário do BPN

Variáveis		RC	Erro padrão	P-valor	I.C. 95% (RC)	
					Inferior	Superior
Fumante	Sim	-	-	-	-	-
	Não	0,645	0,134	0,002	0,492	0,845
Sexo	Masculino	-	-	-	-	-
	Feminino	1,217	0,104	0,062	0,990	1,495
Número de consultas pré-natal	< 6 consultas	-	-	-	-	-
	≥ 6 consultas	0,703	0,170	0,054	0491	1,006
Pré-natal	Adequado	-	-	-	-	-
	Inadequado	1,563	0,124	<0,001	1,222	1,999
Região	Norte /Nordeste	-	-	-	-	-
	Demais regiões	1,393	0,117	0,007	1,100	1,763
Número de filhos	Primeiro filho	-	-	-	-	-
	2 ou 3 filhos	0,748	0,111	0,009	0,602	0,929
	Mais de 3 filhos	0,931	0,147	0,628	,695	1,247
Constante		0,118	0,220	<0,001	0,076	0,183

Após o processo de imputação, a prevalência de BPN entre as 6058 crianças estudadas foi de 8,6%. As variáveis que mostraram uma associação significativa à variável resposta, quando controladas pelas demais, foram: "Fumantes", "Sexo", "Número de consultas pré-natal", "Região", "Pré-natal" e "Número de filhos".

Analisando as RC apresentadas na Tabela 6, observamos que a chance de recém-nascidos com peso inferior a 2,500 Kg é 35% menor para as mães que não são fumantes, quando comparada com filhos de mães fumantes. A prevalência de mães fumantes em nossa amostra foi de 15,9%, o que é um valor considerável uma vez que o tabagismo contribui negativamente para a saúde da criança e da mãe.

Quanto ao sexo do recém nascido, o sexo feminino apresentou chance 21,7% maior de BPN comparada ao masculino. Os resultados encontrados na literatura mostram que a

prevalência de BPN, em geral, é maior no sexo masculino (Ribeiro, Guimarães, Lima, Sarinho, & Coutinho, 2009). Entretanto, resultados semelhantes ao nosso também são encontrados na literatura, como por exemplo em Maia & Souza (2010), que afirmam que esta relação pode ser explicada pelo fato de meninas apresentarem peso mais baixo que meninos, para a mesma idade gestacional.

O número de consultas pré-natal realizadas é um importante fator, porque reflete o acesso das mães aos serviços de saúde. O preconizado pelo Ministério da Saúde é realizar um mínimo de 6 consultas pré-natal, e a chance de recém-nascidos com baixo peso é 30% menor quando as mães atendem esse requisito, comparada com nascidos de mães que realizaram menos de seis consultas. Em nossa base de dados, a prevalência de BPN para as mães que realizaram seis ou mais consultas foi de 7,5%, contra 12,6%, mostrando como um pré-natal adequado afeta positivamente a saúde da criança.

Quanto à adequação do pré-natal, a variável foi estatisticamente significante (p-valor < 0,001) e mostra que recém-nascidos de mães que realizaram um pré-natal inadequado tem chance de BPN 56,3% maior, quando comparados com nascidos de mães cujo pré-natal foi adequado. Nesta amostra, a prevalência de BPN em nascidos de mães que tiveram um pré-natal inadequado foi de 11,2% contra 6,4% com pré-natal adequado, mostrando mais uma vez como o pré-natal é um importante fator de prevenção para o BPN e para riscos a saúde materno-infantil.

A região também foi significante no modelo. Crianças nascidas nas regiões Sudeste, Sul e Centro-Oeste apresentaram chance de BPN 39,3% maior que àquelas das regiões Norte e Nordeste. O esperado era ter um resultado contrário, uma vez que as regiões Norte e Nordeste apresentam diferenças demográficas e socioeconômicas, e muitas vezes, carência dos serviços de saúde. As prevalências entre as regiões na nossa amostra foram 7,5% no N e NE e 9,4% nas demais regiões.

E finalmente, o número de filhos também se mostrou significante para a categoria mães de 2 ou 3 filhos (p-valor=0,009). A chance de BPN nessa categoria foi 25% menor, comparada com a chance mães primíparas. Mães ‘de primeira viagem’ não tem tanta maturidade para os cuidados à gestação, enquanto mães de sucessivas gestações não tem o organismo totalmente recuperado para fornecer ao feto os nutrientes necessários. Porém, este último fator tem mais relevância para o intervalo entre os partos, variável que se mostrou estatisticamente significante em outros modelos, considerando as bases imputadas. A

proporção de recém-nascidos com baixo peso na nossa amostra foi de 9,3% para mães primíparas, 7,5% para mães de 2 ou 3 filhos e 10,4% para mães de mais de três filhos.

Apesar de não serem significativas no modelo agrupado, algumas variáveis tiveram associação significativa com os modelos das bases imputadas e têm influência na ocorrência do baixo peso ao nascer. Mulheres solteiras, separadas e/ou viúvas e o tipo de parto ser cesárea, também mostraram ser fatores de risco para o BPN.

O BPN é um importante indicador de saúde, que merece atenção, uma vez que é influenciado por diversos fatores, como sugerem nossos resultados. Os resultados encontrados na literatura não são unânimes quanto a estes fatores, sendo em geral diferentes de acordo com o tipo e local de estudo.

