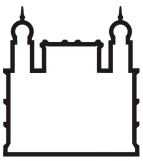


Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

ESCOLA NACIONAL DE SAÚDE PÚBLICA
SERGIO AROUCA
ENSP

Vanessa Eufrauzino Pacheco

Comparação de Métodos para Tratamento de Dados Faltantes em Inquéritos Epidemiológicos com Amostragem Complexa

Rio de Janeiro
2018

Vanessa Eufrauzino Pacheco

Comparação de Métodos para Tratamento de Dados Faltantes em Inquéritos Epidemiológicos com Amostragem Complexa

Dissertação apresentada ao Programa de Pós-graduação em Epidemiologia em Saúde Pública, da Escola Nacional de Saúde Pública Sérgio Arouca, na Fundação Oswaldo Cruz, como requisito parcial para obtenção do título de Mestre em Ciências. Área de concentração: Métodos Quantitativos em Epidemiologia e Saúde Pública

Orientador: Prof. Cleber Nascimento do Carmo

Coorientadora: Prof. Ana Paula Esteves Pereira

Rio de Janeiro

2018

Catalogação na fonte
Fundação Oswaldo Cruz
Instituto de Comunicação e Informação Científica e Tecnológica
Biblioteca de Saúde Pública

P116c Pacheco, Vanessa Eufrauzino
Comparação de métodos para tratamento de dados faltantes em inquéritos epidemiológicos com amostragem complexa. / Vanessa Eufrauzino Pacheco. -- 2018.
83 f. : graf.; tab.

Orientadores: Cleber Nascimento do Carmo e Ana Paula Esteves Pereira.
Dissertação (Mestrado) – Fundação Oswaldo Cruz, Escola Nacional de Saúde Pública Sergio Arouca, Rio de Janeiro, 2018.

1. Coleta de Dados. 2. Inquéritos Epidemiológicos. 3. Dados Faltantes. 4. Modelos Logísticos. 5. Distribuição Aleatória. 6. Epidemiologia Descritiva. I. Título.

CDD – 22.ed. – 001.433

Vanessa Eufrauzino Pacheco

Comparação de Métodos para Tratamento de Dados Faltantes em Inquéritos Epidemiológicos com Amostragem Complexa

Dissertação apresentada ao Programa de Pós-graduação em Epidemiologia em Saúde Pública, da Escola Nacional de Saúde Pública Sérgio Arouca, na Fundação Oswaldo Cruz, como requisito parcial para obtenção do título de Mestre em Ciências. Área de concentração: Métodos Quantitativos em Epidemiologia e Saúde Pública

Aprovada em: 10 de abril de 2018.

Banca Examinadora

Profa. Dra. Ludmilla da Silva Viana Jacobson
IME-Universidade Federal Fluminense

Profa.Dra.Célia Landmann Szwarcwald
ICICT-Fiocruz

Prof. Dr. Cleber Nascimento do Carmo(Orientador)
ENSP-Fiocruz

Rio de Janeiro

2018

Agradecimentos

São muitas pessoas que entram e saem de nossas vidas, algumas para nos deixar com um bom aprendizado e outras apenas para perturbar mesmo.

Para aquelas que ficam, se misturam com a nossa essência e fazem de nossa vida mais que um dia após o outro, uma experiência ímpar gostaria de deixar bem explícito meus mais calorosos agradecimentos.

A Deus que sem Ele não sou nada, que me deu forças pra sempre continuar, mesmo quando estava muito difícil.

Ao orientador Cleber, que entrou nesse projeto cego, acreditou e aceitou de bom grado e com carinho e a coorientadora Ana Paula por me deixar com os nervos a flor da pele na reta final para que o trabalho saísse maravilhoso.

A mamãe Adalgisa e a Karen e Karine, por entenderem os motivos de minhas muitas ausências, a toda ajuda que me deram sem entender um terço do que eu queria fazer, por todo apoio que me deram e me dão até hoje para concretizar meus sonhos acadêmicos. Muito obrigada, vocês são parte da razão do meu viver.

A minha querida Rebecca pela melhor função de 8 linhas implementada no R, além de todos os empurrões para tudo.

Ao Leandro, que existiu na minha vida por pouco tempo, mas é de enorme importância para a minha sanidade mental.

Ao grupo de pesquisa saúde materno-infantil com ênfase às professoras Duca e Silvana que diariamente me dão aula de como ser uma boa profissional, como ser uma boa epidemiologista além de terem sido grandes incentivadoras para eu fazer o mestrado.

Não menos importante nessa trajetória estão Márcia, Mônica, ops: Elaine, Bárbara e todos os amigos que ajudam a diminuir a tensão do dia a dia.

Muito obrigada a todos!

RESUMO

Inquéritos epidemiológicos com amostragem complexa são bastante utilizados devido a redução de custo propiciando o mesmo benefício que uma pesquisa censitária. Entretanto, a ocorrência de perda de dados é um dos problemas que podem afetar esses inquéritos, influenciando os resultados analíticos da pesquisa. A razão para um dado ser faltante é chamado de mecanismo de dados faltantes, definido em três categorias: perda completamente aleatória, perda não completamente aleatória e perda não aleatória. Esse trabalho descreve esses mecanismos e aponta algumas técnicas de tratamento de dados que podem ser aplicadas em uma amostragem complexa, considerando predominantemente desfechos categóricos. A partir do banco de dados Nascer no Brasil, foram simulados bancos de dados com os três tipos de mecanismos de perda e, para o tratamento dos dados, foram utilizados os métodos de análise de casos completos, método do vizinho mais próximo, imputação múltipla por média preditiva e imputação por escore de propensão. Para a comparação dos resultados foram observadas as taxas de recuperação de dados de maneira exata aos originais e diferença quadrática de estimativas de parâmetros de regressão logística e linear. Os métodos de imputação trouxeram mais de 50% dos dados recuperados de maneira exata para os mecanismos perda completamente aleatória e perda não aleatória, já para o mecanismo perda não completamente aleatória, a recuperação foi de aproximadamente 30%. Para as diferenças quadráticas os métodos do vizinho mais próximo e imputação múltipla tiveram resultados equiparáveis.O presente estudo ressaltou a importância da escolha adequada de métodos de imputação para desfechos categóricos e de variáveis para predição de valores, assim como demonstrou diferenças importantes observadas de acordo com o mecanismo de perda utilizado.

.

Palavras-chave: Amostra complexa, Dados faltantes, Escore de propensão, Imputação múltipla , Vizinho mais próximo

ABSTRACT

Epidemiological surveys with complex sampling are widely used because of cost reduction, providing the same benefit as a census survey. However, the occurrence of data loss is one of the problems that can affect these surveys, influencing the analytical results of the research. The reason for a missing data is called the missing data mechanism, defined in three categories: completely random loss, not completely random loss, and non-random loss. This work describes these mechanisms and points out some techniques of data treatment that can be applied in a complex sampling, considering predominantly categorical outcomes. Databases with the three types of loss mechanisms were simulated from the Born in Brazil database and, for the treatment of the data, we used the methods of analysis of complete cases, method of the nearest neighbor, multiple imputation by predictive mean and imputation by propensity score. In order to compare the results, the data recovery rates were observed in an exact manner to the originals and the quadratic difference of logistic and linear regression parameters estimates. The imputation methods brought more than 50 % of the exact recovered data to the mechanisms completely random loss and non-random loss, whereas for the mechanism not completely random loss, the recovery was approximately 30 %. For the quadratic differences, the methods of the closest neighbor and multiple imputation had similar results. The present study emphasized the importance of the adequate choice of imputation methods for categorical outcomes and variables for predicting values, as well as showing important differences observed according to the mechanism of loss used.

Key-words: Complex sample, missing data, Propensity score, Multiple imputation, Nearest neighbor.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

PCA	Perda Completamente Aleatória
PNCA	Perda Não Completamente Aleatória
PNA	Perda Não Aleatória

LISTA DE SÍMBOLOS

β	Letra grega Beta
θ	Letra grega Theta
σ	Letra grega Sigma
Σ	Letra grega maiúscula Sigma que representa somatório
ψ	Letra grega Psi
τ	Letra grega Tau

Sumário

1 INTRODUÇÃO

Um dos objetivos de estudos epidemiológicos é observar quais fatores, naturais ou não, interferem em algum agravo de saúde. Inquéritos de saúde utilizam amostragem complexa com frequência, pois permitem avaliar a saúde de indivíduos com um custo reduzido, menor tempo; além de possibilitar a generalização dos resultados. (SZWARCWALD; DAMACENA, 2008; NUNES, KLÜCK; FACHEL, 2010).

Embora vantajosos para uma avaliação mais rápida, os inquéritos podem trazer um inconveniente que nem sempre é tratado de maneira adequada: a perda ou ausência de respostas. Problemas que acarretam a perda de dados podem ocorrer em qualquer momento da pesquisa, seja na hora da confecção dos instrumentos de coletas, na falta de treinamento adequado aos entrevistadores ou com a recusa do participante em responder a um questionário ou algumas perguntas específicas (RUBIN, 1976; MACIEL,2012). A perda de qualquer informação pode direcionar o efeito de associações incorretamente (REZVAN; LEE; SIMPSON, 2015).

A fim de minimizar parte dessa complicação é comum a restrição do banco de dados aos indivíduos com todas as observações preenchidas. Essa prática, porém, pode acarretar em perda de precisão nas estimativas pontuais, já que a amostra é reduzida, ou em estimativas viesadas, já que os dados restritos podem ser de indivíduos com perfis diferentes dos que permaneceram na amostra (CAMARGOS et al., 2011; MACIEL, 2012; NUNES, KLÜCK; FACHEL, 2010).

Deste modo, torna-se importante a criação de estratégias com o intuito de minimizar efeitos causados por dados faltantes, seja durante o planejamento, com esforços para evitar perda de informações, ou no momento da análise, abordando técnicas estatísticas adequadas para os dados faltantes (BELL, KROMREY; FERRON, 2009).

O presente estudo visa apresentar e discutir alguns métodos de imputação de dados faltantes em estudos epidemiológicos com amostragem complexa, com aplicação em um caso real.

2 AMOSTRAGEM COMPLEXA

A amostragem é uma técnica de seleção de indivíduos ou unidades de uma população para estimar suas propriedades e características quando não é possível a realização de um censo. A utilização de uma amostra e não a população deve-se à pelo menos uma das seguintes razões: população infinita; redução de gasto, economia de tempo, facilidade (BOLFARINE; BUSSAB, 1994).

Um dos primeiros passos na concepção de um inquérito é o planejamento amostral. Existe um número significativo de métodos para a seleção de uma amostra. Definir qual utilizar para que a população seja devidamente representada depende de um conhecimento a priori dos indivíduos que se deseja observar (SZWARCWALD; DAMACENA, 2008).

A amostragem pode ser classificada como determinística: quando não são utilizados mecanismos aleatórios, ou probabilística: quando cada indivíduo ou unidade da população tem uma probabilidade conhecida e não nula de ser selecionado. Para fim desse estudo detalharemos as amostragens probabilísticas dos tipos: aleatória simples, amostragem estratificada e amostragem por conglomerado.

O método de amostragem aleatória simples é o mais conhecido pela facilidade de sua implementação. Nele, todos os indivíduos/unidades de uma população têm a mesma probabilidade de seleção, assumindo que a população do estudo está uniformemente distribuída, ou seja, toda a população não apresenta grandes diferenças em suas características. No momento da escolha do modelo de amostragem, pode ser decidido utilizar a reposição dos indivíduos/unidades no processo, o que implicaria na possibilidade do mesmo indivíduo/unidade ser selecionado mais de uma vez.

Em uma amostragem estratificada, a população é dividida em grupos que possuem indivíduos/unidades com as mesmas características dentro de cada grupo e heterogêneos em relação aos outros grupos e, a partir de qualquer outra técnica de amostragem, seleciona - se uma amostra dentro de cada grupo. Estes grupos são denominados estratos e não têm nenhum indivíduo/unidade que possa estar em dois estratos ao mesmo tempo. Esse tipo de amostragem pode ser uniforme, quando se seleciona o mesmo número de indivíduos/unidades em cada estrato ou proporcional, quando é respeitada a proporção de cada estrato da população e mantida na

amostra.

Em uma amostragem por conglomerados, a população é dividida em grupos (bairros, cidades, estados, escolas e etc.) de maneira a apresentarem a mesma variabilidade populacional dentro de cada grupo e depois é aplicada uma técnica de amostragem a fim de escolher “n"grupos para avaliar todos os indivíduos/unidades dentro de cada grupo. É um método mais vantajoso em relação ao custo e por essa razão é mais frequentemente utilizado em inquéritos populacionais. Este tipo de amostragem pode ser repetido em múltiplos estágios, quando uma amostra de conglomerados é selecionada em cada etapa. Os conglomerados são unidades compostas de subunidades, que vão sendo selecionadas em cada estágio até chegar ao objeto de interesse da pesquisa.

Quando a população é extremamente heterogênea, apenas um método de amostragem não é suficientemente eficaz para extrair indivíduos capazes de representarem a população como um todo; para tal, são combinados vários métodos amostrais para um mesmo delineamento amostral. É chamada amostragem complexa a existência de múltiplos estágios de seleção utilizados para compor uma amostra da população que levem a resultados com o menor erro possível.

Além dos tipos usuais de amostragem e da combinação entre eles, constituindo desenhos complexos de amostragem, outro aspecto merecedor de atenção são as probabilidades desiguais de seleção das unidades/indivíduos em cada um dos estágios. Para compensar essas desigualdades, são atribuídos pesos diferenciados aos diversos elementos da amostra, chamados de fatores naturais de expansão, os quais correspondem ao inverso do produto das probabilidades de inclusão nos diversos estágios de seleção. Finalmente, para ajustar os pesos naturais do desenho e/ou corrigir os problemas originados pela ausência ou recusa de resposta é necessário fazer a calibração para totais conhecidos da população. A calibração dos fatores naturais de expansão consiste em estimar novos pesos para cada indivíduo da amostra, através de ajuste dos pesos naturais do desenho segundo informações de variáveis auxiliares da amostra. Por exemplo, um dos propósitos da calibração é obter distribuições amostrais das variáveis auxiliares semelhantes às distribuições populacionais obtidas pelo censo (SZWARCWALD; DAMACENA, 2008; PESSOA; SILVA, 1998).

A utilização de métodos de análise estatística sem levar em consideração a amostragem complexa pode acarretar em erros inferenciais, fazendo com que o investigador interprete

erroneamente o resultado.

3 DADOS FALTANTES

Uma busca no Pubmed feita em 09 de janeiro de 2018, com os descritores *"Missing data or Missing value"* no título ou abstract proporcionou 8.504 resultados nos últimos cinco anos. Quando se buscou *"(Missing data or Missing value) and complex sample"* 5 resultados foram encontrados no mesmo período. A falta de dados surge com frequência no dia a dia de análises de banco de dados e são consideradas inevitáveis. Entretanto, é importante tratar essas perdas de maneira a melhor adequar à metodologia para análise sem gerar ou, pelo menos, minimizar vieses nos resultados (MACIEL, 2012; BELL; KROMREY; FERRON, 2009).

3.1 MECANISMOS E PADRÕES DE DADOS FALTANTES

Para Rubin e colaboradores (1987) antes de qualquer tratamento em um banco de dados com dados faltantes é importante fazer uma distinção entre os conceitos de "padrão de não resposta" e de "mecanismos geradores de não resposta". O primeiro se refere à configuração dos dados observados e não observados em uma base de dados que podem ser considerados do tipo monotônicos ou não-monotônicos (BLANKERS, 2011), 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.

3.1.1 Padrão de não resposta

O padrão dos dados faltantes implica no tipo de tratamento feito para a análise dos dados. Padrões monotônicos permitem uma maior variedade de métodos, enquanto um padrão não monotônico necessita do uso de métodos mais complexos para essa manipulação.(LITTLE; RUBIN, 1987; SHAFER; GRAHAM, 2012)

É chamado de padrão monotônico quando mais de uma variável possui faltas, entretanto existe uma ordem e o conjunto de dados é considerado uma matriz diagonal ou superior ou inferior, como exemplificado na matriz A. No caso de apenas uma variável possuir faltas, o padrão é dito como monotônico univariado.

$$\mathbf{A} = \left|\begin{array}{c|cccccc} \mathrm{Y} & X_1 & X_2 & X_3 & X_4 & \cdots & X_n \\ \mathrm{y}_1 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & & \mathrm{x} \\ \mathrm{y}_2 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & & - \\ \mathrm{y}_3 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & & - \\ \mathrm{y}_4 & \mathrm{x} & \mathrm{x} & \mathrm{x} & - & & - \\ \vdots & & & & \cdots & & \\ \mathrm{y}_n & \mathrm{x} & \mathrm{x} & - & - & & - \end{array}\right|$$

Fonte: O próprio o autor(2017).

E padrão não-monotônico é definido assim quando mais de uma variável possui faltas, e não existe uma ordem para as perdas, que são distribuídas de maneira aleatória, como exemplificado na matriz B.

$$\mathbf{B} = \left|\begin{array}{c|cccccc} \mathrm{Y} & X_1 & X_2 & X_3 & X_4 & \cdots & X_n \\ \mathrm{y}_1 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & & \mathrm{x} \\ \mathrm{y}_2 & - & \mathrm{x} & \mathrm{x} & - & & \mathrm{x} \\ \mathrm{y}_3 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & & - \\ \mathrm{y}_4 & \mathrm{x} & - & \mathrm{x} & \mathrm{x} & & - \\ \vdots & & & & \cdots & & \\ \mathrm{y}_n & \mathrm{x} & \mathrm{x} & \mathrm{x} & - & & \mathrm{x} \end{array}\right|$$

Fonte: O próprio o autor(2017).

Antes de realizar os possíveis tratamentos aos dados faltantes é importante definir também, além do padrão, os determinantes dessas faltas. A probabilidade de um entrevistado não informar a resposta da variável de interesse pode estar relacionada com o valor do dado que foi omitido e com os valores das respostas de outras variáveis na pesquisa. O conhecimento do mecanismo gerador da não resposta permite a escolha correta da técnica para melhor tratá-las (RUBIN, 1976).

Conforme a classificação de Little e Rubin (1987), abordamos três mecanismos geradores de não resposta: Perda Completamente Aleatória(PCA), Perda Não Completamente Aleatória (PNCA) e Perda Não Aleatória (PNA).

3.1.2 Perda Completamente Aleatória (PCA) dos dados

A PCA, também conhecida por Missing Completely at Random (MCAR), tem melhor compreensão por se tratar de uma perda que pode ocorrer aos dados de qualquer indivíduo da amostra. O conjunto de dados perdidos correspondem a uma subamostra do banco de dados, proporcionando estimativas de parâmetros não viesadas mas com baixa precisão, reduzindo o poder estatístico, já que a amostra se reduz.

Representando estatisticamente essa perda, suponha uma matriz $Y_{nxp} = (Y_{ij})$ [1] e uma matriz indicadora de dados faltantes $L_{nxp} = L_{ij}$ que tem como dado $L_{ij} = 0$ quando Y_{ij} está presente e $L_{ij} = 1$ quando Y_{ij} está ausente. Esse mecanismo é modelado por um conjunto de distribuições de probabilidade $\tau(L|\psi)$, onde ψ é o parâmetro estatístico.

$$P(L = l|Y, \psi) = P(L = l|\psi) \qquad \text{para todo } l, \psi \tag{3.1}$$

Exemplo: A ausência de gestantes em consultas pré-natais quando se fariam coletas de dados antropométricos para uma pesquisa epidemiológica.

3.1.3 Perda Não Completamente Aleatória (PNCA) dos dados

Quando os dados ausentes são do tipo PNCA, também conhecido por Missing at Random (MAR), significa que as respostas faltantes pertencem a um subgrupo específico que pode ser identificado por outros dados, completamente preenchidos, existentes dentro do conjunto de dados, e dentro desses grupos os dados ausentes são do tipo PCA.

$Y_{nxp} = (Y_{obs}, Y_{falt})$ [2] a localização de Y_{ij} dentro do conjunto de dados é $Y_{ij} \in Y_{obs}$ se e somente se $L_{ij} = 0$ e caso contrário $Y_{ij} \in Y_{falt}$.

$$P(L = l|Y, \psi) = P(L = l|Y_{obs}, \psi) \qquad \text{para todo } l, \psi \tag{3.2}$$

Exemplo: A ausência de gestantes adolescentes em consulta pré-natal para o preenchimento de um questionário, durante o período escolar. Ou então a recusa de uma adolescente solteira responder se manteve relações sexuais quando esta está próxima dos pais.