Capítulo 5

Conclusão e Considerações Finais

As variáveis “Fumante”, “Pré-natal” e “Região” foram significantes em todos os modelos, e as diferenças nas estimativas da RC, erro padrão e intervalos de confiança destas variáveis, aliados ao fato dos modelos terem diferentes variáveis explicativas, mostram como o processo de imputação pode influenciar nos resultados e, consequentemente, nas conclusões do pesquisador. Além disso, nossos resultados confirmam a importância do tratamento aos dados e os riscos de se ignorar o problema através da adoção de procedimentos que consideram apenas a análise dos dados completos.

É comum, em especial em estudos epidemiológicos, o pesquisador restringir a análise dos dados apenas aos casos completos, ignorando e deletando aqueles em que ocorreram não resposta. Como é o caso do método *Listwise Deletion*, que deleta todas as informações dos casos com dados faltantes, independente da variável onde ocorreu. Essa abordagem ocasiona dois problemas: considerar que os dados são MCAR, quando na verdade eles não são, e levar o pesquisador a conclusões erradas já que as estimativas podem ser viesadas. Além disso, definir modelos de regressão considerando apenas os casos completos pode fazer com que variáveis deixem de ser significativas no modelo, como aconteceu nesse estudo.

Portanto é importante, principalmente para o profissional de estatística, não ignorar os dados faltantes, mas sim realizar uma análise dos mesmos e imputá-los a partir de métodos apropriados, antes da fase da análise dos dados, dos resultados e das conclusões.

Do mesmo modo, quando ignoramos os dados faltantes e consideramos apenas os casos completos, o tamanho da amostra pode diminuir, enfraquecendo também o poder dos testes estatísticos.

É importante ressaltar que no presente estudo empírico, por se tratar de uma aplicação a uma base com dados faltantes reais, não temos como confirmar a qualidade das estimativas calculadas, diferente de trabalhos teóricos encontrados na literatura, em que os dados faltantes são gerados através de simulações por sorteios aleatórios a partir de uma base de dados completa. Nestes estudos teóricos os pesquisadores têm conhecimento do valor real do dado faltante, diferente do nosso em que o dado faltante de fato aconteceu.

É de conhecimento da autora e dos orientadores que a eficiência das estimativas e o erro atribuído aos valores imputados são aspectos importantes do processo de imputação.

Porém, os mesmos não foram abordados nessa monografia por uma questão de tempo e definições (conceitual), mas serão fins de estudo em trabalhos futuros. Outra questão importante não abordada nesta monografia foi a consideração do desenho amostral nas análises e também na aplicação dos procedimentos de imputação.

Por fim, esperamos que o assunto abordado nesta monografia sirva de incentivo para outros pesquisadores na área da Estatística Aplicada a Epidemiologia e desperte a atenção para a necessidade de se fazer uma investigação dos dados faltantes e, caso necessário, aplicar métodos de imputação adequados antes de realizarem suas análises e conclusões.

Bibliografia

(s.d.). Tratto il giorno Dezembro 12, 2013 da DATASUS: http://tabnet.datasus.gov.br/cgi/idb2000/fqd17.htm

Agresti, A. (2013). *Categorical Data Analysis* (3ª ed.). New York: Wiley-Interscience.

Allison, P. D. (2001). *Missing Data (Quantitative Applications in the Social Sciences)* (1ª ed.). Thousand Oaks, CA: SAGE Publications.

Araújo, D. M. (2012). *Fatores associados ao estado nutricional gestacional e desfechos perinatais em usuários do Sistema Único de Saúde (SUS), em dois municípios do estado do Rio de Janeiro (RJ).* Tese de Doutorado em Ciências na área de Epidemiologia em Saúde Pública, Escola Nacional de Saúde Pública (ENSP) - Fiocruz, Rio de Janeiro.

Araújo, S. G., & Sant'Ana, D. M. (2003). Relação entre a idade materna e o peso ao nascer: um estudo da gravidez na adolescência no município de Umuarama, PR, Brasil em 2001. *Ciência, Cuidado e Saúde*, 155:160.

Assunção, F. (2012). *Estratégias para tratamento de variáveis com dados faltantes durante o desenvolvimento de modelos preditivos.* Dissertação de Mestrado em Ciências, Universidade de São Paulo, Instituto de Matemática e Estatística, São Paulo.

Baracho, S. M. (2003). *Tratamento de Dados Ausentes em Estudos Longitudinais.* Dissertação de Mestrado , UFMG, Departamento de Estatística.

Benicio, M. D., Monteiro, C., Souza, J. P., Castilho, E. A., & Lamonica, I. M. (1985). Análise multivariada de fatores de risco para o baixo peso ao nascer em nascidos vivos do município de São Paulo, SP (Brasil) . *Revista de Saúde Pública*, 311-20.

Bernstein IM, M. J. (2005). Maternal smoking and its association with birth weight. *Obstet Gynecol*, 986-991.

Bisceski, A., Rodrigues, D., Simon, J., Silva, M., Engel, M., Covatti, P., et al. (2012). CARACTERISTICAS EPIDEMIOLÓGICAS DA SAÚDE MATERNO-INFANTIL. *Revista de Enfermagem*, 79-88.