A partir do momento em que se descobre a qual grupo os dados ausentes pertencem, é possível fazer inferência para o parâmetro de interesse, encontrando estimativas não viesadas, assim como no mecanismo PCA.

3.1.4 Perda Não Aleatória (PNA) dos dados

O mecanismo PNA ocorre de maneira sistemática, quando uma resposta ou um conjunto de respostas não é respondido por um grupo de indivíduos e não tem como ser identificados por

[1] Onde n representa o número de indivíduos, p o número de variáveis, i representa o i-ésimo indivíduo da amostra da j-ésima variável.

[2] obs=valor observado na amostra; falt = valor faltante na amostra

outras respostas inclusas no banco de dados. Esse mecanismo é o que acarreta mais viés no momento das análises de dados. Como a probabilidade da resposta depende da própria questão não respondida e um conjunto de características não observadas nos dados existentes, então temos:

$$P(L = l|Y, \psi) = P(L = l|Y_{falt}, \psi) \qquad \text{para todo } l, \psi \tag{3.3}$$

Exemplo: Uma grávida usuária de drogas não respondendo a uma questão sobre utilização de drogas ilícitas por saber que isso pode ocasionar problemas de saúde ao bebê.

3.2 MÉTODOS PARA TRATAMENTO DE DADOS FALTANTES

Os softwares atuais tem como programação padrão a análise do conjunto de dados completos. Entretanto, a maioria das pesquisas possui alguma perda de dados. Tendo em vista essa necessidade de adaptação dos dados para as inferências, várias técnicas de tratamento de dados faltantes são utilizadas para minimizar possíveis prejuízos.

3.2.1 Análise de Casos Completos

Uma das maneiras mais usuais no tratamento dos dados é a restrição das informações apenas aos indivíduos que tenham todas as características preenchidas. A facilidade de implementação desse método auxilia no processo de disseminação desse tipo de abordagem entretanto, um dos maiores erros é quando não se explora o mecanismo das perdas dos indivíduos descartados permanentemente.

Quando é feito o descarte das informações, a amostra é reduzida se tornando uma subamostra o que pode influenciar na representatividade da população. Se o mecanismo que rege as perdas for do tipo PCA o maior problema que pode existir é de precisão dos parâmetros estimados que podem estar contidos em intervalos de maior amplitude. Entretanto, se o mecanismo for ou do tipo PNCA ou PNA, onde não resposta é sistemática, pode motivar tanto a imprecisão quanto a invalidade da estimativa estudada.

3.2.2 Análise de Casos Disponíveis

Essa técnica se assemelha ao caso anterior para as análises e também com as mesmas preocupações sobre as inferências. A diferença é que a análise de casos disponíveis não exclui

nenhum sujeito do estudo, tudo é relativo a cada análise. O método, utilizado por softwares, como por exemplo o SPSS, restringe o n para cada análise específica. Quando têm-se um conjunto de análises de um mesmo banco de dados que se utilizou esse método de tratamento, o tamanho de amostra para cada análise será diferente devido às perdas em cada variável.

Para tentar minimizar as falhas que possam ocorrer na restrição do banco de dados, algumas técnicas para a completude são utilizadas. Estas técnicas, chamadas de imputação de dados, consistem em completar os dados a partir de valores estimados estatisticamente para então analisar o conjunto de dados obtido (valores observados mais os valores imputados) de maneira convencional (GRAHAM,2012; GARCÍA-PEÑA,ARCINIEGAS-ALARCÓN & BARBIN, 2014).

Várias formas de imputação têm sido propostas na literatura, as quais se enquadram em um dos dois tipos de imputação: simples ou múltipla. Através do método de imputação simples cada valor ausente é substituído por um único valor imputado. Há várias formas de se fazer isto, mas, deve-se ter cautela ao utilizar a imputação simples, pois esta técnica distorce a incerteza dos dados. Já a imputação múltipla consiste em substituir cada perda por mais de um valor imputado. Os conjuntos de dados completados são analisados e usados para estimar um valor plausível que representa a incerteza sobre o valor a ser imputado.

3.2.3 Imputação simples

A imputação simples (ou única) consiste em preencher o valor que falta por um valor estimado de maneira que se obtenha um conjunto de dados completos, para fazer as inferências de maneira usual. Vários métodos para a imputação simples são implementados sendo os principais apresentados a seguir.(BARACHO, 2003; ZHANG, 2003; NUNES et al, 2010; CASTRO, 2014)

3.2.3.1 Medidas de tendência central

As medidas são utilizadas por serem de fácil obtenção em um conjunto de dados disponíveis, média e mediana para variáveis quantitativas e moda para variáveis qualitativas. Quando não existem dados discrepantes que influenciem na média ou se o mecanismo de perda que rege os dados for do tipo PCA, a média se torna um bom valor para a estimação.(BARACHO, 2003; ZHANG, 2003; NUNES et al, 2010)

De acordo com Little e Rubin (2002), entre as desvantagens da imputação por média estão: o tamanho da amostra é superestimado, a variância é subestimada, a correlação é negati-

vamente tendenciosa e a distribuição de novos valores é incorreta.

3.2.3.2 Regressões

A partir de variáveis que possam ser associadas àquela variável com perda, é possível buscar valores preditos para completar os dados faltantes utilizando o método de regressão, seja ela logística ou linear. A predição de dados ausentes através deste método parece bem razoável pois, além de levar em conta as estimativas dos efeitos fixos, também leva em conta o efeito aleatório predito.(NUNES et al, 2010)

3.2.3.3 Hot dec e Cold dec

Essa técnica implica em utilizar doadores dentro de um banco de dados que possam de alguma maneira se assemelhar em outras características ao indivíduo que possua dado faltante. Doadores são aqueles indivíduos que possuem todas as respostas preenchidas e suas informações são utilizadas para completar os dados que faltarem. A escolha desse doador é aleatória, entretanto ele é selecionado dentro de um grupo específico de indivíduos que possuam as mesmas características pré determinadas das do indivíduo com dado faltante. O método denominado hot dec utiliza um doador que esteja presente na mesma base de dados, sendo assim, utiliza informações correntes à pesquisa, já o método cold dec utiliza outras bases de dados, sejam de uma mesma pesquisa de um tempo pregresso, ou de uma pesquisa similar. (ANDRIDGE; LITTLE, 2010)

3.2.3.4 Imputação via escore de propensão

Escore de propensão é um método de pareamento que utiliza a probabilidade predita de pertencimento a algum grupo baseado em características observadas. Geralmente se utiliza de regressão logística para a obtenção dessas probabilidades preditas. A inclusão ao contexto de imputação se dá porque esse método permite preencher cada falta de informação com um valor observado de um indivíduo que tenha a mesma probabilidade de apresentar a ausência daquela informação. Essa técnica está inclusa ao método hot dec, entretanto utilizando mais de um método estatístico para a obtenção do doador.(GOULÃO, 2013) Para a utilização desse método alguns passos são seguidos:

- Cria-se variável indicadora para a não resposta (0-sim; 1-não).

- Ajusta-se um modelo logístico com as variáveis associadas. É atribuída uma probabilidade predita para cada observação.

- Ordena os escores e agrupa-os para que se possa, dentro de cada grupo, fazer uma amostragem com reposição dos valores presentes, entretanto a amostragem é na dimensão dos ausentes para completar todos os valores não preenchidos.

3.2.4 Imputação múltipla

Uma técnica criada por Rubin (1976) tem estado em evidência devido aos bons resultados trazidos em sua utilização, além da implementação estar sendo facilitada por avanços tecnológicos. A imputação múltipla tem como método aplicar a cada dado faltante não apenas um, mas m valores utilizando, em geral, modelos de regressão para esse fim e com isso criando M base de dados completas.

O primeiro passo para a aplicação desse método de manipulação é investigar a proporção de faltas, para definir a quantidade de bases necessárias a serem criadas. A fração de informação ausente, definida por α , implica no tamanho de M que será criado e esse, por sua vez implica na eficiência relativa (na escala da variância) de uma estimativa pontual da análise do conjunto de dados. Essa eficiência Schafer (1997) definiu como sendo aproximadamente $\frac{1+\alpha}{M} - 1$. Para uma visualização mais clara sobre a eficiência temos a tabela a seguir.

Tabela 3.1 – Tabela de eficiência (%) em relação ao número de imputações

	α				
M	0,1	0,3	0,5	0,7	0,9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

Fonte: Baracho (2003, p.9).

Entretanto, o tamanho de M não necessariamente tem que ser grande (BARACHO, 2003; GRAHAM; OLCHOWSKI; GILREATH, 2007). Como visto na tabela 3.1 aumentando o número de imputações, não aumenta proporcionalmente a eficiência, então é possível aplicar o método de imputação múltipla com um número razoavelmente pequeno para M. 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 (GRAHAM; OLCHOWSKI; GILREATH,

2007,ARCINIEGAS-ALARCÓN; DIAS,2009). 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.

A partir do tamanho definido, o próximo passo é escolher o método estatístico para a criação dos conjuntos de dados, seja por regressão linear ou logística (média preditiva), método de regressão linear bayesiana, métodos MCMC (Markov Chain Monte Carlo), árvores de decisão, entre outros.

Após os M banco de dados imputados e analisados de maneira usual, faz-se uma medida geral, chamada de pool, com todas as estatísticas observadas para cada banco. Essa é uma regra de Rubin onde

$$\overline{\theta} = \frac{1}{M}\Sigma_{i=1}^{M}\hat{\theta}_i, \tag{3.4}$$

$$\overline{\sigma} = \frac{1}{M}\Sigma_{i=1}^{M}\hat{\sigma}_i, \tag{3.5}$$

$$\sigma = \frac{1}{M-1}\Sigma_{i=1}^{M}\hat{X}_i - \overline{X} \tag{3.6}$$

onde:

- $\overline{\theta}$ é a média dos parâmetros $\hat{\theta}$ observados nos M bancos;
- $\overline{\sigma}$ é a média das variâncias obtidas dentro dos M bancos imputados;
- σ é a variância entre conjunto imputado.

As regras de Rubin (Rubin rules) são regras simples que podem ser aplicadas independentemente do método utilizado para a criação dos dados imputados (SCHAFER; GRAHAM, 2002).

4 REFERENCIAL TEÓRICO

Na busca por maiores detalhes sobre o tema, foram feitas buscas de estudos anteriores, simulados ou não, nos diversos canais de artigos acadêmicos. Foram encontrados diversos estudos de simulação de perda que utilizaram um padrão ouro para comparar aos métodos de imputação por eles utilizados.

Baracho (2010) utilizou dados de um estudo longitudinal sobre a eficácia de uma droga no tratamento de episódios depressivos para a testar alguns métodos de imputação. Foi imputada a variável nível de depressão, medida em alguns momentos durante o estudo. Os métodos de imputação testados foram:

- Imputação pela última observação → Foi observada a última resposta do paciente, completando os dados com essa informação;
- Imputação por média por tempo → Foi calculada a média do nível de depressão para cada indivíduo em momentos distintos completando os dados com essa informação;
- Média por indivíduos → Foi calculada a média de todos os indivíduos em um tempo específico e utilizado esse valor para completar o conjunto de dados;
- Imputação simples por regressão → Calcula-se a média preditiva para cada observação faltante a partir de parâmetros estimados; e por fim
- Imputação múltipla → que foi detalhado na sessão anterior, além da análise de dados completos (aac).

A conclusão observada foi a da necessidade de se utilizar um método adequado. Como foi encontrada associação significativa para todos os modelos de regressão utilizados, o autor não concluiu sobre qual foi o melhor método de imputação, apenas concluindo que o efeito da droga não influenciou no nível de depressão.

Para o estudo de Silva (2012), foram simulados bancos de dados com perdas de 10%, 20% e 30% de um estudo sobre árvores de eucalipto. A variável imputada foi a altura dessas árvores aos 5 anos de idade e os métodos escolhidos para a imputação foram:

- Imputação múltipla com enfoque bayesiano por MCMC;

- Método de imputação múltipla por decomposição de valores singulares:

Para comparar a eficiência do método, foram comparadas as médias e desvios dos valores imputados em relação aos valores originais. Foi observado maior diferença do erro quadrático médio quanto maior a proporção de perda. Para perdas de no máximo 20%, foi visto que a imputação múltipla foi superior à decomposição de valores singulares. Com 30% dos dados faltantes, os métodos se equiparam nos resultados

Já Heijden e colaboradores (2006), em seu estudo sobre embolia pulmonar, fizeram uma comparação de métodos para tratamento de dados faltantes com utilizando os seguintes métodos:

- Análise de casos completos (ACC);
- Método do indicador de dado faltante $\rightarrow$ Que insere variáveis do tipo dummy para indicar a falta (1-sim) ou a presença (0-não) da informação ;
- Imputação simples com média;
- Imputação múltipla por média preditiva.

A partir da prevalência entre os grupos com e sem dados faltantes, foi detectado que o mecanismo de não resposta não foi o PCA. Para a comparação entre os métodos, foram utilizados os coeficientes de regressão assim como o erro padrão e área da curva ROC. A direção dos coeficientes para todos os métodos foi a mesma. Na análise de casos completos em comparação aos outros métodos a estimativa do beta foi inflacionada, o que ocorre quando a perda não é aleatória. Esse estudo não tinha um padrão ao qual comparar, assumindo assim como uma limitação para essa publicação.

No estudo de Nunes e colaboradores (2010) utilizou-se dados de pacientes de uma coorte sobre a utilização de albumina para comparação entre imputação única e imputação múltipla. Abaixo são descritos os métodos utilizados:

- Análise de casos completos (ACC);
- Mediana;
- O método do valor normal $\rightarrow$ Utiliza o limite inferior da faixa de normalidade da variável a ser imputada (a albumina)

- Imputação múltipla $\rightarrow$ sendo utilizada a abordagem bayesiana, e dois modelos para a comparação, p primeiro construído a partir de todas as variáveis com possível relação com a variável com perda e o segundo com todas as variáveis do anterior, mais o desfecho da análise original.

Como resultado foram observadas semelhanças entre os coeficientes de regressão pelos métodos de imputação. Em comparação à análise de casos completos, a utilização do método de imputação simples se torna mais eficiente. Entretanto esse método não levou em consideração a variabilidade amostral devido a utilização das medidas de tendência central. A imputação múltipla, indiferente do modelo de estimação utilizado, e mesmo com toda a sua complexidade, foi o melhor método para o tratamento de dados faltantes para este estudo.

A partir de um inquérito de saúde domiciliar realizado em Belo Horizonte e que têm plano amostral complexo, Camargos e colaboradores (2011) imputaram a variável IMC com apenas o método de imputação múltipla, entretanto com modelos diferentes: os modelos de imputação múltipla seguiram uma hierarquia ao acrescer uma variável ao modelo anterior se iniciando pelas variáveis idade, cor e escolaridade. Para o 2^{o} modelo foram incluídos os desfechos ao modelo anterior e aos modelos subsequentes foram inclusas as variáveis diretamente ligadas ao IMC. A análise de casos completos (ACC) foi comparada aos quatro modelos de imputação múltipla.

Como resultado, a ACC apresentou em média maior desvio entre as abordagens, seguida pela imputação sem o desfecho na hora da imputação. As melhores predições foram as que utilizaram as variáveis diretamente ligadas ao IMC, como teste visual de silhueta, que obtiveram menor desvio da razão de chances na análise final.

Nos estudos citados acima, a imputação múltipla foi considerada mais vantajosa dentre as principais técnicas de tratamento de dados faltantes e o melhor método para predição de diversas variáveis. Uma vez que os valores faltantes foram imputados, as técnicas de análise para dados completos puderam ser utilizadas, sendo aplicável a vários níveis de perda, e com fácil implementação. Entretanto, a variação das informações permaneceu devido ao número de valores imputados.

A literatura é crescente no que diz respeito aos métodos de imputação, entretanto ainda escassa do tocante sobre imputação em amostragem complexa. É necessário se preocupar também com esse tipo de dados já que o aumento no uso desse delineamento amostral é crescente.

Neste sentido, procura-se resposta para a seguinte pergunta: qual método de tratamento de dados faltantes permitiria, a partir de uma amostragem complexa, estimar os parâmetros de maneira mais válida possível?

4.1 ESTUDO NASCER NO BRASIL

O estudo "Nascer no Brasil: Inquérito Nacional sobre Parto e Nascimento" foi pensado em 2009 e realizado entre os anos de 2011 e 2012 pela Fundação Oswaldo Cruz, em parceria com diversas outras instituições científicas e importantes pesquisadores e descreveu a atenção à gestação e ao parto para 23.894 mulheres em 266 hospitais no país.

Como objetivo a pesquisa estimou a prevalência de cesarianas e outras intervenções obstétricas e neonatais, descreveu complicações maternas e neonatais de acordo com o tipo de parto, descreveu a motivação das mulheres para opção pelo tipo de parto, as consequências no que diz respeito a prematuridade, a saúde mental materna pós-parto. Descreveu também a estrutura hospitalar oferecida e a relacionou com os desfechos obstétricos e neonatais.

O desenho amostral da pesquisa foi feito estratificando os hospitais por macrorregião (Norte, Nordeste, Sudeste, Sul, Centro-oeste), localização (capital, não capital) e tipo de hospital (público, privado, ou misto). Como base para o cálculo do tamanho amostral, foram considerados os nascimentos no ano de 2007, e para a seleção de hospitais foram considerados os que realizaram mais de 500 partos. A amostra foi realizada de maneira a detectar diferença de 14% sobre a proporção de cesarianas entre os tipos de hospitais.

A seleção dos hospitais foi feita com probabilidade proporcional ao tamanho, levando em consideração o número de partos no ano de 2007, sendo definida como primeiro estágio. No segundo estágio, foram selecionados o número de dias (mínimo de sete) que seriam necessários em cada hospital a fim de alcançar o total de 90 mulheres pré-estipulado. No terceiro estágio foi feita uma amostragem inversa, que determinou a quantidade de entrevistas diárias (máximo de 12) com o intuito de alcançar as 90 puérperas em pelo menos sete dias consecutivos. O número de entrevistas diárias também dependeu do número de turnos de pesquisa e do número de entrevistadores disponíveis por turno em cada hospital. Para definir os números de turnos de pesquisa, foi usado o número médio de nascidos vivos por hospital em 2007. Quatro combinações foram definidas: se fosse apenas um entrevistador, então ou ele permaneceria em um turno com quatro entrevistas diárias ou dois turnos para seis entrevistas; se fossem utilizados

dois entrevistadores então seria um turno com oito entrevistas diárias ou dois turnos com doze entrevistas por dia. O número de ordem da puérpera era de acordo com a entrada no hospital. O método completo da amostragem encontra-se descrito em Vasconcelos et al (2014).

Participaram do estudo todas as mulheres admitidas nas maternidades selecionadas por ocasião do parto e seus conceptos, vivos ou mortos, com peso ao nascer $\geq$ 500 g e/ou idade gestacional $\geq$ 22 semanas de gestação. Foram critérios de exclusão: puérpera com transtorno mental grave, que não permita a comunicação com o entrevistador; indígenas ou estrangeiras que não compreendam o idioma português; surda/muda; e mulheres internadas por decisão judicial, para interrupção da gravidez. As mães foram entrevistadas face a face no pós-parto imediato, e por telefone, 45 a 60 dias após o parto. Também foram anotados dados do prontuário hospitalar da mãe e do recém - nascido. As informações dos prontuários foram obtidas após o término da internação da puérpera e do recém-nascido. Foram captadas informações do cartão da gestante e de exame de ultrassonografia por meio de fotografia e também utilizadas para obtenção de informações sobre o pré - natal e a idade gestacional na ocasião da entrevista. Além disso, em cada unidade de saúde incluída na pesquisa foi realizada entrevista com o gestor da maternidade e chefias específicas para coletar informações sobre estrutura e processo de trabalho.

Para garantir a qualidade dos dados coletados, foram elaborados manuais com descrições detalhadas dos procedimentos para seleção da população do estudo e coleta de dados; assim como foram ministrados treinamentos padronizados a toda a equipe em cada estado. Além disso, o preenchimento dos questionários foi realizado de forma digital facilitando o correto preenchimento dos campos. Os supervisores de campo replicam os questionários preenchidos previamente, em uma amostra aleatória de 2,5% das mães entrevistadas. Antes de iniciar a coleta, cada estabelecimento participante recebeu a visita do coordenador estadual e/ou do supervisor que entregou uma carta ao gestor, com cópia resumida do projeto e do parecer do Comitê de Ética em Pesquisa. O estudo foi aprovado no Conselho de Ética em Pesquisa - CEP da Escola Nacional de Saúde Pública - ENSP, sob o CAAE 0096.0.031.000-10. (Leal & Gama, 2012)

4.2 JUSTIFICATIVA

A definição de uma abordagem para a análise de um banco de dados com perda de informação é um dos principais problemas em modelagem estatística de dados epidemiológicos.

Explorar abordagens analíticas adequadas para analisar conjuntos de dados com observações incompletas é uma questão que pode ser bastante delicada, pois a utilização de métodos inapropriados pode levar a conclusões errôneas sobre o evento na população. Métodos específicos para solucionar problema de dados faltantes estão sendo discutidos de maneira ativa nos últimos anos.

A técnica de imputação vem sendo muito utilizada ultimamente pela facilidade de implementação, pelo bom resultado que proporciona e pela flexibilidade na utilização das variáveis, sejam numéricas, categóricas, fatores ou desfecho. No entanto, quando se trata sobre amostragem complexa existe pouca literatura e a dificuldade de inserção do desenho amostral é um dos principais fatores.

Este trabalho pretende ampliar a discussão sobre imputação de dados no processo de modelagem dos dados em estudos epidemiológicos levando em consideração a amostragem complexa do estudo "Nascer no Brasil: Estudo Nacional Sobre Parto e Nascimento". Pioneiro na abordagem da assistência ao parto e nascimento e de abrangência nacional, sua relevância para a produção de conhecimento na área perinatal foi inegável. Deste modo, discutir métodos de imputação utilizando a amostra do "Nascer no Brasil"torna-se ainda mais importante e oportuno.

5 OBJETIVOS

5.1 OBJETIVO GERAL

Avaliar, a partir de análises estatísticas, métodos de tratamento de dados faltantes que minimizem os vieses nas estimações de parâmetros em inquérito epidemiológico com amostragem complexa.

5.2 OBJETIVOS ESPECÍFICOS

- Descrever os mecanismos de perda de informações.
- Descrever as principais estratégias de imputação de dados faltantes.
- Comparar os métodos de imputação descritos a partir de uma aplicação à dados reais de pesquisa epidemiológica com plano amostral complexo - Nascer no Brasil.

6 MÉTODOS

No presente estudo foi utilizado um recorte do banco de dados da pesquisa Nascer no Brasil cujos dados são provenientes de um estudo longitudinal com delineamento amostral complexo. A população escolhida foi de puérperas adolescentes. Após o recorte inicial, os dados totalizaram 4.571 puérperas adolescentes, com idades entre 12 e 19 anos.

A escolha dessa população se deu pelo fato do conjunto de dados dessa população conter o mínimo de faltas em variáveis que, com base na literatura, possam ter algumas associações.

6.1 O BANCO DE DADOS PARA A ANÁLISE

6.1.1 Preparação dos bancos com perdas

Três mecanismos de perda foram simulados para testar os métodos de imputação dos dados e nos passos seguintes seguem as descrições de como foram criadas as perdas dos dados:

1- Mecanismo PCA: amostragem de perda aleatória simples utilizando os estratos de seleção da pesquisa como um filtro, trazendo a randomização dentro de cada estrato e proporcionando às adolescentes a mesma probabilidade de inclusão.

2- Mecanismo PNCA: amostragem de perda estratificada utilizando como filtro, além da variável estrato de seleção, as variáveis escolaridade (utilizando a baixa escolaridade como motivo para a não resposta), cor da pele (utilizando a cor da pele parda ou preta como motivo para a não resposta) e região (utilizando não morar nem na região Sudeste e nem na Sul como motivo para a não resposta).

3- Mecanismo PNA: amostragem de perda estratificada utilizando como filtro, além da variável estrato de seleção, as variáveis escolaridade (utilizando a baixa escolaridade como motivo para a não resposta), planejamento da gravidez atual (utilizando não queria engravidar como motivo para a não resposta) e situação do bebê no momento da entrevista (utilizando ser natimorto ou óbito neonatal como motivo para a não resposta).

Foram extraídas amostras de 30% de perda[1], de maneira aleatória dentro de cada conjunto, respeitando as singularidades de cada mecanismos. O comando sample do software R

[1] 30% para Rubin é considerado o máximo de perdas para que não influencie nas análises.

project foi o meio de processamento para essas amostragens. As variáveis que perderam observações estão dispostas na tabela 6.1.

Tabela 6.1 – Variáveis com necessidade de tratamento de dados e respectivos mecanismos de perda.

Variáveis para imputação	PCA	PNCA	PNA
Adequação da escolaridade	X		
Classe socioenonômica	X		
Cor da pele	X		
Peregrinou para internação	X		
Adequação do pré-natal	X	X	X
Idade gestacional	X	X	
Intenção da gravidez	X	X	
Fumo na gravidez	X		X
Uso de álcool na gravidez	X		X

Fonte: O próprio o autor(2017).

Para o mecanismo PNCA, as variáveis escolhidas para perda foram: adequação do pré-natal, idade gestacional no parto e intenção de gravidez. Por suposição, existiria maior dificuldade de acesso à informações sobre a assistência pré-natal e idade gestacional no parto para as mulheres com menor escolaridade, pretas ou pardas e de regiões menos favorecidas (variáveis utilizadas na associação com essas perdas).

Já no mecanismo PNA, as variáveis escolhidas para perda foram: fumo na gravidez, uso de álcool na gravidez e adequação do pré-natal. Por suposição, mulheres com baixa escolaridade, que não quiseram engravidar e com bebês em situação grave (variáveis utilizadas na associação com essas perdas) podem não querer responder sobre o uso de álcool e fumo, além da maior dificuldade de acesso à informações sobre a assistência pré-natal para essas mulheres.

6.2 ESTRATÉGIAS PARA O TRATAMENTO DOS DADOS

As análises foram feitas em diversos cenários: O recorte do banco, com as 4.571 foi analisado com e sem as informações do plano amostral. O intuito dessas análises foi observar os resultados de associações para concluir se há ou não discrepância quando se utiliza as informações do plano de amostragem.

6.2.1 As imputações

Foram testados três métodos de imputação: imputação pelo vizinho mais próximo, imputação múltipla por média preditiva (testado com M=5 e M=10 a fim de comparações) e im-

putação por escore de propensão, já descritos anteriormente.

Para o método do vizinho mais próximo e imputação múltipla, e considerando cada um dos três mecanismos de simulação de perda (PCA, PNCA e PNA), os dados foram imputados das seguintes maneiras:

- foram utilizadas as variáveis socioeconômicas e demográficas (região geográfica; idade; tipo de financiamento, se pública ou privada e escolaridade do chefe), SEM a utilização do peso e das variáveis que compõem o desenho amostral;
- foram utilizadas as variáveis socioeconômicas e demográficas (região geográfica; idade; tipo de financiamento, se pública ou privada e escolaridade do chefe), COM a utilização do peso e das variáveis que compõem o desenho amostral;
- foram utilizadas todas as variáveis do banco consideradas nos modelos de regressão como preditoras, SEM a utilização do peso e das variáveis que compõem o desenho amostral.
- foram utilizadas todas as variáveis do banco consideradas nos modelos de regressão como preditoras, COM a utilização do peso e das variáveis que compõem o desenho amostral.

O método do escore de propensão não permite que tenha dados ausentes nos preditores na regressão logística, então não foi possível a utilização de todas as variáveis, como feito para os métodos do vizinho mais próximo e imputação múltipla. Sendo assim, para o método do escore de propensão, os dados foram imputados das seguintes maneiras:

- foram utilizadas as variáveis socioeconômicas e demográficas (região geográfica; idade; tipo de financiamento, se pública ou privada e escolaridade do chefe), SEM a utilização do peso e das variáveis que compõem o desenho amostral.
- foram utilizadas as variáveis socioeconômicas e demográficas (região geográfica; idade; tipo de financiamento, se pública ou privada e escolaridade do chefe), COM a utilização do peso e das variáveis que compõem o desenho amostral.

6.2.2 Modelo para comparações

Serão avaliados os estimadores das razões de chance (RC) simples e seus respectivos intervalos de confiança de 95% para o modelo de regressão logistica que tem como desfecho a

peregrinação e dos coeficientes betas ($\hat{\beta}$) com seus respectivos intervalos de confiança de 95% para o modelo de regressão linear que tem como desfecho (Y) da idade gestacional. A escolha das variáveis do modelo de peregrinação se baseou em estudos que demonstraram um risco duas vezes maior deste desfecho entre as adolescentes mais pobres, com um pré-natal inadequado, baixa escolaridade e sem situação conjugal estável (MONTESCHIO et al, 2014; BARNASTEFANO et al, 2010; MENEZES et al, 2006). Já a prematuridade, se associa com pré-natal inadequado; baixa escolaridade; baixa classe socioeconômica, hábitos maternos inadequados e algumas intercorrências obstétricas não inclusas neste estudo (ALMEIDA et al, 2012; MARTINS et al, 2011).

O modelo teórico é descrito abaixo:

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \tag{6.1}$$

$$RC = \frac{e^{\beta_0+\beta_1 X_{1i}}}{1+e^{\beta_0+\beta_1 X_{1i}}} \tag{6.2}$$

onde: β_0 é o intercepto do modelo

β_1 é a variação que a variável explicativa aplica no desfecho

X_{1i} é a variável explicativa referente ao i-ésimo indivíduo

ε_i é o erro aleatório

Para maiores informações e descrições dos modelos apresentados nas equações 6.1 e 6.2 podem ser vistas em Paula (2004), Morettin e Bussab (2005) e Cordeiro e Demétrio (2008).

A tabela 6.2 expõe as variáveis utilizadas para os modelos de regressão para as comparações.

Tabela 6.2 – Variáveis para a composição dos modelos de regressão logística e linear.

Desfechos	Características
Peregrinação (regressão logística)	Fatores: escolaridade materna, cor da pele materna, classe econômica do domicílio, adequação do pré natal.
Idade gestacional (regressão linear)	Fatores: escolaridade materna, cor da pele materna, classe econômica do domicílio, adequação do pré-natal, fumo durante a gestação, uso de bebida alcoólica durante a gestação e intenção da gravidez.

Fonte: O próprio o autor (2017).

Os modelos propostos tiveram um referencial teórico do sumario executivo do Nascer no Brasil, que exibe um conjunto de artigos sobre a pesquisa (LEAL et al, 2014)

Como critério de comparação das estimativas encontradas para os parâmetros supracitados após a imputação, primeiramente foi calculada a diferença quadrática entre o valor estimado imputado e o valor estimado com os dados completos. O passo seguinte foi pontuar de 1 a 7 todas as estimativas por menor diferença para cada variável, por fim foram somados os pontos de cada método.

PASSO 1: Calcular a diferença quadrática entre as estimativas

$$\text{diferença quadrática} = (\text{estimativa imputada - estimativa do padrão ouro})^2 \tag{6.3}$$

PASSO 2: Observar se a diferença do método é maior ou menor que a anterior e atribuir um ponto de 1 a 7 sendo 1 a menor diferença quadrática e 7 a maior diferença quadrática;

PASSO 3: Somar as pontuações de todas as variáveis de cada método e por fim comparar as pontuações finais obtidas, tendo melhor desempenho o método com menor pontuação.

Nas tabelas 6.3, 6.4 e 6.5 são exibidas as variáveis e em quais etapas elas foram utilizadas.

Tabela 6.3 – Variáveis e etapas em que foram utilizadas para o mecanismo de simulação de perda PCA

Variáveis	Perdas simuladas	Amostragem para simular as perdas	Imputação das perdas (socioeconomicas)	Imputação das perdas (totais)	Regressão peregrinação	Regressão idade gestacional
Amostragem						
Estrato de seleção	x	x	x	x	x	x
Peso amostral			x	x	x	x
Código de unidade primária			x	x	x	x
Socioeconomicas e demográficas						
Região			x	x		
Hospital			x	x		
Classe econômica do domicílio	x		x	x	x	x
Financiamento do parto			x	x		
Idade materna			x	x		
Cor da pele materna	x		x	x	x	x
Adequação da escolaridade materna	x			x	x	x
Escolaridade do chefe da família			x	x		
Gravidez						
Intenção da gravidez	x			x		x
Adequação do pré-natal	x			x	x	x
Fumo na gravidez	x			x		x
Uso de álcool na gravidez	x			x		x
Parto						
Tipo de parto				x		
Acompanhante na internação				x		
Situação do bebê ao nascer				x		
Desfechos						
Peregrinação	x			x	x	
Idade gestacional	x			x		x

Fonte: O próprio o autor (2017).

Tabela 6.4 – Variáveis e etapas em que foram utilizadas para o mecanismo de simulação de perda PNCA

Variáveis	Perdas simuladas	Amostragem para simular as perdas	Imputação das perdas (socioeconomicas)	Imputação das perdas (totais)	Regressão peregrinação	Regressão idade gestacional
Amostragem						
Estrato de seleção		x	x	x	x	x
Peso amostral			x	x	x	x
Código de unidade primária			x	x	x	x
Socioeconomicas e demográficas						
Região		x	x	x		
Hospital			x	x		
Classe econômica do domicílio			x	x		
Financiamento do parto			x	x		
Idade materna			x	x		
Cor da pele materna		x	x	x		
Adequação da escolaridade materna		x		x		
Escolaridade do chefe da família			x	x		
Gravidez						
Intenção da gravidez	x			x		x
Adequação do pré-natal	x			x	x	
Fumo na gravidez				x		
Uso de álcool na gravidez				x		
Parto						
Tipo de parto				x		
Acompanhante na internação				x		
Situação do bebê ao nascer				x		
Desfechos						
Peregrinação				x	x	
Idade gestacional	x			x		x

Fonte: O próprio o autor (2017).

Tabela 6.5 – Variáveis e etapas em que foram utilizadas para o mecanismo de simulação de perda PNA

Variáveis	Perdas simuladas	Amostragem para simular as perdas	Imputação das perdas (socioeconomicas)	Imputação das perdas (totais)	Regressão peregrinação	Regressão idade gestacional
Amostragem						
Estrato de seleção		x	x	x	x	x
Peso amostral			x	x	x	x
Código de unidade primária			x	x	x	x
Socioeconomicas e demográficas						
Região			x	x		
Hospital			x	x		
Classe econômica do domicílio			x	x		
Financiamento do parto			x	x		
Idade materna			x	x		
Cor da pele materna			x	x		
Adequação da escolaridade materna				x		
Escolaridade do chefe da família			x	x		
Gravidez						
Intenção da gravidez		x		x		
Adequação do pré-natal	x			x	x	
Fumo na gravidez	x			x		x
Uso de álcool na gravidez	x			x		x
Parto						
Tipo de parto				x		
Acompanhante na internação				x		
Situação do bebê ao nascer		x				
Desfechos						
Peregrinação				x	x	
Idade gestacional				x		x

Fonte: O próprio o autor (2017).

Este trabalho foi submetido e aprovado ao Comitê de Ética em Pesquisa -CEP- da Escola Nacional de Saúde Pública - ENSP, sob o CAAE 70939317.0.0000.5240. O parecer se encontra em anexo a partir da página 73.

7 RESULTADOS

Este capítulo é dividido em duas seções. Na primeira são apresentados os dados do conjunto dos dados completos em comparação ao conjunto de dados sem qualquer imputação (para cada um dos três mecanismos de perda), além das associações a partir de regressões. Na segunda seção são apresentados os dados do conjunto dos dados completos em comparação ao conjunto de dados imputados pelos métodos descritos anteriormente, além das associações a partir de regressões.

7.1 Características dos dados

Após o recorte inicial, os dados totalizaram 4.571 puérperas adolescentes, com idades entre 12 e 19 anos, sendo 70% na faixa etária de 16 a 19 anos, majoritariamente das regiões nordeste e sudeste e atendidas pelo sistema único de saúde. Grande parte dessas adolescentes se autodeclararam pardas e afirmaram terem companheiro, 62,3% e 68,5%, respectivamente. A frequência de meninas que não tiveram nenhuma gestação anterior foi de 80%. A escolaridade está adequada para mais de 50% das puérperas em todas as faixas etárias com exceção da faixa de 19 anos. Foi considerada escolaridade adequada a contagem de anos completos de estudo de um ingressante aos 9 anos de idade na primeira série do ensino fundamental e sem reprovação(SAMPAIO, 2007).

A tabela 7.1 exibe a dimensão do banco de dados e a proporção de perda para cada variável segundo os três diferentes mecanismos de perda amostrados: PCA, PNCA e PNA. Para o método PCA, como os 30% de perda foram distribuídos para as nove variáveis pré-definidas, as perdas foram relativamente baixas quando se avaliou variável por variável (aproximadamente 9%). Já para os métodos PNCA e PNA, os 30% de perda foram distribuídos em apenas três variáveis.

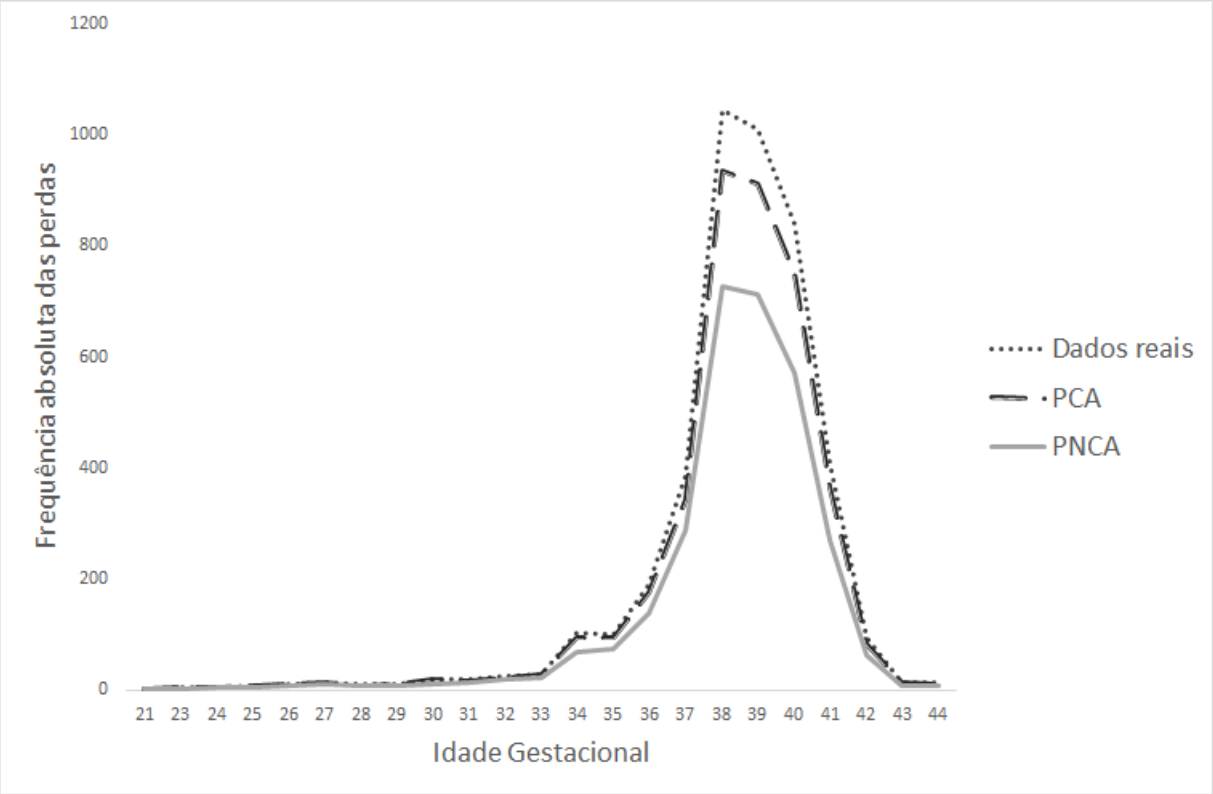
Tabela 7.1 – Distribuição de frequências absolutas das variáveis categóricas utilizadas nas análises.

Variáveis	Dados completos	PCA	PNCA	PNA
Adequação da escolaridade				
Inadequada	3.054	2.803	*	*
Adequada	1.517	1.395	*	*
Raça/cor				
Branca	1.242	1.159	*	*
Preta	405	376	*	*
Parda/morena/mulata	2.923	2.654	*	*
Classe socioeconômica				
Classe D+E	1.576	1.430	*	*
Classe C	2.465	2.271	*	*
Classe A+B	494	463	*	*
Fumou durante a gestação				
Não	4.148	3.627	*	2.945
Sim	413	378	*	321
Bebeu durante a gestação				
Não	3.947	3.627	*	2.945
Sim	505	469	*	378
Intenção da gravidez				
Sim, naquele momento	1.574	1.466	1.069	
Queria esperar mais tempo	1.520	1.413	1.014	*
Não queria engravidar	1.445	1.355	966	*
Não soube informar	31	28	23	*
Adequação do pré-natal				
Inadequado	1.018	948	609	661
Parcialmente adequado	1.059	968	773	840
Adequado	2.410	2.237	1.684	1.759
Não fez pré-natal/ não soube informar	83	73	56	55
Peregrinação				
Não	3.591	3.328	*	*
Sim	974	895	*	*

*Mesmos valores do dados completos.