Brasil, M. (2009). *Pesquisa Nacional de Demografia e Saúde da Criança e da Mulher*. Tratto da http://bvsms.saude.gov.br/bvs/pnds/banco_dados.php

Brasil, M. (2009). *Pesquisa Nacional de Demografia e Saúde da Criança e da Mulher*. Tratto da http://bvsms.saude.gov.br/bvs/pnds/banco_dados.php

Brasil, M. (2009). *Pesquisa Nacional de Demografia e Saúde da Criança e da Mulher – PNDS 2006 : dimensões do processo reprodutivo e da saúde da criança/ Ministério da Saúde, Centro Brasileiro de Análise e Planejamento* (1ª ed.). Brasília: Ministério da Saúde.

Caçola, P., & Bobbio, T. (2010). Baixo peso ao nascer e alterações no desenvolvimento motor: a realidade atual. *Revista Paulista de Pediatria*, 70-76.

da Silva, M. (2012). *Imputação múltipla: comparação e eficiência em experimentos multiambientais.* Dissertação de Mestrado em Ciências: Estatística e Experimentação Agronômica, Universidade de São Paulo , Escola Superior de Agricultura Luiz de Queiroz, Piracicaba.

DATASUS. (s.d.). Tratto il giorno Dezembro 2013 da Departamento de Informática do SUS: http://tabnet.datasus.gov.br/cgi/idb2000/fqd17.htm

Enders, C. K. (2010). *Applied missing data analysis.* New York.

Engels, J., & Diehr, P. (2003). Imputation of missing longitudinal data: a comparison of methods. *Journal of Clinical Epidemiology*, p. 968-976.

Faculty & Staff. (2014). *Logistic Regression: Statnotes, from North Carolina State University, Public Administration Program*. Tratto il giorno Maio 2014 da http://faculty.chass.ncsu.edu/: http://faculty.chass.ncsu.edu/garson/PA765/logistic.htm

Figueira, C. (2006). *Modelos de Regressão Logística.* Dissertação de Mestrado em Matemática, Universidade Federal do Rio Grande do Sul, Instituto de Matemática, Porto Alegre.

Folha de São Paulo. (s.d.). Acesso em Dezembro de 2013, disponível em http://www1.folha.uol.com.br/equilibrioesaude/1014267-cresce-numero-de-bebes-nascidos-com-baixo-peso-no-pais.shtml

Franciotti, D. L., Mayer, G. N., & Cancelier, A. L. (2010). Fatores de risco para baixo peso ao nascer: um estudo de caso-controle. *Arquivos Catarinenses de Medicina, 39*(3).

Halpern, R., Schaefer, E., Pereira, A., Arnt, E., Bezerra, J., & Pinto, L. (1996). Fatores de risco para baixo peso ao nascer em uma comunidade rural do sul do Brasil. *Jornal de Pediatria*, p. 369-373.

Indicadores de morbidade e fatores associados. (s.d.). Tratto il giorno Dezembro 12, 2013 da DATASUS: http://tabnet.datasus.gov.br/cgi/idb2000/fqd17.htm

Lippi, U. G., Andrade, A. S., Bertagnon, J. D., & Melo, E. (1989). Fatores obstétricos associados ao baixo peso ao nascer. *Revista de Saúde Pública*, 382-387.

Little, R. (1982, Junho). Models for nonresponse in sample surveys. *Journal os the American Statistical Association* , p. 237-250.

Little, R., & Rubin, D. (1983). On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *The American Statistician*, 218-220.

Maia, R. d., & Souza, J. P. (2010). Fatores associados ao baixo peso ao nascer em município do norte do Brasil. *Revista Brasileira de Crescimento e Desenvolvimento Humano*, 735:744.

Maia, V. O. (2005). *Gravidez na adolescência: variáveis que influenciam peso ao nascer e índice de Apgar dos conceptos.* Dissertação de Mestrado em Tocoginecologia , Universidade de Pernambuco, Faculdade de Ciências Médicas, Recife.

Mariotoni, G. G., & Barros Filho, A. (1998, Abril). A gravidez na adolescência é fator de risco para o baixo peso ao nascer? *Jornal de Pediatria*, p. 107-113.

Mariotoni, G., & Filho, A. (2000). Peso ao nascer e características maternas ao longo de 25 anos na Maternidade de Campinas. *Jornal de Pediatria*, p. 55-64.

Melo, A. M., Kassar, S. B., Lira, P. I., Coutinho, S. B., Eickmann, S. H., & Lima, M. C. (2013). Characteristics and factores associated with health care in children younger than 1 year with very low birth weight. *Jornal de Pediatria*, pp. 75-82.

Meng, X.-L. (2000, Dezembro). Missing Data: Dial M for ??? *Journal of the American Statistical Association*, p. 1325-1330.

Minamisava, R., Barbosa, M., Malagoni, L., & Andraus, L. (2004). Fatores associados ao baixo peso ao nascer no Estado de Goiás. *Revista Eletrônica de Enfermagem*, 336-349.

Nascimento, L., & Gotlieb, S. (2001). Fatores de risco para o baixo peso ao nascer, com base em informações da declaração de nascido vivo em Guaratinguetá, SP, no ano de 1998. *Informe Epidemiológico do SUS*, 113-120.

Nunes, L. N. (2007). *Métodos de imputação de dados aplicados na área da saúde.* Tese de Doutorado em Epidemiologia, UFRGS, Faculdade de Medicina, Porto Alegre.

Nunes, L. N., Klück, M. M., & Fachel, J. G. (2009). Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. *Caderno de Saúde Pública*, 268:278.