Fonte: O próprio o autor (2018).

A perda da variável contínua “Idade Gestacional"também foi importante. O gráfico 7.1 delineia as curvas de resposta desta variável para os conjuntos de dados com perda pelos métodos PCA e PNCA, em comparação ao conjunto de dados completo.



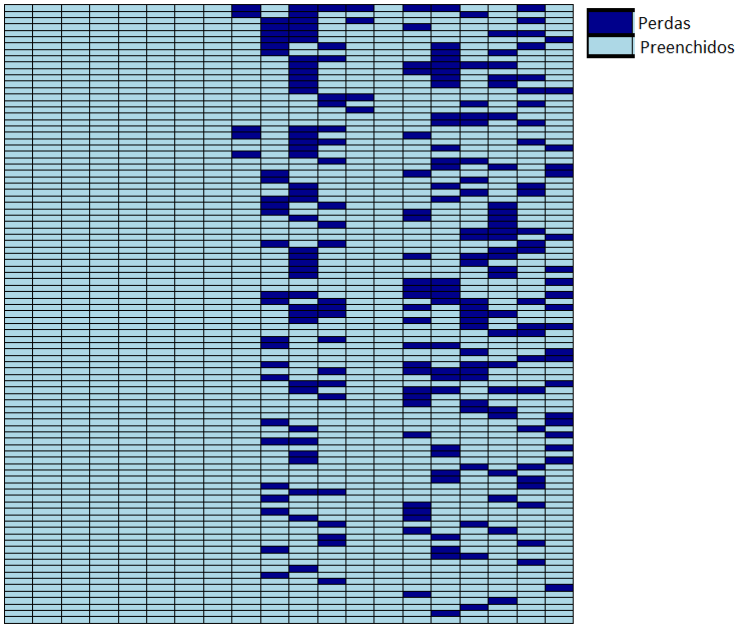
Fonte: O próprio autor (2018)

Figura 7.1 – Distribuição das respostas por idade gestacional.

Como esperado, dentro do conjunto PNCA ocorreu uma maior perda da variável idade gestacional devido à apenas três variáveis terem sido selecionadas para as perdas (intenção da gravidez, orientações no pré-natal e idade gestacional) (Figura 7.1).

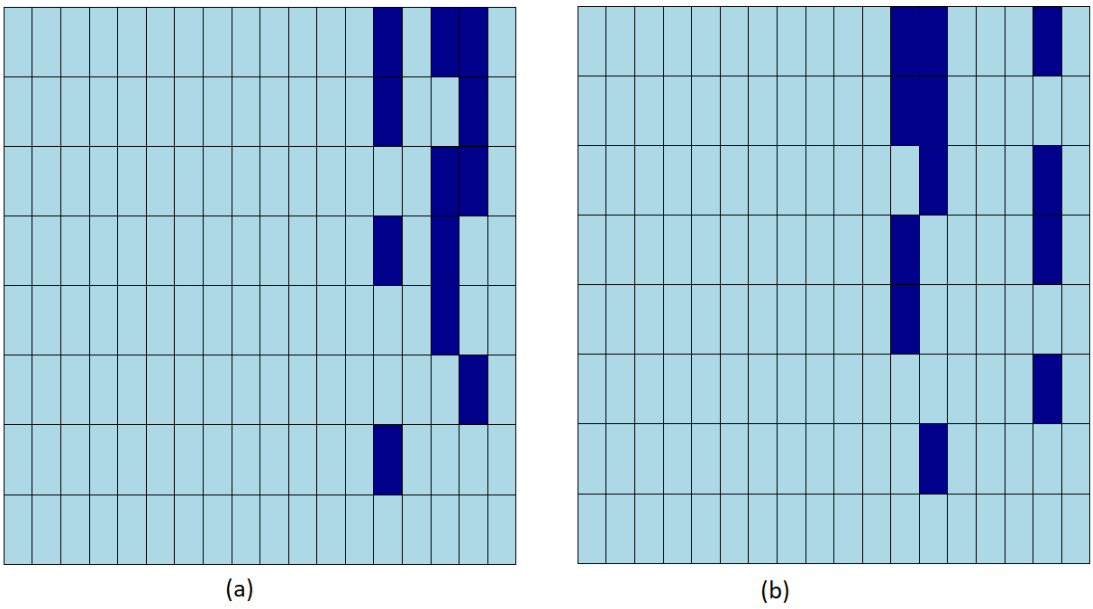
O padrão de distribuição das perdas (em azul escuro) em cada uma das sub amostras é demonstrado nos gráficos 7.2 e 7.3. Como a amostragem foi com reposição, ou seja, a mesma adolescente pode perder mais de uma informação, observam-se linhas com mais de uma perda em todas as sub amostras.

Como descrito na metodologia, as perdas na sub amostra PCA, apresentadas no gráfico 7.2, foram distribuídas em um número maior de variáveis para ter maior aleatoriedade dentre as não respostas. Já as perdas pelos métodos PNCA e PNA, apresentadas no gráfico 7.3, tiveram um número mais restrito de variáveis com perdas por terem sido simuladas de maneira a serem sistemáticas, cada uma a sua maneira, sendo possível observar um padrão não monotônico nas perdas.



Fonte: O próprio autor (2018).

Figura 7.2 – Projeção da distribuição da perda dos dados pelo mecanismo PCA em cada variável.



Fonte: O próprio autor (2018).

Figura 7.3 – Projeção da distribuição da perda dos dados pelos mecanismos (a):PNCA e (b):PNA em cada variável.

Na tabela 7.2 apresentamos as estimativas de RC ($\hat{RC}$) simples para o desfecho "peregrinação"segundo cada variável explicativa, tanto para o conjunto de dados completo quanto

para os três conjuntos de dados com perda. As $\hat{RC}$ simples dos três conjuntos de dados com perda não foram muito discrepantes e ficaram relativamente próximas aos valores modelados para o conjunto de dados completo. No entanto, como esperado, os três mecanismos de perda apresentaram um aumento no intervalo de confiança das $\hat{RC}$.

Para o mecanismo de perda PCA, a variável “adequação do pré-natal"foi a que mais se afastou da $\hat{RC}$ obtida pelo conjunto de dados completos, dentre as quatro variáveis incluídas no modelo de regressão. Para os mecanismos de perda PNCA e PNA, apenas foi possível comparar as estimativas para a variável “adequação do pré-natal", já que, para o desfecho “peregrinação", essa foi a única incluída na amostragem de perda e no modelo de regressão. Em comparação à obtida pelo conjunto de dados completo, a RC estimada pelo mecanismo de perda PNA se afastou mais do que a estimada pelo mecanismo PNCA.

Tabela 7.2 – $\hat{RC}$ simples para peregrinação dos três mecanismos de perda em comparação aos dados completos utilizando a informação do plano amostral.

Variáveis	Dados completos $\hat{RC}$ IC(95%)	PCA $\hat{RC}$ IC(95%)	PNCA $\hat{RC}$ IC(95%)	PNA $\hat{RC}$ IC(95%)
Raça/cor				
Branca	1	1		
Preta	1,44(0,93;2,23)	1,38(0,84-2,29)		
Mista	1,51(1,18;1,93)	1,51(1,16-1,97)		
Classe socioeconômica				
Classe D+E	1	1		
Classe C	0,81(0,63;1,03)	0,83(0,65-1,06)		
Classe A+B	0,47(0,33;0,69)	0,51(0,35-0,73)		
Adequação do pré-natal				
Inadequado	1	1	1	1
Parcialmente adequado	1,03(0,79;1,35)	1,15(0,81-1,62)	1,17(0,82-1,67)	1,31(0,91-1,90)
Adequado	0,59(0,39;0,87)	1,41(1,03-1,93)	1,26(0,92-1,71)	1,38(1,00-1,91)
Escolaridade adequada				
Não	1	1		
Sim	1,09(0,88;1,35)	1,12(0,89-1,42)		

Fonte: O próprio o autor (2018).

Na tabela 7.3 observamos, para os três mecanismos de perda, valores de estimados de $\beta(\hat{\beta})$ para idade gestacional na mesma direção do observado para os dados completos. Entretanto, para este desfecho, as estimativas para os dados com perda, em comparação às estimativas dos dados completos, foram mais díspares que as observadas para o desfecho peregrinação apresentado na tabela 7.2. É interessante notar que dentro de cada variável existe sempre uma categoria que se aproxima ao máximo da estimativa do dado completo.

Tabela 7.3 – Coeficientes do modelo de regressão para idade gestacional dos três mecanismos de perda em comparação aos dados completos utilizando as informações do plano amostral.

Variáveis	Dados completos $\hat{\beta}$ (IC95%)	PCA $\hat{\beta}$ (IC95%)	PNCA $\hat{\beta}$ (IC95%)	PNA $\hat{\beta}$ (IC95%)
Raça/cor				
Intercepto	38,44(38,24;38,64)	38,45(38,37;38,63)		
Branca	1	1		
Preta	-0,13(-0,74;0,49)	-0,15(-2,58;0,54)		
Mista	-0,06(-0,33;0,20)	-0,12(-0,32;0,08)		
Classe socioeconômica				
Intercepto	38,31(38,06;38,56)	38,35(38,09;38,61)		
Classe D+E	1	1		
Classe C	0,12(-0,17;0,41)	0,05(-0,27;0,36)		
Classe A+B	0,11(-0,17;0,39)	0,02(-0,28;0,32)		
Adequação do pré-natal				
Intercepto	38,21(37,89;38,54)	38,23(37,85;38,62)	38,22(37,72;38,72)	38,34(37,98;38,71)
Inadequado	1	1	1	1
Parcialmente adequado	0,46(0,11;0,81)	0,45(0,03;0,87)	0,33(-0,22;0,88)	0,33(-0,07;0,73)
Adequado	0,18(-0,18;0,55)	0,12(-0,30;0,55)	0,08(-0,49;0,65)	0,04(-0,37;0,46)
Escolaridade adequada				
Intercepto	38,41(38,26;38,57)	38,40(38,23;38,57)		
Sim	-0,08(-0,32;0,15)	-0,16(-0,44;0,11)		
Não	1	1		
Bebeu durante a gestação				
Intercepto	38,39(38,25;38,54)	38,35(38,19;38,51)		38,39(38,24;38,55)
Sim	-0,04(-0,32;0,31)	<0,001(-0,41;0,41)		-0,14(-0,50;0,21)
Não	1	1		1
Fumou durante a gestação				
Intercepto	38,39(38,26;38,53)	38,35(38,19;38,50)		38,39(38,25;38,53)
Não	1	1		1
Sim	0,05(-0,19;0,29)	0,28(-0,07;0,62)		0,28(-0,10;0,66)
Intenção da gravidez				
Intercepto	38,40 (38,20;38,60)	38,37(38,15;38,58)	38,33(38,02;38,63)	
Queria engravidar	1	1	1	
Queria esperar	0,05 (-0,19;0,29)	0,07(-0,20;0,35)	0,09(-0,38;0,50)	
Não queria engravidar	-0,1 (-0,37;0,16)	-0,07(-0,36;0,23)	0,02(-0,39;0,44)	

Fonte: O próprio o autor (2018).

7.2 Características dos dados após as imputações.

A tabela 7.4 descreve a proporção de valores discordantes dos dados reais segundo os diferentes métodos de imputação para o mecanismo PCA, utilizando ou não o peso amostral para a imputação. A imputação pelo método do vizinho mais próximo, quando apenas as variáveis socioeconômicas são as preditoras, possibilitou a recuperação exata de 60% dos dados faltantes sem a utilização do peso amostral, já com a utilização do peso amostral a recuperação exata foi de 61% . O mesmo método, porém utilizando todas as variáveis, obteve taxas de recuperação exata muito semelhantes, de 59% com o peso amostral e de 60% sem o peso amostral. As imputações múltiplas assim como o escore de propensão tiveram um desempenho pior em relação ao método do vizinho mais próximo. A imputação múltipla teve uma recuperação entre 50% e 55% dos dados, independente da utilização do peso. Já o escore de propensão, que teve o pior desempenho, recuperou apenas 43% dos dados com exatidão, independente da utilização do peso.

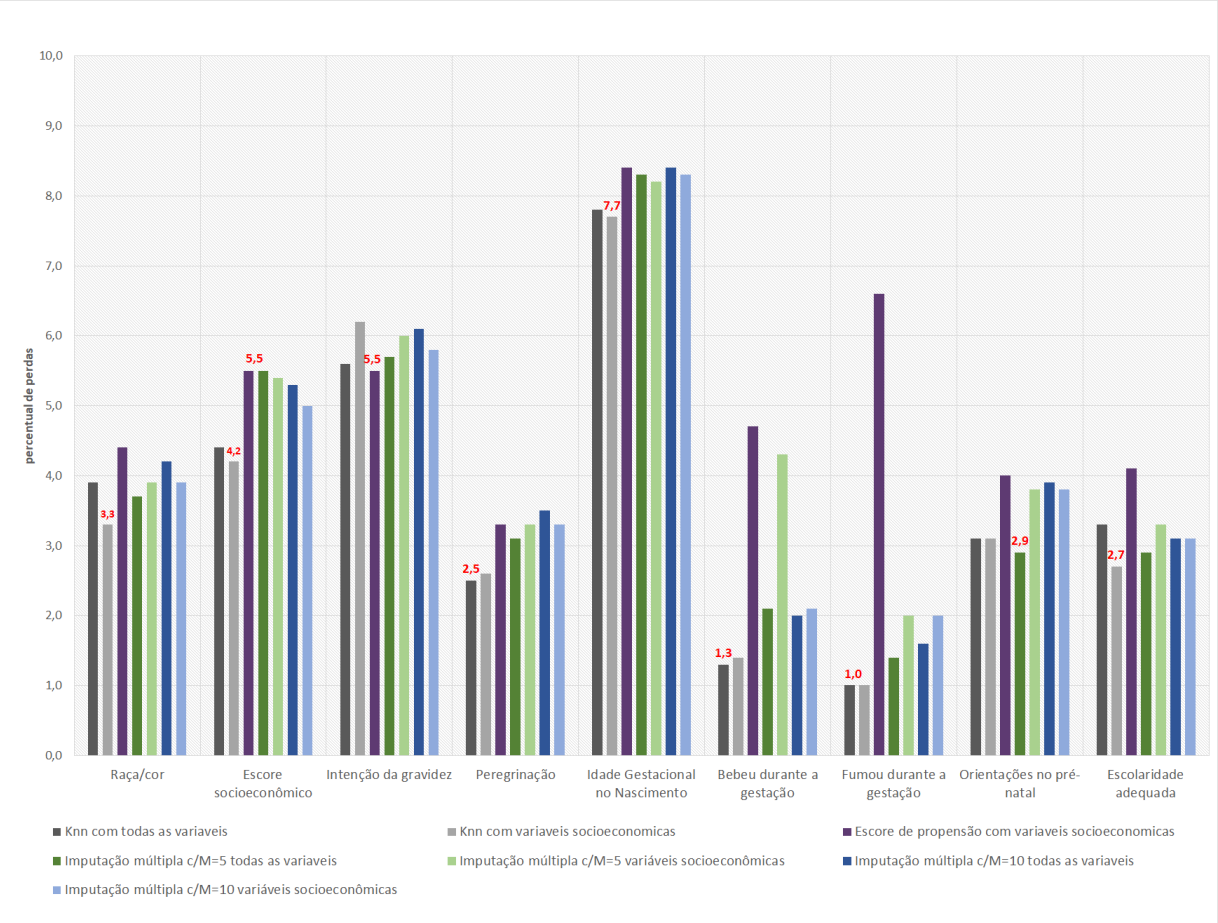
Tabela 7.4 – Frequência absoluta e relativa de perdas imputadas com valores diferentes dos reais: mecanismo de perda PCA.

Variáveis	Raça/Cor	Classe social	Intenção da gravidez	Peregrinação	Semanas gestacionais	Bebeu	Fumou	Adequação do pré-natal	Escolaridade adequada
SEM O PLANO AMOSTRAL									
Método	n(%)	n(%)	n(%)	n(%)	n(%)	n(%)	n(%)	n(%)	n(%)
Total de perdas	399(9,1)	381(8,7)	336(7,7)	378(8,7)	431(9,9)	395(9,1)	445(10,2)	375(8,6)	403(9,3)
Vizinho mais próximo									
Todas as variáveis	169(3,9)	191(4,4)	244(5,6)	109(2,5)	342(7,8)	56(1,3)	42(1,0)	135(3,1)	142(3,3)
Variáveis socioeconômicas	145(3,3)	183(4,2)	269(6,2)	113(2,6)	337(7,7)	61(1,5)	44(1,0)	137(3,1)	118(2,7)
Escore de propensão	194(4,4)	241(5,5)	240(5,5)	145(3,3)	367(8,4)	205(4,7)	289(6,6)	176(4,0)	179(4,1)
Imputação Múltipla									
M=5 e todas as variáveis	161(3,7)	239(5,5)	249(5,7)	136(3,1)	363(8,3)	93(2,1)	60(1,4)	151(2,9)	128(2,6)
M=5 e variáveis socioeconômicas	168(3,9)	235(5,4)	262(6,0)	142(3,3)	357(8,2)	189(4,3)	88(2,0)	167(3,8)	142(3,3)
M=10 e todas as variáveis	184(4,2)	232(5,3)	265(6,1)	153(3,5)	365(8,4)	86(2,0)	69(1,6)	169(3,9)	136(3,1)
M=10 e variáveis socioeconômicas	170(3,9)	216(5,0)	252(5,8)	143(3,3)	361(8,3)	90(2,1)	88(2,0)	167(3,8)	135(3,1)
COM O PLANO AMOSTRAL									
Método	n(%)	n(%)	n(%)	n(%)	n(%)	n(%)	n(%)	n(%)	n(%)
Total de perdas	399(9,1)	381(8,7)	336(7,7)	378(8,7)	431(9,9)	395(9,1)	445(10,2)	403(9,2)	375(8,6)
Vizinho mais próximo									
Todas as variáveis	163(3,7)	187(4,3)	241(5,5)	112(2,6)	337(7,7)	59(1,4)	41(0,9)	140 (3,2)	132(3,0)
Variáveis socioeconômicas	152(3,5)	192(4,4)	258(5,9)	102(2,3)	333(7,6)	59(1,4)	42(1,0)	132(3,0)	118(2,7)
Escore de propensão	204(4,7)	251(5,8)	216(5,0)	140(3,2)	352(8,1)	188(4,3)	282 (6,5)	172(3,9)	193(4,4)
Imputação Múltipla									
M=5 e todas as variáveis	183(4,2)	234(5,4)	254(5,8)	137(3,1)	354(8,1)	102(2,3)	73(1,7)	160(3,7)	141(3,2)
M=5 e variáveis socioeconômicas	176(4,0)	231(5,3)	258(5,9)	150(3,4)	351(8,0)	94(2,2)	84(1,9)	170(3,9)	148(3,4)
M=10 e todas as variáveis	185(4,2)	239(5,5)	259(5,9)	127(2,9)	352(8,1)	92(2,1)	76(1,7)	159(3,6)	134(3,1)
M=10 e variáveis socioeconômicas	159(3,6)	238(5,5)	259(5,9)	139(3,2)	356(8,2)	89(2,0)	71(1,6)	172(3,9)	138(3,2)

Fonte: O próprio o autor (2018).

Nas figuras 7.4 e 7.5 é possível observar de maneira mais clara qual tipo de tratamento teve um melhor desempenho na recuperação de dados (menor percentual de dados imputados com valores diferentes dos reais) dentro de cada variável.

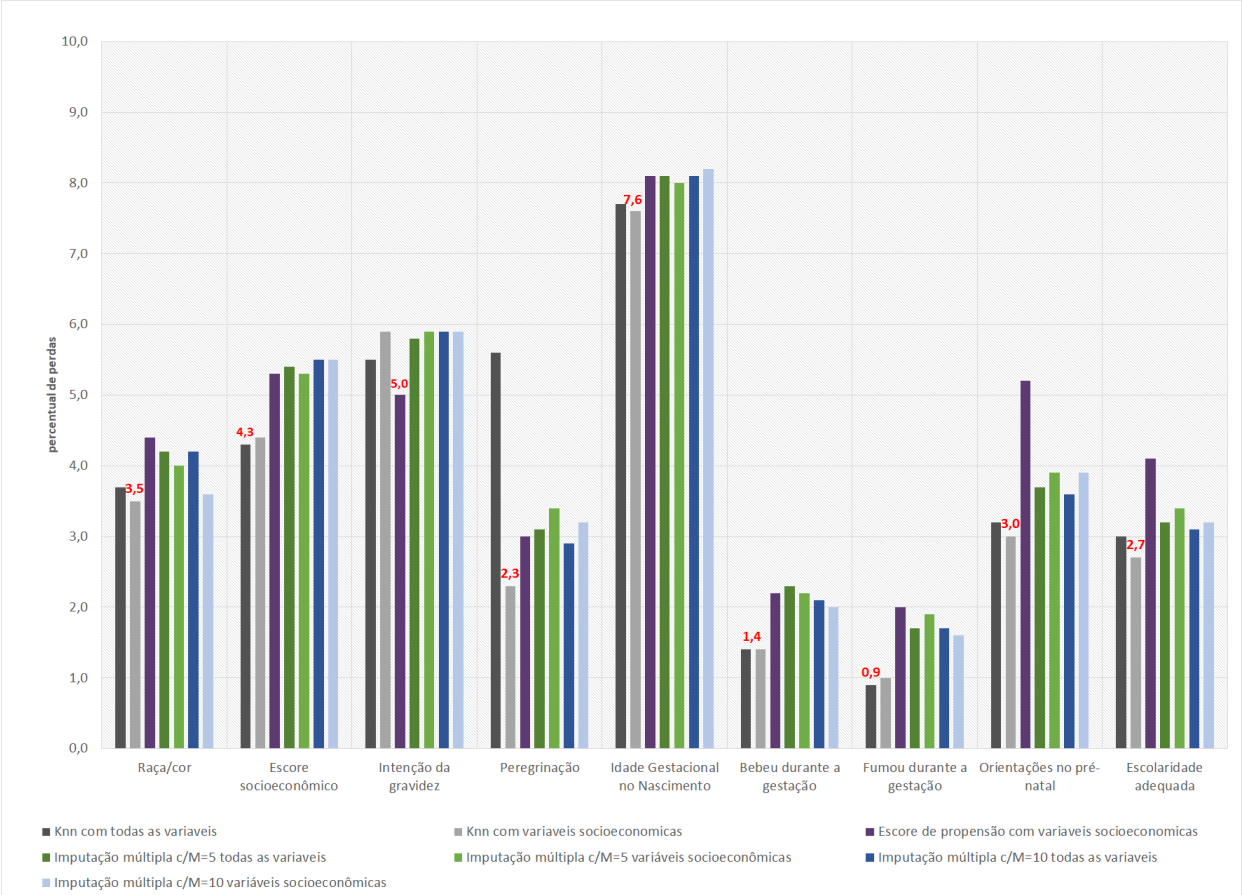
Na figura 7.4, análise sem a informação do plano amostral, o método do vizinho mais próximo se mantém melhor no maior número de variáveis exceto para variável escolaridade adequada, que teve como melhor método a imputação múltipla fazendo o uso de M=10 e todas as variáveis para predição.



Fonte: O próprio autor (2018).

Figura 7.4 – Proporção de dados imputados com valores diferentes dos reais segundo os diversos métodos de imputação sem a utilização do plano amostral para o mecanismo de perda PCA.

Já na figura 7.5, com a informação do plano amostral, o método do vizinho mais próximo teve melhor desempenho em todas as variáveis. A utilização do peso não foi tão significante quando comparamos as proporções de dados recuperados de maneira exata.



Fonte: O próprio autor (2018).

Figura 7.5 – Proporção de dados imputados com valores diferentes dos reais segundo os diversos métodos de imputação com a utilização do plano amostral para o mecanismo de perda PCA.

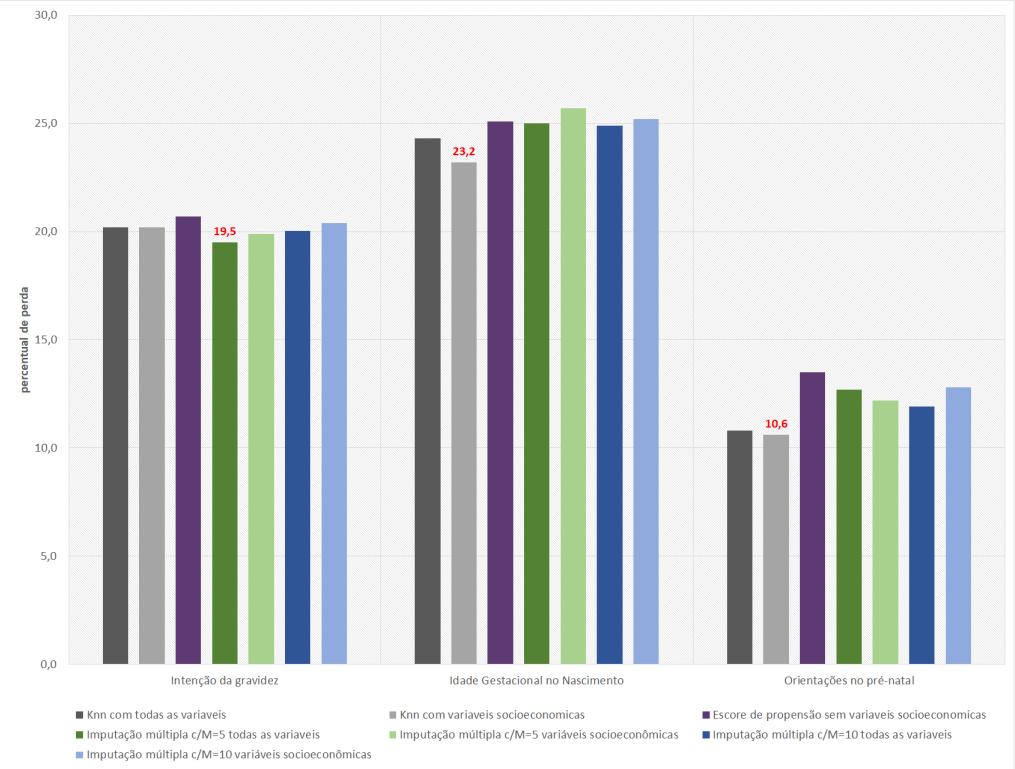
A tabela 7.5 descreve a proporção de valores discordantes dos dados reais segundo os diferentes métodos de imputação para o mecanismo de perda PNCA, utilizando ou não o plano amostral para a imputação. De maneira global, os valores recuperados pelo método do vizinho mais próximo apenas com as variáveis socioeconômicas, assim como no mecanismo PCA, foram os com maior precisão quando comparados com os valores reais, obtendo 39% de recuperação, independente da utilização do plano. Com o uso de todas as variáveis para a predição, foi possível obter 37% dos dados semelhantes aos reais, sem a utilização do plano da amostra, e 38% com o uso do plano amostral. A imputação por escore de propensão teve 33% dos dados recuperados tendo ou não o plano amostral na predição. A imputação múltipla manteve a recuperação dos dados em torno de 34% para os oito cenários, tendo melhor desempenho (35%) na imputação feita com o plano amostral e M=10, sendo indiferente o tipo de variável para a predição ou sem plano, todas as variáveis para a predição.

Tabela 7.5 – Frequência absoluta e relativa de perdas imputadas com valores diferentes dos reais: mecanismo de perda PNCA.

Variáveis	Intenção da gravidez	Idade gestacional	Orientações no pré-natal
SEM PLANO AMOSTRAL			
Método	n(%)	n(%)	n(%)
Total de perdas	1.289 (29,6)	1.321 (30,3)	1.240 (28,4)
Vizinho mais próximo			
Todas as variáveis	879 (20,2)	1.061 (24,3)	473 (10,8)
Variáveis socioeconômicas	881 (20,2)	1.013 (23,2)	462 (10,6)
Escore de propensão	904 (20,7)	1.093 (25,1)	591 (13,5)
Imputação Múltipla			
M=5 e todas as variáveis	851 (19,5)	1.092 (25,0)	553 (12,7)
M=5 e variáveis socioeconômicas	868 (19,9)	1.119 (25,7)	532 (12,2)
M=10 e todas as variáveis	888 (20,0)	1.084 (24,9)	521 (11,9)
M=10 e variáveis socioeconômicas	888 (20,0)	1.101 (25,2)	559 (12,8)
COM PLANO AMOSTRAL			
Método	n(%)	n(%)	n(%)
Total de perdas	1.289 (29,6)	1.321 (30,3)	2.980 (68,3)
Vizinho mais próximo			
Todas as variáveis	850 (19,5)	1.052 (24,1)	487 (11,2)
Variáveis socioeconômicas	881 (20,2)	1.035 (23,7)	450 (10,3)
Escore de propensão	903 (20,7)	1.103 (25,3)	563 (12,9)
Imputação Múltipla			
M=5 e todas as variáveis	881 (20,2)	1.105(25,3)	566 (13,0)
M=5 e variáveis socioeconômicas	872 (20,1)	1.111 (25,5)	578 (13,3)
M=10 e todas as variáveis	864 (19,8)	1.079 (24,7)	570 (13,1)
M=10 e variáveis socioeconômicas	868 (19,9)	1.112 (25,5)	528 (12,1)

Fonte: O próprio o autor (2018).

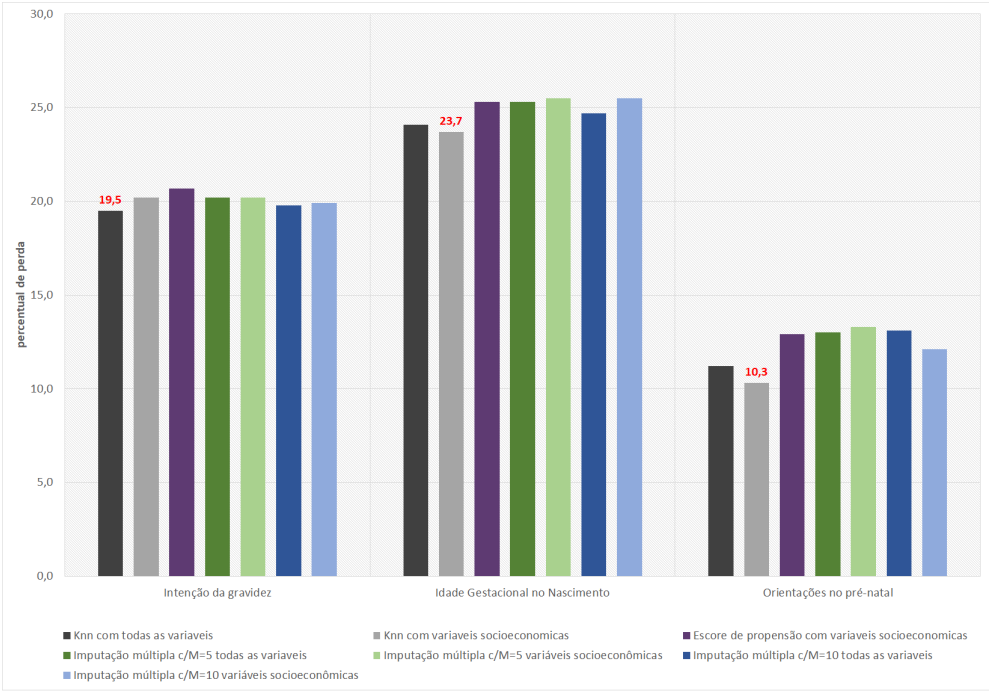
Observando o gráfico 7.6, que exibe as imputações sem a informação do plano amostral, podemos perceber uma uniformidade nas imputações da variável intenção da gravidez e uma maior oscilação entre os métodos dentro da variável idade gestacional. Nas três variáveis o melhor método foi o do vizinho mais próximo com as variáveis sociodemográficas como preditoras.



Fonte: O próprio autor (2018).

Figura 7.6 – Perdas de dados do mecanismo PNCA após a imputação de dados pelos diversos métodos sem a utilização do plano amostral.

No gráfico 7.7, que exibe os dados dos valores imputados utilizando o plano amostral, apresenta uma variação um pouco maior entre as variáveis dentro de cada imputação. A variável intenção da gravidez tem como melhor desempenho a imputação múltipla com M=5 e todas as outras variáveis como preditoras. O método do vizinho mais próximo foi melhor para as outras duas variáveis, tendo como preditoras apenas as socioeconômicas.



Fonte: O próprio autor (2018).

Figura 7.7 – Perdas de dados do mecanismo PNCA após a imputação de dados pelos diversos métodos com a utilização do plano amostral.

Na tabela 7.5 estão apresentados os resultados da imputação para o conjunto de dados com perda não aleatória (PNA). Assim como nos casos anteriores, o vizinho mais próximo teve o melhor desempenho. Seja utilizando o plano amostral ou não, com todas as variáveis como preditoras o total de dados recuperados foi de 76%.

Quando observamos o percentual do mesmo método imputado com as variáveis socioeconômicas, mas sem o plano, o percentual permaneceu 73%. Já para a imputação com o desenho amostral, o total imputado de maneira precisa foi de 72%.

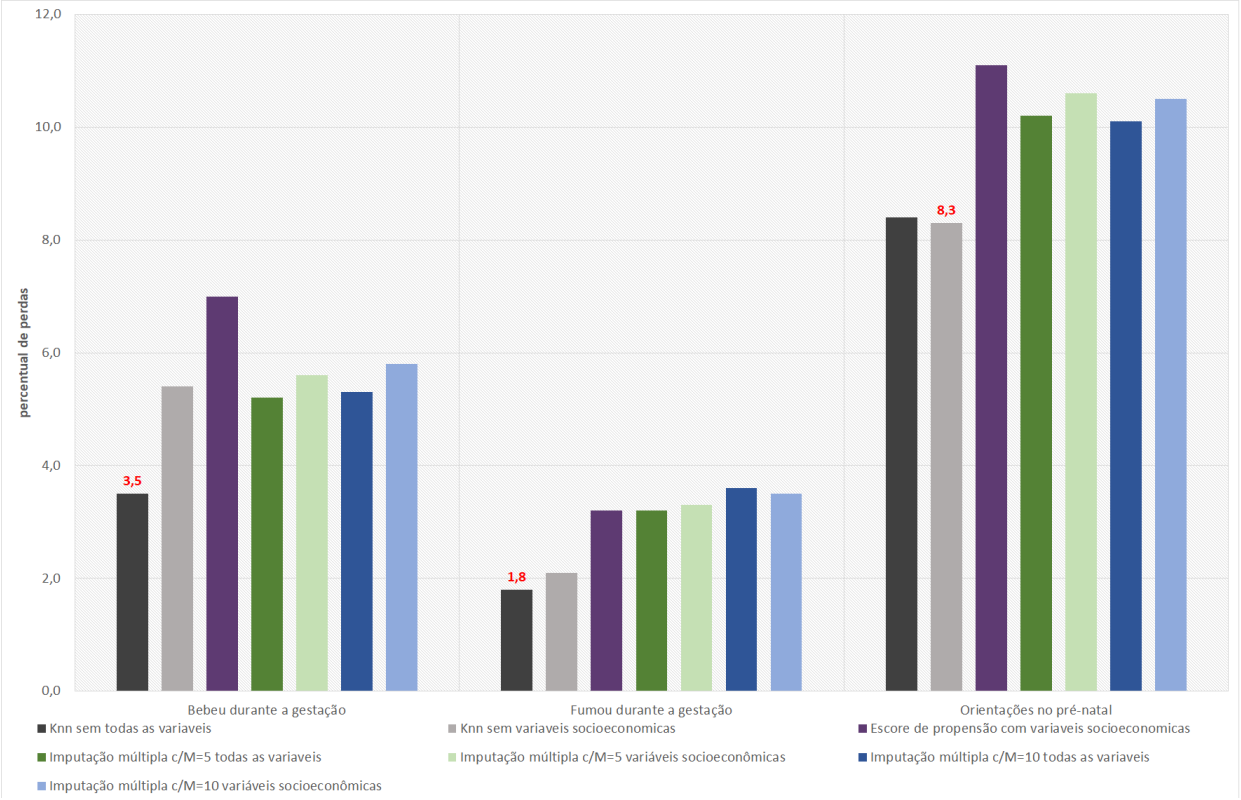
O escore de propensão teve 63% dos dados recuperados utilizando o plano e 61%sem a utilização do mesmo. Já os métodos de imputação múltipla tiveram como maior recuperação 68% utilizando o plano e todas as variáveis como preditoras e M=5, e como menor recuperação foi 65% que utilizou o M = 10, apenas as variáveis socioeconômicas como preditora e não utilizou o peso amostral.

Tabela 7.6 – Frequência absoluta e relativa de dados imputados com valores diferentes dos reais no conjunto de dados com o mecanismo PNA.

Variáveis	Bebeu durante a gestação	Fumou durante a gestação	Orientações no pré-natal
SEM PLANO AMOSTRAL			
Método	n(%)	n(%)	n(%)
Total de perdas	882(20,2)	796(18,2)	832(19,1)
Vizinho mais próximo			
Todas as variáveis	151(3,5)	80(1,8)	366(8,4)
Variáveis socioeconômicas	234(5,4)	90(2,1)	36(8,3)
Escore de propensão	306(7,0)	138(3,2)	487(11,1)
Imputação Múltipla			
M=5 e todas as variáveis	229(5,2)	140(3,2)	447(10,2)
M=5 e variáveis socioeconômicas	244(5,6)	145(3,3)	463(10,6)
M=10 e todas as variáveis	232(5,3)	155(3,6)	441(10,1)
M=10 e variáveis socioeconômicas	255(5,8)	153(3,5)	460(10,5)
COM PLANO AMOSTRAL			
Método	n(%)	n(%)	n(%)
Total de perdas	882(20,2)	796(18,2)	832(19,1)
Vizinho mais próximo			
Todas as variáveis	149(3,4)	87(2,0)	379(8,7)
Variáveis socioeconômicas	233(5,3)	94(2,2)	378(8,7)
Escore de propensão	317(7,3)	168(3,9)	488(11,2)
Imputação Múltipla			
M=5 e todas as variáveis	222(5,1)	138(3,2)	441(10,1)
M=5 e variáveis socioeconômicas	224(5,1)	152(3,5)	495(11,3)
M=10 e todas as variáveis	229(5,2)	159(3,6)	637(10,0)
M=10 e variáveis socioeconômicas	235(5,4)	147(3,4)	456(10,5)

Fonte: O próprio o autor (2018).

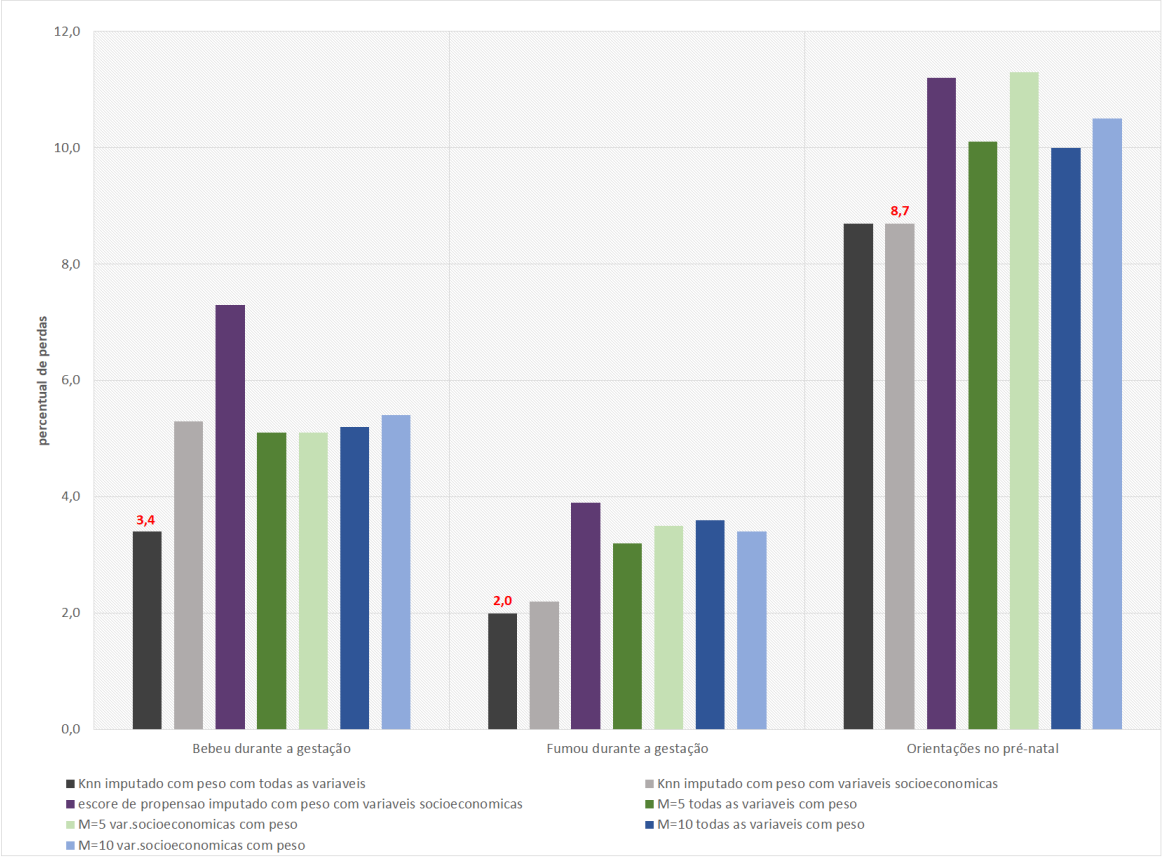
O gráfico 7.8 apresenta as proporções dentro de cada variável quando comparados os valores imputados com os reais. A variável bebeu na gestação foi melhor predita pela imputação múltipla com M=10 e utilizando todas as variáveis. Para as outras duas, o melhor método permaneceu sendo o mesmo das informações globais para esse tipo de perda, método do vizinho mais próximo.