Nunes, L. N., Klück, M. M., & Fachel, J. G. (2010). Comparação de métodos de imputação única e múltipla usando como exemplo um modelo de risco para mortalidade cirúrgica. *Revista Brasileira de Epidemiologia*, 596-606.

Pagano, M., & Gauvreau, K. (2004). *Princípios de Bioestatística* (2ª ed.). (L. S. Paiva, Trad.) São Paulo: Thomson Learning.

Paula, H. A., Salvador, B. C., Barbosa, L., & Cotta, R. M. (2011). Peso ao Nascer e Variáveis Maternas no Âmbito da Promoção da Saúde. *APS*, 67-74.

Ribeiro, A., Guimarães, M., Lima, M., Sarinho, S., & Coutinho, S. (2009). Fatores de risco para a mortalidade neonatal em crianças com baixo peso ao nascer. *Revista de Saúde Pública*, 246-255.

Robert M. Groves, Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, & Roger Tourangeau. (2009). *Survey Methodology* (2ª ed.). John Wiley & Sons.

Rojas, P. B., Francisco, C. C., Siqueira, L. M., & Carminatti, A. d. (2012). Fatores modificáveis associados ao baixo peso ao nascer da gravidez na adolescência. *Arquivos Catarinenses de Medicina*, 64-69.

Rubin, D. (1987). Multiple Imputation for Nonresponse in Surveys. *New York: John Wiley & Sons*.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 581-592.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data.* Londres: CRC Press.

Schafer, J. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 3-15.

Schafer, J. L., & Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods*, p. 147-177.

Sclowitz, I. K. (2007). *Fatores de risco para a recorrência de baixo peso ao nascer em sucessivas gestações.* Tese de Doutorodo em Epidemiologia, Universidade Federal de Pelotas, Pelotas.

Silva, J. (2011). *Métodos de imputação múltipla para GEE em estudos longitudinais.* Dissertação de Pós-graduação em Estatística, Universidade Federal de Minas Gerais, Ciência Exatas, Belo Horizonte.

Surita, F. G., Suarez, M. B., Siani, S., & Silva, J. P. (2011). Fatores associados ao baixo peso ao nascimento entre adolescentes no Sudeste do Brasil. *Revista Brasileira de Ginecologia e Obstetrícia*, 286-291.

Uchimura, T. T., Pelissari, D. M., & Uchimura, N. S. (2008). Baixo peso ao nascer e fatores associados. *Revista Gaúcha de Enfermagem*, 29:33.

Uchimura, T., Pelissari, D., Soares, D., Uchimura, N., Santana, R., & Moraes, C. (2007). Fatores de risco para o baixo peso ao nascer segundo as variáveis da mãe e do recém-nascido, em Maringá-PR, no período de 1996 a 2002. *Ciência, Cuidade e Saúde*, 51-58.

Vieira, Marcel de Toledo. (2011). Material didático da disciplina de Amostragem I e II. Universidade Federal de Juiz de Fora, Departamento de Estatística, Juiz de Fora.

Veroneze, R. (2011). *Tratamento de dados faltantes empregando biclusterização com imputaçã múltipla.* Dissertação de Mestrado em Engenharia Elétrica, Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação, Campinas.

Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of statistical data editing and imputation.* New Jersey: John Wiley & Sons.

Anexo 1

Estatísticas descritivas das imputações

Tabela 1: Estatísticas descritivas da variável Peso

Dados		N	Média	Desvio padrão	Mínimo	Máximo
Dados originais		5504	3,24	28,78	0,50	5,50
Valores imputados	1	554	3,17	27,04	0,66	5,49
	2	554	3,19	26,59	0,51	5,43
	3	554	3,21	26,65	0,52	5,43
	4	554	3,21	28,35	0,83	5,47
	5	554	3,21	25,99	0,59	5,33
Dados preenchidos após a imputação	1	6058	3,23	28,64	0,50	5,50
	2	6058	3,24	28,59	0,50	5,50
	3	6058	3,24	28,59	0,50	5,50
	4	6058	3,24	28,74	0,50	5,50
	5	6058	3,24	28,54	0,50	5,50

Tabela 2: Estatísticas descritivas da variável “Cesárea”

Dados			N	Porcentagem
Dados originais		Sim	2387	39,6
		Não	3635	60,4
Valores imputados	1	Sim	11	30,6
		Não	25	69,4
	2	Sim	10	27,8
		Não	26	72,2
	3	Sim	12	33,3
		Não	24	66,7
	4	Sim	14	38,9
		Não	22	61,1
	5	Sim	15	41,7
		Não	21	58,3
Preencher dados após imputação	1	Sim	2398	39,6
		Não	3660	60,4
	2	Sim	2397	39,6
		Não	3661	60,4
	3	Sim	2399	39,6
		Não	3659	60,4
	4	Sim	2401	39,6
		Não	3657	60,4
	5	Sim	2402	39,7
		Não	3656	60,3

Tabela 3: Estatísticas descritivas da variável “Cor”