Fonte: O próprio autor (2018).

Figura 7.8 – Perdas de dados do mecanismo PNA após a imputação de dados pelos diversos métodos sem a utilização do plano amostral.

O gráfico 7.9 apresenta as proporções dentro de cada variável quando comparados os valores imputados com os reais utilizando o plano na hora da imputação. O método do vizinho mais próximo teve o melhor desempenho nas três variáveis observadas, entretanto para as duas primeiras variáveis dispostas, o melhor método foi aquele que utilizou todas as variáveis como preditoras, já a terceira utilizando apenas variáveis socioeconômicas teve o melhor resultado.



Fonte: O próprio autor (2018).

Figura 7.9 – Perdas de dados do mecanismo PNA após a imputação de dados pelos diversos métodos com a utilização do plano amostral.

Os resultados dos modelos de regressão estimados para a criação de pontuações e das diferenças médias quadráticas estão exibidos no apêndice. As pontuações para a comparação entre os parâmetros estão expostas nas tabelas 7.7, 7.8 e 7.9. Foram destacados os valores com menor ponto em cada variável e a menor soma final.

A perda do tipo PCA sem a utilização do plano teve menor pontuação no método imputação múltipla com M=5 e as variáveis socioeconômicas como preditoras. A pontuação encontrada com o somatório de todos os pontos foi de 29. Já o escore de propensão teve 26 pontos a mais que o método anterior, sendo considerado o pior método por este critério.

Para o mesmo mecanismo, entretanto utilizando o plano amostral, a menor pontuação foi para o método do vizinho mais próximo, utilizando todas as variáveis como preditoras. O escore desse método foi de 35, que teve uma diferença de 19 pontos a maior pontuação, que foi também para o método escore de propensão

Tabela 7.7 – Pontuações criadas a partir das diferenças quadráticas dentro de cada variável e o total para o mecanismo PCA.

Variáveis	Imputação Múltipla M=5 Todas as variáveis	Imputação Múltipla M=5 Socioeconômicas	Imputação Múltipla M=10 Todas as variáveis	Imputação Múltipla M=10 Socioeconômicas	Vizinho mais próximo Todas as variáveis	Vizinho mais próximo Socioeconômicas	Escore de propensão Socioeconômicas
Sem informação do plano amostral							
Modelo logístico							
Raça/cor	7	1	3	6	4	2	5
Classe social	3	4	1	5	2	2	6
Adequação do pré-natal	3	6	1	5	2	4	7
Escolaridade adequada	5	3	3	1	2	4	1
Modelo linear							
Raça/cor	5	4	6	7	1	3	2
Classe social	1	2	4	6	5	3	7
Adequação do pré-natal	7	6	2	1	4	3	5
Escolaridade adequada	4	7	1	3	2	6	5
Bebeu	5	1	4	7	6	3	2
Fumou	2	3	1	6	5	4	8
Intenção da gravidez	6	5	3	4	1	2	7
Total Escore	48	42	29	51	34	36	55
Com informação do plano amostral							
Modelo logístico							
Raça/cor	5	1	6	7	2	3	4
Classe social	2	4	6	2	1	3	7
Adequação do pré-natal	3	7	1	4	2	5	6
Escolaridade adequada	3	4	6	1	5	7	2
Modelo linear							
Raça/cor	6	2	7	1	3	5	4
Classe social	1	2	4	6	3	5	7
Adequação do pré-natal	5	7	2	6	4	3	1
Escolaridade adequada	1	2	6	4	3	8	5
Bebeu	2	7	1	3	6	4	5
Fumou	5	6	2	1	4	3	7
Intenção da gravidez	4	7	5	3	2	1	6
Total Escore	37	49	46	38	35	47	54

Fonte: O próprio o autor (2018).

Para o mecanismo PNCA - apresentado na tabela 7.8, sem a utilização do plano amostral, o método do vizinho mais próximo utilizando as variáveis socioeconômicas teve a menor pontuação (8 pontos) que foi 7 pontos abaixo do maior valor que foi obtido pela imputação múltipla com M = 10 e sem diferença para as variáveis preditoras.

Já com o uso do plano, o método com menor pontuação foi a imputação múltipla, com M=5 e todas as variáveis. A pontuação observada é de 4, diferença de 14 pontos de amplitude para o maior valor que foi obtido pela imputação múltipla com M=10 e pelo método do escore de propensão.

Na tabela 7.9 exibe a pontuação para o mecanismo PNA. Tanto para a abordagem com peso, quanto para a abordagem sem plano, a menor pontuação é observada no método de imputação múltipla , com M=5 e todas as variáveis para a predição. Foram 9 pontos quando não utiliza o plano e 12 quando utiliza para a imputação. Para o primeiro, a amplitude é de 16 pontos de diferença para o maior, já o segundo a amplitude é de 10.

O método do vizinho mais próximo utilizando o plano e todas as variáveis também obteve 12 pontos.

Tabela 7.8 – Pontuações criadas a partir das diferenças quadráticas dentro de cada variável e o total para o mecanismo PNCA.

Variáveis	Imputação Múltipla M=5 Todas as variáveis	Imputação Múltipla M=5 Socioeconômicas	Imputação Múltipla M=10 Todas as variáveis	Imputação Múltipla M=10 Socioeconômicas	Vizinho mais próximo Todas as variáveis	Vizinho mais próximo Socioeconômicas	Escore de propensão Socioeconômicas
Sem informação do plano amostral							
Modelo logístico							
Adequação do pré-natal	6	4	3	1	5	2	7
Modelo linear							
Adequação do pré-natal	3	6	2	7	5	4	1
Intenção da gravidez	1	5	6	7	4	2	3
Total Escore	10	15	11	15	14	8	11
Com informação do plano amostral							
Modelo logístico							
Adequação do pré-natal	2	6	7	4	1	3	5
Modelo linear							
Adequação do pré-natal	1	5	4	2	6	3	8
Intenção da gravidez	1	7	4	6	3	2	5
Total Escore	4	18	15	12	10	8	18

Fonte: O próprio o autor (2018).

Tabela 7.9 – Escores criados a partir das diferenças quadráticas dentro de cada variável e o total para o mecanismo PNA.

Variáveis	Imputação Múltipla		Imputação Múltipla		Vizinho mais próximo		Escore de propensão
	M=5	M=5	M=10	M=10			
	Todas as variáveis	Socioeconômicas	Todas as variáveis	Socioeconômicas	Todas as variáveis	Socioeconômicas	Socioeconômicas
Sem informação do plano amostral							
Modelo logístico							
Adequação do pré-natal	1	6	4	7	3	2	5
Modelo linear							
Adequação do pré-natal	1	6	3	4	7	5	2
Bebeu	1	6	2	5	7	3	4
Fumou	6	7	2	3	4	5	1
Total Escore	9	25	11	19	21	15	12
Com informação do plano amostral							
Modelo logístico							
Adequação do pré-natal	1	4	7	3	2	5	6
Modelo linear							
Adequação do pré-natal	5	7	2	4	6	3	1
Bebeu	2	6	7	4	3	5	1
Fumou	4	5	3	2	1	7	6
Total Escore	12	22	19	13	12	20	14

Fonte: O próprio o autor (2018).

Na tabela 7.10 encontramos as diferenças globais entre todas as estimativas. Interessante notar que, apesar do método do vizinho mais próximo ter tido o melhor desempenho na recuperação dos dados em todos os cenários, as diferenças entre as estimativas dos parâmetros mostra que para cada mecanismo existiu um melhor método, independente da utilização do plano. Para o mecanismo PCA, tendo como critério de melhor desempenho a diferença quadrática, a imputação múltipla assim como o método do vizinho mais próximo quando apenas as variáveis socioeconômicas são inclusas nos modelos para predição tiveram os melhores desempenhos. A utilização do plano amostral se torna mais importante para o vizinho mais próximo proporcionando estimativas de parâmetro bem próximas aos parâmetros reais observados pelas distâncias criadas.

O método que trouxe menor diferença quadrática para as estimativas dos parâmetros $\hat{RC}$ e $\hat{\beta}$ no mecanismo PNCA foi o método do vizinho mais próximo, entretanto, se fosse determinada a escolha da imputação múltipla para completar os dados, poderia ser usada o método utilizando M=5 e todas as variáveis como preditoras ou M=10 e apenas as variáveis socioeconômicas como preditoras, entretanto com a utilização do peso em ambos os casos

Tabela 7.10 – Diferenças quadráticas entre os diversos métodos de imputação de dados para os três mecanismos de perda.

Método	PCA s/plano	PCA c/plano	PNCA s/plano	PNCA c/plano	PNA s/plano	PNA c/plano
Imputação múltipla = 5						
Todas as variáveis	0,37	0,33	0,13	0,09	0,37	0,25
Variáveis socioeconômicas	0,22	0,22	0,16	0,14	0,22	0,30
Imputação múltipla = 10						
Todas as variáveis	0,27	0,34	0,17	0,18	0,45	0,36
Variáveis socioeconômicas	0,34	0,33	0,24	0,09	0,26	0,22
Método do vizinho mais próximo						
Todas as variáveis	0,24	0,23	0,17	0,16	0,34	0,29
Variáveis socioeconômicas	0,24	0,22	0,10	0,08	0,28	0,38
Escore de propensão						
Variáveis socioeconômicas	0,34	0,32	0,11	0,21	0,21	0,19

Fonte: O próprio o autor (2018).

O software R-Project se mostrou muito versátil, permitindo a imputação pelos vários métodos, tanto os presentes neste estudo quanto em estudos passados, e permitindo a inclusão das variáveis que compoe o desenho da amostra e o peso durante os processamentos para a imputação. Os pacotes VIM e MICE, que foram utilizados para as imputações deste trabalho fazem imputação pelo vizinho mais próximo e imputação múltipla, respectivamente. Estes são de fácil implementação e com bastante benefícios pelo que foi observado.

8 DISCUSSÃO

Neste trabalho foram feitas amostragem para a comparação dos métodos de imputação do vizinho mais próximo, escore de propensão e imputação múltipla para os mecanismos de perda PCA, PNCA e PNA. Foram utilizadas proporções de recuperação exata dos dados faltantes, diferenças quadráticas entre as estimativas dos parâmetros dos dados completos e dos dados imputados e pontuações criadas a partir dessas diferenças quadráticas para definir qual a melhor metodologia para o tratamento dos dados utilizados. O método do vizinho mais próximo teve um bom desempenho na recuperação exata dos dados para os três mecanismos, entretanto a imputação múltipla recuperou valores de modo quase tão eficaz quanto o melhor método.

Quando utilizadas as diferenças quadráticas para a indicação do melhor método, foi observado que para cada mecanismo de perda, um método teve maior destaque do que os outros. Para o PCA a imputação múltipla com M=5 e utilizando apenas as variáveis socioeconômicas para a predição. Para o mecanismo PNCA o método do vizinho mais próximo foi o que teve menor diferença quadrática entre os parâmetros estimados. Já para o PNA o escore de propensão, mesmo com a limitação de existir apenas um grupo de variáveis para as predições, teve a menor diferença para as estimativas dos parâmetros de razões de chance e β.

A utilização do método do vizinho mais próximo foi adequada por que, por mais que não possua muitas comparações na literatura utilizando este método, se mostrou tão eficiente quanto o método de imputação múltipla, sendo superior em alguns casos de recuperação de dados, mesmo sendo por poucos pontos percentuais de diferença. Já a metodologia do escore de propensão, por existir a limitação de não poder possuir valores ausentes para a sua utilização, teria que ser estudada com maior aprofundamento.

A literatura mostra maior quantidade de estudos com variáveis do tipo quantitativas com perdas trazendo um grande número de métodos de imputação bayesianos. Esse trabalho teve como maior contribuição trazer comparações entre mecanismos e métodos que ainda não foram comparados nos outros estudos utilizando a inferência clássica, que é mais difundida no contexto das análises de inquéritos de saúde, além de utilizar um número significativo de variáveis categóricas, que são predominantes em pesquisas de saúde.

Assim como nos estudos de Nunes (2010), Blankers (2011)e Van Der Heijden (2006),

entre outros, a imputação múltipla trouxe benefícios para a presente análise, tendo uma boa recuperação dos dados e ainda com uma boa aproximação aos parâmetros a que foram comparados.

Um dos questionamentos que esse estudo traz é a razão da imputação múltipla não ter sido melhor em todos os casos. Como uma possível resposta, o fator epidemiológico possa ter influenciado nos resultados como por exemplo variáveis de confundimento que não foram levadas em consideração. Outra questão seria a utilização da amostragem complexa que, mesmo não sendo de extrema relevância na recuperação dos dados, possa ter algum tipo de atuação sobre os resultados.

Zhang (2003), que fez um estudo teórico e metodológico, explica a importância da escolha das variáveis para a predição na abordagem via escore de propensão, já que a correlação pode trazer deficiências ao modelo logístico de predição. Uenal (2014) define que o melhor aproveitamento do escore de propensão como método de imputação é quando utilizado em variáveis contínuas, com padrão monotônico e com mecanismos do tipo PCA e PNCA.

A importância de se observar o mecanismo que rege o tipo de perda, como todos os autores enfatizam, é um fator considerável na hora da escolha do método. Além disso, ter um conhecimento a priori das variáveis associadas à perda é imprescindível para um bom modelo de predição, gerando uma recuperação dos dados mais próxima da exata possível. Como observado nos estudos da revisão para este trabalho, as variáveis preditoras inclusas no modelo de imputação múltipla são de extrema importância para melhor eficiência e validade do método.

No presente trabalho, a comparação por pontuações foi atrativa pelo fato de a amplitude das diferenças não influenciar o resultado final dessa soma. No entanto, observamos que a menor diferença (melhor pontuação) entre os métodos, aparentemente se deu de maneira aleatória. Além disso, a pontuação criada a partir das diferenças foi mais discordante para a definição do melhor método. Também na análise apresentada têm como maior parte variáveis do tipo categóricas, algumas com mais categorias do que outras, e isso influencia também no total imputado de maneira exata, por que proporcionalmente podem ser próximas, entretanto em valores absolutos pode existir prejuízos em variáveis com muita categoria.

A utilização do plano amostral, mesmo que nestes dados em questão não tenham sido tão significantes para as imputações, pode trazer uma maior eficiência aos métodos definidos. A escolha do método para a imputação se torna controverso neste estudo, mas deixa claro que

qualquer um dos métodos escolhidos não influenciariam de maneira inversa as associações, já que os métodos tiveram resultados próximos tanto entre si quanto entre o padrão ouro.

A imputação é uma maneira satisfatória de tratamento de dados faltantes, beneficiando a análise posterior com o tamanho da amostra, por que não se faz necessária a redução à dados completos e valores plausíveis para a inferência (UENAL et al, 2014; BLANKERS et al, 2011). Entretanto, este não é o único modo de tratamento dos dados, sendo possível também utilizar métodos de ponderação, os quais aumentam a representatividade de outros indivíduos para suprir a perda dos que não responderam. Contudo, os métodos de ponderação necessitam da população completa para o cálculo de probabilidades de inclusão.

A ausência de informação sobre a população de adolescentes no ano de 2007, ano utilizado para cálculo da amostra do Nascer no Brasil, estratificada por regiões, capital e interior e tipo de serviço (público, privado e misto) traz a limitação para averiguar outros tipos de tratamento. Estas estratificações foram as utilizadas para a formação da amostragem da pesquisa principal, tornando a criação dos planos das não respondentes para este trabalho impraticável pela possível diferença das gestantes adolescentes no atendimento em alguns hospitais, por serem consideradas gestantes de risco

9 CONCLUSÃO

O presente estudo esclarece formas de construção de métodos de imputação para uma amostragem complexa quando os dados são predominantemente categóricos, entretanto salienta a necessidade de aperfeiçoamento devido a relevância do assunto. Como trabalho futuro poderia resultar em conclusões mais bem definidas como o melhor procedimento para o conjunto de dados: por exemplo a busca por outros métodos ou a escolha mais requintada das variáveis para a predição dos dados ausentes. Técnicas inovadoras como machine learning prometem grandes predições tendo em vista a complexidade dos algoritmos criados com intuito de captar a maior certeza na hora de prever um resultado. Também seria possível a busca por uma metodologia que não utilize o conjunto de dados reais para definir o melhor método de imputação, já que na realidade não conhecemos os dados perdidos.

REFERÊNCIAS

ALMEIDA, Adriana Carvalho de et al. Fatores de risco maternos para prematuridade em uma maternidade pública de Imperatriz-MA. **Revista Gaúcha de Enfermagem**, v. 33, n. 2, p. 86-94, 2012.

ANDRIDGE, Rebecca R.; LITTLE, Roderick JA. A review of hot deck imputation for survey non−response. **International statistical review**, v. 78, n. 1, p. 40-64, 2010.

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

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

BARBASTEFANO, Patrícia Santos; GIRIANELLI, Vania Reis; DA COSTA VARGENS, Octavio Muniz. O acesso à assistência ao parto para parturientes adolescentes nas maternidades da rede sus. **Revista Gaúcha de Enfermagem**, v. 31, n. 4, p. 708, 2010.

BELL, Bethany A.; KROMREY, Jeffrey D.; FERRON, John M. Missing data and complex samples: The impact of listwise deletion vs. subpopulation analysis on statistical bias and hypothesis test results when data are MCAR and MAR. In: **Proceedings of the Joint Statistical Meetings, Survey Research Methods Section.** 2009. p. 4759-4770.

BLANKERS, Matthijs et al.**E-mental health interventions for harmful alcohol use: research methods and outcomes.** 2011.

BOLFARINE, Heleno; BUSSAB, Wilton O. **Elementos de amostragem.** USP, 1994.

CAMARGOS, Vitor Passos et al. Imputação múltipla e análise de casos completos em modelos de regressão logística: uma avaliação prática do impacto das perdas em covariáveis.**Cad Saúde Pública**, v. 27, n. 12, p. 2299-313, 2011.

CASTRO, Isabela Q. **Uma Aplicação de Métodos de Imputação no Estudo de Fatores Associados ao Baixo Peso ao Nascer** [Trabalho de Conclusão de Curso]. Universidade Federal de Juiz de Fora,; 2014.

CORDEIRO, Gauss Moutinho; DEMÉTRIO, Clarice GB. **Modelos lineares generalizados e extensões.** Sao Paulo, v. 33, 2008.

DIAS, Ana Cristina Garcia; TEIXEIRA, Marco Antonio Pereira. Gravidez na adolescência: um olhar sobre um fenômeno complexo. **Paidéia**, v. 20, n. 45, p. 123-131, 2010.

DIAS, Antonio José Ribeiro; ALBIERI, Sonia. Uso de imputação em pesquisas domiciliares. **Anais**, p. 11-26, 2016.

GARCÍA−PEÑA, Marisol; ARCINIEGAS−ALARCÓN, Sérgio; BARBIN, Décio.Imputação de dados climáticos utilizando a decomposição por valores singulares: uma comparação empírica. **Revista Brasileira de Meteorologia**, v. 29, n. 4, p. 527-536, 2014.

GOULÃO, Beatriz Preto Barrocas Afonso. **Seleção de variáveis na presença de valores omissos: uma aplicação na modelação do índice de massa corporal nos imigrantes africanos e brasileiros.** 2013. Tese de Doutorado.