Dados			N	Porcentagem
Dados originais		Branca	2044	34,1
		Preta	608	10,2
		Parda	3017	50,4
		Amarela	166	2,8
		Indígena	151	2,5
Valores imputados	1	Branca	30	41,7
		Preta	5	6,9
		Parda	33	45,8
		Amarela	2	2,8
		Indígena	2	2,8
	2	Branca	26	36,1
		Preta	8	11,1
		Parda	34	47,2
		Amarela	1	1,4
		Indígena	3	4,2
	3	Branca	27	37,5
		Preta	7	9,7
		Parda	33	45,8
		Amarela	2	2,8
		Indígena	3	4,2
	4	Branca	27	37,5
		Preta	7	9,7
		Parda	32	44,4
		Amarela	5	6,9
		Indígena	1	1,4
	5	Branca	27	37,5
		Preta	7	9,7
		Parda	33	45,8
		Amarela	5	6,9
Dados preenchidos após a imputação	1	Branca	2074	34,2
		Preta	613	10,1
		Parda	3050	50,3
		Amarela	168	2,8
		Indígena	153	2,5
	2	Branca	2070	34,2
		Preta	616	10,2
		Parda	3051	50,4
		Amarela	167	2,8
		Indígena	154	2,5
	3	Branca	2071	34,2
		Preta	615	10,2
		Parda	3050	50,3
		Amarela	168	2,8
		Indígena	154	2,5
	4	Branca	2071	34,2
		Preta	615	10,2
		Parda	3049	50,3
		Amarela	171	2,8
		Indígena	152	2,5
	5	Branca	2071	34,2
		Preta	615	10,2
		Parda	3050	50,3
		Amarela	171	2,8
		Indígena	151	2,5

Tabela 4: Estatísticas descritivas da variável “Intervalo”

Dados		N	Média	Desvio padrão	Mínimo	Máximo
Dados originais		6032	33,58	1938,78	9,00	266,00
Valores imputados	1	26	48,22	1542,44	10,79	128,15
	2	26	48,00	1748,30	9,47	118,44
	3	26	50,80	1627,71	9,51	245,40
	4	26	49,77	1544,51	10,40	226,52
	5	26	57,48	1645,95	12,03	226,42
Preencher dados após imputação	1	6058	33,65	1937,92	9,00	266,00
	2	6058	33,65	1938,62	9,00	266,00
	3	6058	33,67	1938,51	9,00	266,00
	4	6058	33,66	1938,10	9,00	266,00
	5	6058	33,70	1939,57	9,00	266,00

Tabela 5: Estatísticas descritivas da variável “Plano de saúde”

Dados			N	Porcentagem
Dados originais		Sim	945	15,6
		Não	5106	84,4
Valores imputados	1	Sim	2	28,6
		Não	5	71,4
	2	Sim	1	14,3
		Não	6	85,7
	3	Sim	3	42,9
		Não	4	57,1
	4	Sim	2	28,6
		Não	5	71,4
	5	Não	7	100,0
Preencher dados após imputação	1	Sim	947	15,6
		Não	5111	84,4
	2	Sim	946	15,6
		Não	5112	84,4
	3	Sim	948	15,6
		Não	5110	84,4
	4	Sim	947	15,6
		Não	5111	84,4
	5	Sim	945	15,6
		Não	5113	84,4

Tabela 6: Estatísticas descritivas da variável "Meses de gravidez na última consulta pré-natal"

Dados		N	Média	Desvio padrão	Mínimo	Máximo
Dados originais		5757	8,60	41,02	1,00	10,00
Valores imputados	1	301	8,31	38,95	1,86	9,98
	2	301	8,25	35,92	1,18	9,99
	3	301	8,30	38,17	2,27	9,99
	4	301	8,37	37,66	1,59	9,94
	5	301	8,27	40,23	2,02	9,95
Preencher dados após imputação	1	6058	8,59	40,99	1,00	10,00
	2	6058	8,59	40,88	1,00	10,00
	3	6058	8,59	40,96	1,00	10,00
	4	6058	8,59	40,90	1,00	10,00
	5	6058	8,59	41,07	1,00	10,00

Tabela 7: Estatísticas descritivas da variável "Quantas consultas pré-natal realizou"

Dados		N	Média	Desvio padrão	Mínimo	Máximo
Dados originais		5533	7,87	179,53	1,00	32,00
Valores imputados	1	525	7,80	137,33	1,01	28,93
	2	525	7,20	146,35	1,02	28,46
	3	525	7,49	149,23	1,03	29,59
	4	525	7,67	142,09	1,00	28,93
	5	525	7,61	136,58	1,04	26,88
Preencher dados após imputação	1	6058	7,87	176,26	1,00	32,00
	2	6058	7,83	177,11	1,00	32,00
	3	6058	7,85	177,17	1,00	32,00
	4	6058	7,86	176,61	1,00	32,00
	5	6058	7,86	176,24	1,00	32,00

Tabela 8: Estatísticas descritivas da variável "Meses de gravidez na 1ª consultas pré-natal"

Dados		N	Média	Desvio padrão	Mínimo	Máximo
Dados originais		5754	2,24	77,91	0,00	9,00
Valores imputados	1	304	2,41	56,78	,00	9,00
	2	304	2,96	62,03	,02	8,85
	3	304	2,79	56,96	,07	8,71
	4	304	2,74	55,66	,02	8,88
	5	304	2,75	57,17	,02	8,89
Preencher dados após imputação	1	6058	2,24	77,00	0,00	9,00
	2	6058	2,26	77,46	0,00	9,00
	3	6058	2,25	77,15	0,00	9,00
	4	6058	2,25	77,08	0,00	9,00
	5	6058	2,25	77,14	0,00	9,00

Tabela 9: Estatísticas descritivas da variável “Tomou injeção contra tétano”