GRAHAM, John W. Missing data: analysis and design. **Statistics for social and behavioral sciences.** New York, NY: Springer. doi, v. 10, p. 978-1, 2012.

GRAHAM, John W.; OLCHOWSKI, Allison E.; GILREATH, Tamika D.How many imputations are really needed? Some practical clarifications of multiple imputation theory. **Prevention Science**, v. 8, n. 3, p. 206-213, 2007.

KMETIC, Andrew et al. Multiple imputation to account for missing data in a survey: estimating the prevalence of osteoporosis. **Epidemiology**, v. 13, n. 4, p. 437-444, 2002.

LEAL, Maria do Carmo; GAMA, Silvana Granado Nogueira da. **Nascer no Brasil: inquérito nacional sobre parto e nascimento.** Rio de Janeiro: ENSP/Fiocruz, 2012.

LEAL, Maria do Carmo et al. Sumário Executivo Temático da Pesquisa "Nascer no Brasil". **Inquérito Nacional sobre o Parto e Nascimento**, 2014.

LITTLE, Roderick JA; RUBIN, Donald B. **Statistical analysis with missing data**. John Wiley & Sons, 2014.

MACIEL, Pricila Henkes. **Estudos longitudinais para avaliação de custo na área da saúde: como tratar dados faltantes e censuras.** Monografia (Graduação em Estatística) - Departamento de Estatística, Universidade Federal do Rio Grande do Sul 2012.

MARTINS, Marília da Glória et al . Associação de gravidez na adolescência e prematuridade. **Rev. Bras. Ginecol. Obstet.**, Rio de Janeiro , v. 33, n. 11, p. 354-360, Nov. 2011 .

MENEZES, Daniela Contage Siccardi et al. Avaliação da peregrinação anteparto numa amostra de puérperas no Município do Rio de Janeiro, Brasil, 1999/2001. **Cadernos de Saúde Pública**, v. 22, p. 553−559, 2006.

MONTESCHIO, Lorenna Viccentine Coutinho et al. Acesso de parturientes para a assistência ao parto em hospitais universitários: caracterização e fatores associados. **Revista Gaúcha de**

Enfermagem, v. 35, n. 1, p. 22−30, 2014.

NUNES, Luciana Neves; KLUCK, Mariza Machado; FACHEL, Jandyra Maria Guimarães. Comparação de métodos de imputação única e mútipla usando como exemplo um modelo de risco para mortalidade cirúrgica. **Revista brasileira de epidemiologia**, São Paulo. Vol. 13, no. 4 (dez. 2010), p. 596-606., 2010.

OLIVEIRA, Elaine Fernandes Viellas de; GAMA, Silvana Granado Nogueira da; SILVA, Cosme Marcelo Furtado Passos da. Gravidez na adolescência e outros fatores de risco para mortalidade fetal e infantil no Município do Rio de Janeiro, Brasil. **Cad. Saúde Pública**, v. 26, n. 3, p. 567-578, 2010.

PAULA, Gilberto Alvarenga. **Modelos de regressão: com apoio computacional**. São Paulo: IME-USP, 2004.

PESSOA, Djalma Galvão Carneiro; SILVA, Pedro Luis Nascimento. **Análise de dados amostrais complexos.** São Paulo: Associação Brasileira de Estatística, v. 112, 1998.

REZVAN, Panteha Hayati; LEE, Katherine J.; SIMPSON, Julie A. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. **BMC medical research methodology**, v. 15, n. 1, p. 30, 2015.

RUBIN, Donald B. Inference and missing data. **Biometrika**, p. 581-592, 1976.

SAMPAIO, Carlos Eduardo Moreno; NESPOLI, Vanessa. Índice de adequação idade-anos de escolaridade. **Revista Brasileira de Estudos Pedagógicos**, v. 85, n. 209-10-11, 2007.

SCHAFER, Joseph L. **Analysis of incomplete multivariate data.** Chapman & Hall , 1997.

SCHAFER, Joseph L.; GRAHAM, John W. Missing data: our view of the state of the art. **Psychological methods**, v. 7, n. 2, p. 147, 2002.

SILVA, Maria Joseane Cruz da. **Imputação múltipla: comparação e eficiência em experimentos multiambientais.** 2012. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agronomia "Luiz de Queiroz", Universidade de São Paulo.

SZWARCWALD, Célia Landmann; DAMACENA, Giseli Nogueira. Complex sampling design in population surveys: planning and effects on statistical data analysis. **Revista Brasileira de Epidemiologia**, v. 11, p. 38-45, 2008.

UENAL, Hatice; MAYER, Benjamin; DU PREL, Jean-Baptist. Choosing appropriate methods for missing data in medical research: A decision algorithm on methods for missing data. **Journal of Applied Quantitative Methods**, v. 9, n. 4, 2014.

VAN DER HEIJDEN, Geert JMG et al. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical

example. **Journal of clinical epidemiology**, v. 59, n. 10, p. 1102-1109, 2006.

VASCONCELLOS, Mauricio Teixeira Leite de et al. **Desenho da amostra Nascer no Brasil: Pesquisa Nacional sobre Parto e Nascimento.** 2014.

ZHANG, Paul. Multiple imputation: theory and method. **International Statistical Review**, v. 71, n. 3, p. 581-592, 2003.

APÊNDICE A – Tabelas com as estimativas das razões de chance e dos β.

Tabela A.1 – Comparação da estimativa para as estimativas das razões de chance simples do mecanismo PCA entre os três métodos vs dados reais.

Variáveis	Padrão ouro	Imputação Múltipla				Vizinho mais próximo		Escore de propensão
		M=5	M=10	M=5	M=10			
		Todas as variáveis	Socioeconômicas	Todas as variáveis	Socioeconômicas	Todas as variáveis	Socioeconômicas	Socioeconômicas
Sem a utilização do plan amostral								
Raça/cor								
Branca	1	1	1	1	1	1	1	1
Preta	1,44 (0,93;2,22)	1,27 (0,8;1,99)	1,44 (0,92;2,23)	1,39 (0,9;2,15)	1,29 (0,85;1,97)	1,37 (0,87;2,16)	1,32 (0,83;2,09)	1,36 (0,89;2,07)
Mistas	1,51 (1,18;1,93)	1,33 (1,05;1,69)	1,49 (1,18;1,89)	1,46 (1,16;1,84)	1,4 (1,11;1,76)	1,45 (1,13;1,86)	1,5 (1,16;1,92)	1,41 (1,10;1,79)
Escore socioeconômico								
Classe D+E	1	1	1	1	1	1	1	1
Classe C	0,81 (0,63;1,03)	0,81 (0,66;1)	0,83 (0,67;1,03)	0,84 (0,67;1,04)	0,88 (0,71;1,1)	0,8 (0,64;1)	0,78 (0,61;0,99)	0,87 (0,7;1,08)
Classe A+B	0,47 (0,33;0,69)	0,53 (0,38;0,74)	0,53 (0,38;0,75)	0,47 (0,33;0,68)	0,56 (0,4;0,78)	0,5 (0,35;0,7)	0,48 (0,34;0,69)	0,53 (0,39;0,74)
Adequações no pré-natal								
Inadequado	1	1	1	1	1	1	1	1
Parcialmente	1,03 (0,79;1,35)	0,99 (0,77;1,27)	0,89 (0,68;1,17)	1,01 (0,78;1,3)	0,93 (0,7;1,24)	0,98 (0,75;1,27)	0,92 (0,69;1,22)	1,04 (0,8;1,35)
Adequado	0,59 (0,39;0,87)	0,65 (0,44;0,95)	0,63 (0,43;0,93)	0,61 (0,41;0,92)	0,67 (0,45;1)	0,61 (0,4;0,93)	0,6 (0,39;0,91)	0,74 (0,51;1,08)
Escolaridade adequada								
Sim	1,09 (0,88;1,35)	1,04 (0,84;1,28)	1,11 (0,91;1,36)	1,11 (0,91;1,35)	1,09 (0,88;1,34)	1,1 (0,89;1,37)	1,12 (0,92;1,37)	1,09 (0,88;1,34)
Não	1	1	1	1	1	1	1	1
Com a utilização do plano amostral								
Raça/cor								
Branca	1	1	1	1	1	1	1	1
Preta	1,44 (0,93;2,22)	1,53 (1;2,35)	1,44 (0,95;2,17)	1,34 (0,86;2,1)	1,29 (0,82;2,03)	1,43 (0,91;2,25)	1,38 (0,89;2,15)	1,36 (0,87;2,12)
Parda/morena/mulata	1,51 (1,18;1,93)	1,54 (1,2;1,97)	1,49 (1,18;1,88)	1,43 (1,13;1,81)	1,38 (1,08;1,76)	1,53 (1,19;1,97)	1,46 (1,14;1,88)	1,46 (1,15;1,84)
Escore socioeconômico								
Classe D+E	1	1	1	1	1	1	1	1
Classe C	0,81 (0,63;1,03)	0,84 (0,68;1,04)	0,78 (0,62;0,98)	0,84 (0,68;1,03)	0,82 (0,65;1,03)	0,78 (0,62;0,99)	0,83 (0,65;1,05)	0,9 (0,73;1,12)
Classe A+B	0,47 (0,33;0,69)	0,5 (0,36;0,7)	0,54 (0,39;0,76)	0,63 (0,46;0,86)	0,57 (0,41;0,8)	0,49 (0,34;0,69)	0,51 (0,36;0,72)	0,68 (0,5;0,93)
Adequação do pré-natal								
Inadequado	1	1	1	1	1	1	1	1
Parcialmente	1,03 (0,79;1,35)	0,97 (0,75;1,26)	0,94 (0,71;1,24)	1 (0,76;1,3)	0,95 (0,74;1,22)	0,96 (0,73;1,25)	0,91 (0,69;1,21)	1 (0,78;1,28)
Adequado	0,59 (0,39;0,87)	0,66 (0,44;0,97)	0,68 (0,47;1)	0,64 (0,43;0,97)	0,65 (0,44;0,96)	0,63 (0,42;0,95)	0,58 (0,38;0,89)	0,71 (0,49;1,01)
Escolaridade adequada								
Sim	1,09 (0,88;1,35)	1,07 (0,87;1,33)	1,06 (0,85;1,31)	1,05 (0,86;1,3)	1,09 (0,89;1,34)	1,13 (0,92;1,4)	1,16 (0,95;1,42)	1,08 (0,88;1,34)
Não	1	1	1	1	1	1	1	1

Fonte: O próprio o autor (2018).

Tabela A.2 – Comparação da estimativa para os parâmetros β do mecanismo PCA entre os três métodos vs dados reais sem a utilização do plano amostral nas imputações.

Variáveis	padrão ouro	Imputação Múltipla				Vizinho mais próximo		Escore de propensão
		M=5	M=10	M=5	M=10			
		Todas as variáveis	Socioeconômicas	Todas as variáveis	Socioeconômicas	Todas as variáveis	Socioeconômicas	Socioeconômicas
Raça/cor								
Intercepto	338,44 (38,24;38,64)	38,41 (38,22;38,6)	38,46 (38,28;38,65)	38,5 (38,31;38,68)	38,49 (38,3;38,68)	38,45 (38,26;38,64)	38,47 (38,28;38,65)	38,39 (38,2;38,58)
Branca	1	1	1	1	1	1	1	1
Preta	-0,13 (-0,75;0,49)	-0,02 (-0,62;0,57)	-0,16 (-0,78;0,45)	-0,17 (-0,76;0,43)	-0,27 (-0,88;0,34)	-0,12 (-0,71;0,47)	-0,2 (-0,82;0,43)	-0,1 (-0,71;0,5)
Mistas	-0,06 (-0,33;0,2)	-0,03 (-0,28;0,21)	-0,13 (-0,39;0,12)	-0,17 (-0,44;0,09)	-0,19 (-0,43;0,06)	-0,08 (-0,34;0,18)	-0,08 (-0,34;0,17)	-0,07 (-0,33;0,18)
Escore socioeconômico								
Intercepto	338,31 (38,07;38,56)	38,38 (38,14;38,62)	38,3 (38,08;38,53)	38,33 (38,09;38,57)	38,31 (38,07;38,55)	38,37 (38,14;38,6)	38,34 (38,11;38,57)	38,31 (38,07;38,55)
Classe D+E	1	1	1	1	1	1	1	1
Classe C	0,12 (-0,17;0,41)	-0,01 (-0,3;0,28)	0,09 (-0,18;0,35)	0,05 (-0,23;0,34)	0,06 (-0,21;0,34)	0,04 (-0,23;0,3)	0,07 (-0,19;0,34)	0,05 (-0,22;0,32)
Classe A+B	0,11 (-0,17;0,39)	0,07 (-0,2;0,34)	0,13 (-0,14;0,41)	0,08 (-0,18;0,35)	0 (-0,29;0,29)	0,04 (-0,23;0,3)	0,08 (-0,18;0,34)	0 (-0,28;0,28)
Adequação do pn								
Intercepto	38,21 (37,89;38,54)	38,23 (38,08;38,38)	38,22 (38,07;38,37)	38,25 (38,1;38,4)	38,23 (38,08;38,38)	38,23 (38,09;38,38)	38,25 (38,1;38,4)	38,2 (38,05;38,35)
Inadequadas	1	1	1	1	1	1	1	1
Parcialmente adequadas	0,46 (0,11;0,81)	0,45 (0,23;0,67)	0,38 (0,18;0,59)	0,33 (0,11;0,54)	0,28 (0,06;0,5)	0,49 (0,28;0,69)	0,42 (0,21;0,62)	0,36 (0,16;0,56)
Adequadas	0,18 (-0,18;0,55)	0,67 (0,35;0,98)	0,64 (0,32;0,96)	0,59 (0,27;0,91)	0,56 (0,24;0,88)	0,62 (0,3;0,94)	0,61 (0,29;0,92)	0,63 (0,31;0,96)
Escolaridade adequada								
Intercepto	38,41 (38,26;38,57)	38,43 (38,28;38,57)	38,42 (38,28;38,56)	38,4 (38,25;38,55)	38,39 (38,24;38,53)	38,42 (38,28;38,57)	38,44 (38,3;38,59)	38,39 (38,24;38,53)
Sim	-0,08 (-0,32;0,15)	-0,13 (-0,36;0,1)	-0,17 (-0,42;0,07)	-0,09 (-0,32;0,15)	-0,13 (-0,36;0,11)	-0,09 (-0,33;0,14)	-0,15 (-0,38;0,08)	-0,15 (-0,39;0,09)
Não	1	1	1	1	1	1	1	1
Bebeu								
Intercepto	338,39 (38,25;38,54)	38,38 (38,24;38,52)	38,37 (38,23;38,5)	38,37 (38,24;38,5)	38,34 (38,2;38,48)	38,39 (38,25;38,52)	38,39 (38,26;38,53)	38,34 (38,21;38,48)
Sim	-0,04 (-0,38;0,31)	0,03 (-0,3;0,36)	-0,02 (-0,37;0,32)	0,02 (-0,33;0,38)	0,1 (-0,26;0,45)	0,05 (-0,3;0,4)	0,02 (-0,34;0,38)	-0,05 (-0,4;0,29)
Não	1	1	1	1	1	1	1	1
Fumou								
Intercepto	38,39 (38,26;38,53)	38,39 (38,26;38,52)	38,37 (38,24;38,5)	38,38 (38,25;38,51)	38,34 (38,21;38,47)	38,39 (38,26;38,52)	38,39 (38,27;38,52)	38,35 (38,22;38,48)
Sim	-0,04 (-0,27;0,19)	-0,06 (-0,31;0,18)	0 (-0,25;0,25)	-0,04 (-0,27;0,19)	0,06 (-0,16;0,28)	0,05 (-0,2;0,29)	0,03 (-0,21;0,27)	0,1 (-0,25;0,46)
Não	1	1	1	1	1	1	1	1
Intenção da gravidez								
Intercepto	38,4 (38,2;38,6)	38,35 (38,15;38,54)	38,34 (38,16;38,53)	38,37 (38,18;38,56)	38,33 (38,14;38,52)	38,4 (38,21;38,58)	38,39 (38,2;38,59)	38,28 (38,09;38,48)
Queria engravidar	1	1	1	1	1	1	1	1
Queria esperar	0,05 (-0,19;0,29)	0,13 (-0,12;0,37)	0,1 (-0,14;0,34)	0,08 (-0,16;0,33)	0,11 (-0,13;0,36)	0,06 (-0,18;0,31)	0,06 (-0,18;0,3)	0,16 (-0,08;0,4)
Não queria engravidar	-0,1 (-0,37;0,16)	-0,02 (-0,27;0,23)	-0,04 (-0,3;0,23)	-0,08 (-0,34;0,19)	-0,07 (-0,33;0,2)	-0,09 (-0,35;0,17)	-0,07 (-0,33;0,19)	-0,01 (-0,27;0,26)

Fonte: O próprio o autor (2018).

Tabela A.3 – Comparação da estimativa para os parâmetros β do mecanismo PCA entre os três métodos vs dados reais com a utilização do plano amostral nas imputações,

Variáveis	padrão ouro	Imputação Múltipla				Vizinho mais próximo		Escore de propensão
		M=5	M=10	M=5	M=10			
		Todas as variáveis	Socioeconômicas	Todas as variáveis	Socioeconômicas	Todas as variáveis	Socioeconômicas	Socioeconômicas
Raça/cor								
Intercepto	38,44 (38,24;38,64)	38,46 (38,26;38,65)	38,45 (38,25;38,66)	38,47 (38,29;38,66)	38,45 (38,26;38,65)	38,43 (38,24;38,63)	38,45 (38,26;38,64)	38,44 (38,25;38,64)
Branca	1	1	1	1	1	1	1	1
Preta	-0,13 (-0,75;0,49)	-0,16 (-0,78;0,46)	-0,16 (-0,76;0,45)	-0,21 (-0,82;0,4)	-0,12 (-0,75;0,51)	-0,08 (-0,67;0,51)	-0,07 (-0,67;0,54)	-0,12 (-0,73;0,48)
Mistas	-0,06 (-0,33;0,2)	-0,12 (-0,37;0,14)	-0,09 (-0,35;0,16)	-0,11 (-0,37;0,15)	-0,09 (-0,35;0,17)	-0,05 (-0,3;0,2)	-0,09 (-0,34;0,16)	-0,12 (-0,38;0,14)
Escore socioeconômico								
Intercepto	38,31 (38,07;38,56)	38,31 (38,08;38,53)	38,3 (38,04;38,56)	38,35 (38,13;38,57)	38,37 (38,12;38,61)	38,35 (38,12;38,58)	38,36 (38,13;38,59)	38,34 (38,12;38,57)
Classe D+E	1	1	1	1	1	1	1	1
Classe C	0,12 (-0,17;0,41)	0,09 (-0,18;0,36)	0,14 (-0,16;0,44)	0,06 (-0,2;0,33)	0,03 (-0,26;0,33)	0,07 (-0,19;0,34)	0,03 (-0,25;0,3)	0,03 (-0,24;0,3)
Classe A+B	0,11 (-0,17;0,39)	0,1 (-0,17;0,37)	0,06 (-0,25;0,37)	0,04 (-0,2;0,29)	0,03 (-0,25;0,31)	0,07 (-0,19;0,33)	0,08 (-0,19;0,34)	-0,02 (-0,28;0,25)
Orientações no pn								
Intercepto	38,21 (37,89;38,54)	38,21 (38,06;38,36)	38,23 (38,07;38,38)	38,26 (38,11;38,4)	38,27 (38,12;38,42)	38,23 (38,08;38,38)	38,23 (38,08;38,38)	38,23 (38,08;38,38)
Inadequado	1	1	1	1	1	1	1	1
Parcialmente	0,46 (0,11;0,81)	0,47 (0,26;0,68)	0,41 (0,21;0,62)	0,39 (0,18;0,59)	0,24 (0;0,47)	0,49 (0,28;0,69)	0,46 (0,26;0,66)	0,39 (0,18;0,59)
Adequado	0,18 (-0,18;0,55)	0,65 (0,33;0,97)	0,69 (0,38;1)	0,55 (0,24;0,87)	0,62 (0,31;0,94)	0,63 (0,31;0,95)	0,59 (0,28;0,91)	0,48 (0,14;0,82)
Escolaridade adequada								
Intercepto	38,41 (38,26;38,57)	38,4 (38,26;38,55)	38,42 (38,27;38,57)	38,43 (38,29;38,57)	38,42 (38,28;38,57)	38,43 (38,29;38,58)	38,43 (38,28;38,57)	38,4 (38,25;38,55)
Sim	-0,08 (-0,32;0,15)	-0,11 (-0,35;0,12)	-0,11 (-0,34;0,12)	-0,14 (-0,37;0,1)	-0,12 (-0,36;0,12)	-0,11 (-0,35;0,12)	-0,14 (-0,37;0,1)	-0,13 (-0,36;0,11)
Não	1	1	1	1	1	1	1	1
Bebeu								
Intercepto	38,39 (38,25;38,54)	38,37 (38,24;38,5)	38,37 (38,23;38,52)	38,4 (38,26;38,53)	38,39 (38,25;38,53)	38,39 (38,26;38,53)	38,38 (38,25;38,52)	38,36 (38,22;38,49)
Sim	-0,04 (-0,38;0,31)	-0,03 (-0,38;0,31)	0,04 (-0,3;0,39)	-0,06 (-0,38;0,27)	-0,07 (-0,43;0,28)	0,02 (-0,33;0,37)	0,01 (-0,35;0,37)	0,01 (-0,34;0,36)
Não	1	1	1	1	1	1	1	1
Fumou								
Intercepto	38,39 (38,26;38,53)	38,37 (38,24;38,5)	38,4 (38,27;38,54)	38,39 (38,26;38,52)	38,39 (38,26;38,52)	38,4 (38,27;38,52)	38,38 (38,26;38,51)	38,36 (38,23;38,49)
Sim	-0,04 (-0,27;0,19)	0,03 (-0,2;0,26)	-0,16 (-0,39;0,07)	0,01 (-0,21;0,23)	-0,03 (-0,29;0,23)	0,03 (-0,22;0,27)	0,02 (-0,22;0,27)	0,28 (-0,07;0,64)
Não	1	1	1	1	1	1	1	1
Intenção da gravidez								
Intercepto	38,4 (38,2;38,6)	38,37 (38,18;38,56)	38,33 (38,14;38,52)	38,38 (38,19;38,57)	38,39 (38,2;38,58)	38,4 (38,21;38,59)	38,4 (38,21;38,59)	38,35 (38,16;38,54)
Queria engravidar	1	1	1	1	1	1	1	1
Queria esperar	0,05 (-0,19;0,29)	0,07 (-0,17;0,32)	0,14 (-0,1;0,38)	0,09 (-0,14;0,33)	0,05 (-0,18;0,29)	0,05 (-0,19;0,29)	0,04 (-0,2;0,27)	0,05 (-0,19;0,29)
Não queria engravidar	-0,1 (-0,37;0,16)	-0,08 (-0,35;0,18)	0,02 (-0,24;0,27)	-0,08 (-0,34;0,19)	-0,07 (-0,32;0,18)	-0,08 (-0,33;0,18)	-0,09 (-0,35;0,17)	-0,04 (-0,3;0,22)

Fonte: O próprio o autor (2018).