Dados			N	Porcentagem
Dados originais		Sim	4185	71,6
		Não	1656	28,4
Valores imputados	1	Sim	153	70,5
		Não	64	29,5
	2	Sim	132	60,8
		Não	85	39,2
	3	Sim	142	65,4
		Não	75	34,6
	4	Sim	147	67,7
		Não	70	32,3
	5	Sim	155	71,4
		Não	62	28,6
Preencher dados após imputação	1	Sim	4338	71,6
		Não	1720	28,4
	2	Sim	4317	71,3
		Não	1741	28,7
	3	Sim	4327	71,4
		Não	1731	28,6
	4	Sim	4332	71,5
		Não	1726	28,5
	5	Sim	4340	71,6
		Não	1718	28,4

Tabela 10: Estatísticas descritivas da variável “Pré-natal”

Dados			N	Porcentagem
Dados originais		Adequado	3016	54,9
		Inadequado	2477	45,1
Valores imputados	1	Adequado	213	37,7
		Inadequado	352	62,3
	2	Adequado	186	32,9
		Inadequado	379	67,1
	3	Adequado	189	33,5
		Inadequado	376	66,5
	4	Adequado	213	37,7
		Inadequado	352	62,3
	5	Adequado	214	37,9
		Inadequado	351	62,1
Preencher dados após imputação	1	Adequado	3229	53,3
		Inadequado	2829	46,7
	2	Adequado	3202	52,9
		Inadequado	2856	47,1
	3	Adequado	3205	52,9
		Inadequado	2853	47,1
	4	Adequado	3229	53,3
		Inadequado	2829	46,7
	5	Adequado	3230	53,3
		Inadequado	2828	46,7

Anexo 2

Tabelas das variáveis categóricas após a imputação

Tabela 1: Distribuição da variável "Peso ao nascer" após a imputação

Número de imputação		N	%
Dados originais	Peso adequado	5101	84,2
	Baixo peso aos nascer	403	6,7
	Dados faltantes	554	9,1
	Total	6058	100,0
1	Peso adequado	5533	91,3
	Baixo peso aos nascer	525	8,7
2	Peso adequado	5537	91,4
	Baixo peso aos nascer	521	8,6
3	Peso adequado	5533	91,3
	Baixo peso aos nascer	525	8,7
4	Peso adequado	5525	91,2
	Baixo peso aos nascer	533	8,8
5	Peso adequado	5550	91,6
	Baixo peso aos nascer	508	8,4
Agrupado	Peso adequado	5536	91,38
	Baixo peso aos nascer	522	8,62
	Total	6058	100,0

Tabela 2: Distribuição da variável "Cor/raça" após a imputação

Número de imputação		N	%
Dados originais	Branca	2044	33,7
	Não branca	3948	65,2
	Dados faltantes	66	1,1
	Total	6058	100,0
1	Branca	2074	34,2
	Não branca	3984	65,8
2	Branca	2070	34,2
	Não branca	3988	65,8
3	Branca	2071	34,2
	Não branca	3987	65,8
4	Branca	2071	34,2
	Não branca	3987	65,8
5	Branca	2071	34,2
	Não branca	3987	65,8
Agrupado	Branca	2071	34,2
	Não branca	3987	65,8
	Total	6058	100,0

Tabela 3: Distribuição da variável “Cesárea” após a imputação

Número de imputação		N	%
Dados originais	Sim	2387	39,4
	Não	3635	60,0
	Dados faltantes	36	0,6
	Total	6058	100,0
1	Sim	2398	39,6
	Não	3660	60,4
2	Sim	2397	39,6
	Não	3661	60,4
3	Sim	2399	39,6
	Não	3659	60,4
4	Sim	2401	39,6
	Não	3657	60,4
5	Sim	2402	39,7
	Não	3656	60,3
Agrupado	Sim	2399	39,6
	Não	3659	60,4
	Total	6058	100,0

Tabela 4: Distribuição da variável “Intervalo entre os partos” após a imputação

Número de imputação		N	%
Dados originais	Primeiro filho	2319	38,3
	Intervalo curto	447	7,4
	Intervalo longo	3266	53,9
	Dados faltantes	26	0,4
	Total	6058	100,0
1	Primeiro filho	2319	38,3
	Intervalo curto	450	7,4
	Intervalo longo	3289	54,3
2	Primeiro filho	2319	38,3
	Intervalo curto	453	7,5
	Intervalo longo	3286	54,2
3	Primeiro filho	2319	38,3
	Intervalo curto	449	7,4
	Intervalo longo	3290	54,3
4	Primeiro filho	2319	38,3
	Intervalo curto	451	7,4
	Intervalo longo	3288	54,3
5	Primeiro filho	2319	38,3
	Intervalo curto	452	7,5
	Intervalo longo	3287	54,3
Agrupado	Primeiro filho	2319	38,3
	Intervalo curto	451,0	7,4
	Intervalo longo	3288	54,3
	Total	6058	100,0

Tabela 5: Distribuição da variável “Plano de saúde” após a imputação

Número de imputação		N	%
Dados originais	Sim	945	15,6
	Não	5106	84,3
	Dados faltantes	7	0,1
	Total	6058	100,0
1	Sim	947	15,6
	Não	5111	84,4
2	Sim	946	15,6
	Não	5112	84,4
3	Sim	948	15,6
	Não	5110	84,4
4	Sim	947	15,6
	Não	5111	84,4
5	Sim	945	15,6
	Não	5113	84,4
Agrupado	Sim	946,6	15,6
	Não	5111,4	84,4
	Total	6058	100,0

Tabela 6: Distribuição da variável “Número de consultas pré-natal” após a imputação

Número de imputação		N	%
Dados originais	Menos de 6 consultas	1162	19,2
	6 ou mais consultas	4371	72,2
	Dados faltantes	525	8,7
	Total	6058	100,0
1	Menos de 6 consultas	1329	21,9
	6 ou mais consultas	4729	78,1
2	Menos de 6 consultas	1350	22,3
	6 ou mais consultas	4708	77,7
3	Menos de 6 consultas	1342	22,2
	6 ou mais consultas	4716	77,8
4	Menos de 6 consultas	1323	21,8
	6 ou mais consultas	4735	78,2
5	Menos de 6 consultas	1340	22,1
	6 ou mais consultas	4718	77,9
Agrupado	Menos de 6 consultas	1337	22,1
	6 ou mais consultas	4721	77,9
	Total	6058	100,0

Tabela 7: Distribuição da variável “Meses de gravidez no 1º pré-natal” após a imputação

Número de imputação		N	%
Dados originais	Até 3 meses	4742	78,3
	4 meses ou mais	1012	16,7
	Dados faltantes	304	5,0
	Total	6058	100,0
1	Até 3 meses	4912	81,1
	4 meses ou mais	1146	18,9
2	Até 3 meses	4886	80,7
	4 meses ou mais	1114	19,3
3	Até 3 meses	4880	80,6
	4 meses ou mais	1178	19,4
4	Até 3 meses	4905	81,0
	4 meses ou mais	1153	19,0
5	Até 3 meses	4903	80,9
	4 meses ou mais	1155	19,1
Agrupado	Até 3 meses	4897	80,8
	4 meses ou mais	1161	19,2
	Total	6058	100,0

Tabela 8: Distribuição da variável “Injeção antitetânica” após a imputação

Número de imputação		N	%
Dados originais	Sim	4185	69,1
	Não	1656	27,3
	Dados faltantes	217	3,6
	Total	6058	100,0
1	Sim	4338	71,6
	Não	1720	28,4
2	Sim	4317	71,3
	Não	1741	28,7
3	Sim	4327	71,4
	Não	1731	28,6
4	Sim	4332	71,5
	Não	1726	28,5
5	Sim	4340	71,6
	Não	1718	28,4
Agrupado	Sim	4331	71,5
	Não	1727	28,5
	Total	6058	100,0

Tabela 9: Distribuição da variável “Pré-natal” após a imputação

Número de imputação		N	%
Dados originais	Adequado	3016	49,8
	Inadequado	2477	40,9
	Dados faltantes	565	9,3
	Total	6058	100,0
1	Adequado	3229	53,3
	Inadequado	2829	46,7
2	Adequado	3202	52,9
	Inadequado	2856	47,1
3	Adequado	3205	52,9
	Inadequado	2853	47,1
4	Adequado	3229	53,3
	Inadequado	2829	46,7
5	Adequado	3230	53,3
	Inadequado	2828	46,7
Agrupado	Adequado	3219	53,1
	Inadequado	2839	46,9
	Total	6058	100,0

Anexo 3

Modelos de Regressão Logística

Tabela 1: Modelo de Regressão Logística para a base de dados originais

Variáveis		OR	Erro padrão	P-valor	I.C. 95% (OR)	
					Inferior	Superior
Fumante	Sim	-	-	-	-	-
	Não	0,651	0,143	0,003	0,492	0,862
Intervalo interpartal	Primeiro filho	-	-	<0,001	-	-
	Intervalo curto	1,083	0,206	0,699	0,723	1,623
	Intervalo longo	0,635	0,120	<0,001	0,502	0,803
Número de consultas pré-natal	< 6 consultas	-	-	-	-	-
	≥ 6 consultas	0,727	0,152	0,035	0,540	0,978
Pré-natal	Adequado	-	-	-	-	-
	Inadequado	1,644	0,138	<0,001	1,253	2,155
Região	Norte e Nordeste	-	-	-	-	-
	Demais regiões	1,463	0,121	0,002	1,154	1,855
Constante		0,104	0,225	<0,001	-	-

Tabela 2: Modelo de Regressão Logística para a base de dados imputados *m*=1

Variáveis		OR	Erro padrão	P-valor	I.C. 95% (OR)	
					Inferior	Superior
Fumante	Sim	-	-	-	-	-
	Não	0,601	0,112	<0,001	0,483	0,749
Sexo	Masculino	-	-	-	-	-
	Feminino	1,205	0,093	0,045	1,004	1,445
Intervalo interpartal	Primeiro filho	-	-	0,006	-	-
	Intervalo curto	1,176	0,166	0,328	0,850	1,627
	Inervalo longo	0,779	0,099	0,012	0,642	0,947
Número de consultas pré-natal	< 6 consultas	-	-	-	-	-
	≥ 6 consultas	0,730	0,122	0,010	0,575	0,928
Pré-natal	Adequado	-	-	-	-	-
	Inadequado	1,571	0,112	<0,001	1,260	1,958
Região	Norte e Nordeste	-	-	-	-	-
	Demais regiões	1,315	0,098	0,005	1,086	1,593
Constante		0,124	0,188	<0,001	-	-

Tabela 3: Modelo de Regressão Logística para a base de dados imputados *m*=2

Variáveis		OR	Erro padrão	P-valor	I.C. 95% (OR)	
					Inferior	Superior
Estado conjugal	Em união/Casada	-	-	-	-	-
	Sem união/Separada/Viúva	1,323	0,119	0,019	1,047	1,672
Fumante	Sim	-	-	-	-	-
	Não	0,728	0,117	0,007	0,579	0,917
Sexo	Masculino	-	-	-	-	-
	Feminino	1,242	0,094	0,021	1,033	1,494
Número de consultas pré-natal	< 6 consultas	-	-	-	-	-
	≥ 6 consultas	0,677	0,123	0,002	0,532	0,863
Pré-natal	Adequado	-	-	-	-	-
	Inadequado	1,520	0,114	<0,001	1,216	1,901
Região	Norte e Nordeste	-	-	-	-	-
	Demais regiões	1,414	0,100	0,001	1,162	1,719
Constante		0,088	0,187	<0,001	-	-

Tabela 4: Modelo de Regressão Logística para a base de dados imputados *m*=3

Variáveis		OR	Erro padrão	P-valor	I.C. 95% (OR)	
					Inferior	Superior
Estado conjugal	Em união/Casada	-	-	-	-	-
	Sem união/Separada/Viúva	1,303	0,122	0,030	1,026	1,654
Fumante	Sim	-	-	-	-	-
	Não	0,696	0,117	0,002	0,553	0,876
Cesárea	Sim	-	-	-	-	-
	Não	0,799	0,098	0,022	0,659	0,967
Intervalo interpartal	Primeiro filho	-	-	0,022	-	-
	Intervalo curto	1,124	0,173	0,500	0,801	1,577
	Inervalo longo	0,794	0,101	0,022	0,652	0,967
Número de consultas pré-natal	< 6 consultas	-	-	-	-	-
	≥ 6 consultas	0,676	0,124	0,002	0,530	0,863
Pré-natal	Adequado	-	-	-	-	-
	Inadequado	1,527	0,114	<0,001	1,220	1,911
Região	Norte e Nordeste	-	-	-	-	-
	Demais regiões	1,478	0,102	<0,001	1,210	1,805
Situação do domicílio	Urbano	-	-	-	-	-
	Rural	1,206	0,098	0,055	0,996	1,462
Constante		0,119	0,208	<0,001	-	-

Tabela 5: Modelo de Regressão Logística para a base de dados imputados *m*=4

Variáveis		OR	Erro padrão	P-valor	I.C. 95% (OR)	
					Inferior	Superior
Fumante	Sim	-	-	-	-	-
	Não	0,627	0,113	<0,001	0,502	0,782
Pré-natal	Adequado	-	-	-	-	-
	Inadequado	1,816	0,095	<0,001	1,507	2,188
Região	Norte e Nordeste	-	-	-	-	-
	Demais regiões	1,313	0,097	0,005	1,085	1,588
Número de filhos	Primeiro filho	-	-	0,009	-	-
	2 ou 3 filhos	0,753	0,108	0,009	0,609	0,932
	Mais de 3 filhos	1,007	0,129	0,959	0,781	1,297
Constante		0,098	0,153	<0,001	-	-

Tabela 6: Modelo de Regressão Logística para a base de dados imputados *m*=5

Variáveis		OR	Erro padrão	P-valor	I.C. 95% (OR)	
					Inferior	Superior
Fumante	Sim	-	-	-	-	-
	Não	0,617	0,115	<0,001	0,492	0,773
Sexo	Masculino	-	-	-	-	-
	Feminino	1,273	0,095	0,011	1,056	1,535
Intervalo interpartal	Primeiro filho	-	-	0,016	-	-
	Intervalo curto	1,170	0,172	0,361	0,835	1,638
	Inervalo longo	0,796	0,102	0,025	0,652	0,972
Número de consultas pré-natal	< 6 consultas	-	-	-	-	-
	≥ 6 consultas	0,545	0,132	<0,001	0,421	0,705
Meses de gravidez no 1º pré-natal	≤ 3 meses	-	-	-	-	-
	> 3 meses	0,709	0,136	0,012	0,543	0,926
Pré-natal	Adequado	-	-	-	-	-
	Inadequado	1,569	0,120	<0,001	1,240	1,986
Região	Norte e Nordeste	-	-	-	-	-
	Demais regiões	1,434	0,101	<0,001	1,176	1,750
Constante		0,139	0,199	<0,001	-	-

Tabela 7: Modelo de Regressão Logística para as estimativas combinadas

Variáveis		OR	Erro padrão	P-valor	I.C. 95% (OR)	
					Inferior	Superior
Fumante	Sim	-	-	-	-	-
	Não	0,645	0,134	0,002	0,492	0,845
Sexo	Masculino	-	-	-	-	-
	Feminino	1,217	0,104	0,062	0,990	1,495
Número de consultas pré-natal	< 6 consultas	-	-	-	-	-
	≥ 6 consultas	0,703	0,170	0,054	0,491	1,006
Pré-natal	Adequado	-	-	-	-	-
	Inadequado	1,563	0,124	<0,001	1,222	1,999
Região	Norte e Nordeste	-	-	-	-	-
	Demais regiões	1,393	0,117	0,007	1,100	1,763
Número de filhos	Primeiro filho	-	-	-	-	-
	2 ou 3 filhos	0,748	0,111	0,009	0,602	0,929
	Mais de 3 filhos	0,931	0,147	0,628	0,695	1,247
Constante		0,118	0,220	<0,001	0,076	0,183