Tabela A.4 – Comparação da estimativa para as razões de chance simples do mecanismo PNCA entre os três métodos vs dados reais sem a utilização do plano amostral nas imputações,

Variáveis	padrão ouro	Imputação Múltipla				Vizinho mais próximo		Escore de propensão
		M=5	M=10	M=5	M=10			
		Todas as variáveis	Socioeconômicas	Todas as variáveis	Socioeconômicas	Todas as variáveis	Socioeconômicas	Socioeconômicas
Sem a utilização do plano amostral								
Orientações no pré-natal								
Inadequadas	1	1	1	1	1	1	1	1
Parcialmente adequadas	1,03 (0,79;1,35)	0,95 (0,72;1,26)	0,88 (0,67;1,16)	1,14 (0,87;1,49)	0,95 (0,73;1,24)	0,96 (0,73;1,27)	0,97 (0,73;1,28)	1,07 (0,82;1,4)
Adequadas	0,59 (0,39;0,87)	0,58 (0,41;0,83)	0,64 (0,46;0,9)	0,77 (0,53;1,12)	0,66 (0,47;0,92)	0,56 (0,37;0,86)	0,65 (0,42;1,00)	0,72 (0,52;1,01)
Com a utilização do plano amostral								
Orientações no pré-natal								
Inadequadas	1	1	1	1	1	1	1	1
Parcialmente adequadas	1,03 (0,79;1,35)	0,99 (0,77;1,27)	0,97 (0,74;1,27)	0,98 (0,75;1,28)	1,01 (0,78;1,3)	1,04 (0,79;1,36)	1,02 (0,79;1,3)	1,09 (0,84;1,41)
Adequadas	0,59 (0,39;0,87)	0,6 (0,4;0,89)	0,69 (0,48;0,99)	0,77 (0,54;1,11)	0,65 (0,47;0,94)	0,64 (0,41;0,99)	0,74 (0,48;1,14)	0,75 (0,53;1,07)

Fonte: O próprio o autor (2018).

Tabela A.5 – Comparação da estimativa para os parâmetros β do mecanismo PNCA entre os três métodos vs dados reais sem a utilização do plano amostral nas imputações,

Variáveis	padrão ouro	Imputação Múltipla				Vizinho mais próximo		Escore de propensão
		M=5	M=10	M=5	M=10			
		Todas as variáveis	Socioeconômicas	Todas as variáveis	Socioeconômicas	Todas as variáveis	Socioeconômicas	Socioeconômicas
Sem a utilização do plano amostral								
Orientações no pn								
Intercepto	38,21 (37,89;38,54)	38,19 (38,04;38,34)	38,22 (38,06;38,37)	38,28 (38,11;38,44)	38,19 (38,04;38,34)	38,31 (38,17;38,45)	38,33 (38,19;38,48)	38,26 (38,12;38,41)
Inadequadas	1	1	1	1	1	1	1	1
Parcialmente adequadas	0,46 (0,11;0,81)	0,35 (0,16;0,55)	0,43 (0,22;0,65)	0,22 (0;0,43)	0,25 (0,03;0,46)	0,36 (0,15;0,58)	0,25 (0,02;0,48)	0,33 (0,13;0,53)
Adequadas	0,18 (-0,18;0,55)	0,44 (0,09;0,79)	0,51 (0,22;0,8)	0,26 (-0,05;0,57)	0,5 (0,21;0,78)	0,46 (0,14;0,79)	0,34 (0,02;0,67)	0,22 (-0,22;0,67)
Intenção da gravidez								
Intercepto	38,40 (38,2;38,6)	38,36 (38,17;38,55)	38,3 (38,11;38,5)	38,27 (38,06;38,49)	38,22 (38;38,44)	38,38 (38,19;38,57)	38,46 (38,27;38,65)	38,3 (38,09;38,51)
Queria engravidar	1	1	1	1	1	1	1	1
Queria esperar	0,05 (-0,19;0,29)	-0,03 (-0,31;0,25)	0,09 (-0,17;0,35)	0,08 (-0,2;0,37)	0,07 (-0,2;0,35)	0,05 (-0,21;0,3)	-0,05 (-0,3;0,2)	0,15 (-0,12;0,42)
Não queria engravidar	-0,10 (-0,37;0,16)	-0,14 (-0,44;0,15)	0,06 (-0,22;0,33)	0,14 (-0,12;0,4)	0,15 (-0,1;0,41)	0,09 (-0,17;0,36)	-0,09 (-0,35;0,18)	-0,01 (-0,3;0,29)
Com a utilização do plano amostral								
Orientações no pn								
Intercepto	38,21 (37,89;38,54)	38,15 (38,01;38,3)	38,15 (38,01;38,3)	38,3 (38,17;38,44)	38,24 (38,09;38,39)	38,33 (38,21;38,46)	38,34 (38,19;38,48)	38,3 (38,16;38,45)
Inadequadas	1	1	1	1	1	1	1	1
Parcialmente adequadas	0,46 (0,11;0,81)	0,42 (0,22;0,62)	0,44 (0,24;0,64)	0,18 (-0,07;0,43)	0,29 (0,07;0,5)	0,22 (-0,01;0,45)	0,29 (0,11;0,46)	0,08 (-0,19;0,35)
Adequadas	0,18 (-0,18;0,55)	0,43 (0,13;0,74)	0,53 (0,25;0,81)	0,19 (-0,11;0,49)	0,32 (0,01;0,63)	0,45 (0,13;0,77)	0,32 (-0,03;0,68)	0,33 (0;0,66)
Intenção da gravidez								
Intercepto	38,4 (38,2;38,6)	38,32 (38,13;38,51)	38,27 (38,07;38,47)	38,36 (38,17;38,55)	38,4 (38,21;38,58)	38,4 (38,22;38,59)	38,45 (38,26;38,63)	38,33 (38,13;38,52)
Queria engravidar	1	1	1	1	1	1	1	1
Queria esperar	0,05 (-0,19;0,29)	0,06 (-0,18;0,29)	0,07 (-0,2;0,33)	0,01 (-0,24;0,25)	-0,11 (-0,37;0,16)	0,04 (-0,2;0,28)	0 (-0,25;0,24)	0,14 (-0,1;0,38)
Não queria engravidar	-0,1 (-0,37;0,16)	-0,19 (-0,49;0,11)	0 (-0,26;0,27)	-0,01 (-0,26;0,23)	-0,1 (-0,4;0,19)	-0,01 (-0,26;0,24)	-0,09 (-0,35;0,17)	-0,05 (-0,3;0,21)

Fonte: O próprio o autor (2018).

Tabela A.6 – Comparação da estimativa para as razões de chance simples do mecanismo PNA entre os três métodos vs dados reais sem a utilização do plano amostral nas imputações,

Sem a utilização do plano amostral								
Variáveis	padrão ouro	Imputação Múltipla				Vizinho mais próximo		Escore de propensão
		M=5	M=10	M=5	M=10			
		Todas as variáveis	Socioeconômicas	Todas as variáveis	Socioeconômicas	Todas as variáveis	Socioeconômicas	Socioeconômicas
Orientações no pré-natal								
Inadequadas	1	1	1	1	1	1	1	1
Parcialmente adequadas	1,03 (0,79;1,35)	1,01 (0,79;1,31)	0,96 (0,74;1,26)	0,93 (0,72;1,22)	1,11 (0,84;1,45)	0,96 (0,73;1,26)	1,07 (0,81;1,41)	1,06 (0,82;1,36)
Adequadas	0,59 (0,39;0,87)	0,61 (0,41;0,91)	0,7 (0,49;1)	0,6 (0,39;0,92)	0,74 (0,52;1,07)	0,64 (0,42;0,98)	0,65 (0,42;1,01)	0,71 (0,49;1,04)
Com a utilização do plano amostral								
Orientações no pré-natal								
Inadequadas	1	1	1	1	1	1	1	1
Parcialmente adequadas	1,03 (0,79;1,35)	0,99 (0,77;1,27)	0,97 (0,74;1,27)	0,98 (0,75;1,28)	1,01 (0,78;1,3)	1,04 (0,79;1,36)	1,02 (0,79;1,3)	1,09 (0,84;1,41)
Adequadas	0,59 (0,39;0,87)	0,6 (0,4;0,89)	0,69 (0,48;0,99)	0,77 (0,54;1,11)	0,65 (0,47;0,94)	0,64 (0,41;0,99)	0,74 (0,48;1,14)	0,75 (0,53;1,07)

Fonte: O próprio o autor (2018).

Tabela A.7 – Comparação da estimativa para os parâmetros β do mecanismo PNA entre os três métodos vs dados reais sem a utilização do plano amostral nas imputações.

Variáveis	padrão ouro	Imputação Múltipla				vizinho mais próximo		Escore de propensão
		M=5	M=10	M=5	M=10			
		Todas as variáveis	Socioeconômicas	Todas as variáveis	Socioeconômicas	Todas as variáveis	Socioeconômicas	Socioeconômicas
Sem a utilização do plano amostral								
Orientações no pré-natal								
Intercepto	38,21 (37,89;38,54)	38,24 (38,08;38,39)	38,23 (38,07;38,39)	38,26 (38,11;38,42)	38,27 (38,11;38,42)	38,25 (38,09;38,4)	38,28 (38,13;38,43)	38,29 (38,15;38,43)
Inadequadas	1	1	1	1	1	1	1	1
Parcialmente adequadas	0,46 (0,11;0,81)	0,35 (0,15;0,56)	0,41 (0,19;0,62)	0,29 (0,08;0,5)	0,28 (0,04;0,52)	0,39 (0,18;0,61)	0,25 (0,05;0,45)	0,21 (0;0,42)
Adequadas	0,18 (-0,18;0,55)	0,76 (0,45;1,07)	0,73 (0,42;1,04)	0,6 (0,29;0,91)	0,61 (0,29;0,92)	0,73 (0,37;1,09)	0,65 (0,31;1)	0,53 (0,15;0,9)
Bebeu durante a gestação								
Intercepto	38,39 (38,25;38,54)	38,41 (38,28;38,55)	38,41 (38,27;38,55)	38,40 (38,25;38,55)	38,41 (38,26;38,56)	38,41 (38,26;38,55)	38,4 (38,26;38,55)	38,40 (38,26;38,54)
Sim	-0,04 (-0,38;0,31)	-0,16 (-0,44;0,11)	-0,17 (-0,47;0,12)	-0,09 (-0,39;0,22)	-0,14 (-0,44;0,16)	-0,17 (-0,52;0,18)	-0,11 (-0,45;0,22)	-0,11 (-0,4;0,18)
Não	1	1	1	1	1	1	1	1
Fumou durante a gestação								
Intercepto	338,39 (38,26;38,53)	38,4 (38,26;38,53)	38,42 (38,3;38,55)	38,39 (38,26;38,53)	38,4 (38,26;38,53)	38,4 (38,26;38,53)	38,4 (38,26;38,53)	38,39 (38,25;38,53)
Sim	-0,04 (-0,27;0,19)	-0,10 (-0,34;0,14)	-0,37 (-0,76;0,02)	-0,03 (-0,27;0,21)	-0,07 (-0,35;0,2)	-0,09 (-0,35;0,17)	-0,09 (-0,34;0,17)	-0,04 (-0,28;0,20)
Não	1	1	1	1	1	1	1	1
Com a utilização do plano amostral								
Adequação do pré-natal								
Intercepto	38,21 (37,89;38,54)	38,23 (38,08;38,39)	38,23 (38,07;38,38)	38,26 (38,11;38,41)	38,27 (38,11;38,43)	38,25 (38,1;38,41)	38,33 (38,09;38,57)	38,26 (38,1;38,42)
Inadequado	1	1	1	1	1	1	1	1
Parcialmente adequado	0,46 (0,11;0,81)	0,43 (0,22;0,63)	0,43 (0,24;0,63)	0,31 (0,11;0,5)	0,26 (0,05;0,48)	0,41 (0,15;0,54)	0,04 (-0,25;0,32)	0,35 (0,15;0,56)
Adequado	0,18 (-0,18;0,55)	0,67 (0,36;0,99)	0,74 (0,43;1,05)	0,58 (0,28;0,87)	0,59 (0,28;0,9)	0,71 (0,36;1,06)	0,15 (-0,22;0,51)	0,51 (0,18;0,84)
Bebeu durante a gestação								
Intercepto	38,39 (38,25;38,54)	38,4 (38,25;38,55)	38,41 (38,27;38,55)	38,14 (38,26;38,55)	38,4 (38,26;38,55)	38,4 (38,4;38,54)	38,35 (38,14;38,56)	38,39 (38,24;38,54)
Sim	-0,04 (-0,38;0,31)	-0,11 (-0,43;0,2)	-0,14 (-0,44;0,15)	-0,15 (-0,46;0,16)	-0,12 (-0,41;0,17)	-0,11 (-0,45;0,24)	0,05 (-0,34;0,45)	-0,04 (-0,38;0,31)
Não	1	1	1	1	1	1	1	1
Fumou durante a gestação								
Intercepto	38,39 (38,26;38,53)	38,4 (38,27;38,54)	38,41 (38,28;38,55)	38,4 (38,26;38,53)	38,39 (38,26;38,52)	38,39 (38,25;38,53)	38,39 (38,18;38,6)	38,41 (38,28;38,54)
Sim	-0,04 (-0,27;0,19)	-0,12 (-0,37;0,13)	-0,18 (-0,44;0,08)	-0,1 (-0,35;0,15)	-0,01 (-0,26;0,24)	-0,05 (-0,31;0,2)	-0,44 (-0,84;-0,05)	-0,22 (-0,59;0,16)
Não	1	1	1	1	1	1	1	1

Fonte: O próprio o autor (2018).

ANEXO A – Aprovação do comitê ético em pesquisa para o projeto Nascer no Brasil.

Ministério da Saúde

FIOCRUZ

Fundação Oswaldo Cruz
Escola Nacional de Saúde Pública Sergio Arouca
Comitê de Ética em Pesquisa

ESCOLA NACIONAL DE SAÚDE PÚBLICA
SERGIO AROUCA
ENSP

Rio de Janeiro, 10 de junho de 2010.

O Comitê de Ética em Pesquisa da Escola Nacional de Saúde Pública Sergio Arouca – CEP/ENSP, constituído nos Termos da Resolução CNS nº 196/96 e, devidamente registrado na Comissão Nacional de Ética em Pesquisa - CONEP, recebeu, analisou e emitiu parecer sobre a documentação referente ao Protocolo de Pesquisa, conforme abaixo, discriminado:

PROTOCOLO DE PESQUISA CEP/ENSP - Nº 92/10
CAAE: 0096.0.031.000-10

Título do Projeto: "Nascer no Brasil: inquérito nacional sobre parto e nascimento (título inicial: Inquérito epidemiológico sobre as conseqüências da cesariana desnecessária no Brasil)"

Classificação no Fluxograma: Grupo III

Pesquisadora Responsável: Maria do Carmo Leal

Instituição onde se realizará: Escola Nacional de Saúde Pública Sergio Arouca - ENSP/Fiocruz

Data de recebimento no CEP-ENSP: 26 / 04 / 2010

Data de apreciação: 11 / 05 / 2010

Parecer do CEP/ENSP: Aprovado.

Ressaltamos que a pesquisadora responsável por este Protocolo de Pesquisa deverá apresentar a este Comitê de Ética um relatório das atividades desenvolvidas no período de 12 meses a contar da data de sua aprovação (*item VII.13.d., da resolução CNS/MS Nº 196/96*) de acordo com o modelo disponível na página do CEP/ENSP na internet.

Esclarecemos, que o CEP/ENSP deverá ser informado de quaisquer fatos relevantes (incluindo mudanças de método) que alterem o curso normal do estudo, devendo a pesquisadora justificar caso o mesmo venha a ser interrompido.

PROF. SERGIO REGO
Coordenador do Comitê de Ética em Pesquisa
CEP/ENSP

ANEXO B – Parecer consubstanciado do CEP

ESCOLA NACIONAL DE SAÚDE PÚBLICA SERGIO AROUCA - ENSP/ FIOCRUZ

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: COMPARAÇÃO DE MÉTODOS PARA TRATAMENTO DE DADOS FALTANTES EM INQUERITOS EPIDEMIOLÓGICOS COM AMOSTRAGEM COMPLEXA

Pesquisador: VANESSA EUFRAUZINO PACHECO

Área Temática:

Versão: 1

CAAE: 70939317.0.0000.5240

Instituição Proponente: FUNDACAO OSWALDO CRUZ

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 2.200.738

Apresentação do Projeto:

Parecer do projeto de pesquisa "Comparação de Métodos para Tratamento de Dados Faltantes em Inquéritos Epidemiológicos com Amostragem Complexa", de Vanessa Eufrauzino Pacheco, mestranda do Programa de Pós-graduação em Epidemiologia em Saúde Pública, orientada por Cleber Nascimento do Carmo, qualificado em 09/02/2017, com financiamento próprio no valor de R$ 860,00.

O projeto será desenvolvido com parte das informações coletadas na amostra de puérperas adolescentes do banco de dados da pesquisa Nascer no Brasil, estudo longitudinal com delineamento amostral complexo.

Segundo a pesquisadora, "Os inquéritos epidemiológicos com amostragem completa são bastante utilizados devido a redução de custo e propiciando o mesmo benefício que uma pesquisa censitária. Entretanto a ocorrência de perda de dados é um dos problemas podem afetar esses inquéritos, influenciando os resultados analíticos da pesquisa. [...] A técnica de imputação vem sendo muito utilizada ultimamente pela facilidade de implementação, pelo bom resultado que proporciona, pela exibilidade na utilização das variáveis, sejam numéricas, categóricas, desfecho,

Endereço: Rua Leopoldo Bulhões, 1480 - Térreo
Bairro: Manguinhos **CEP:** 21.041-210
UF: RJ **Município:** RIO DE JANEIRO
Telefone: (21)2598-2863 **Fax:** (21)2598-2863 **E-mail:** cep@ensp.fiocruz.br

preditores, etc. Este trabalho pretende ampliar a discussão sobre imputação de dados em estudos epidemiológicos, justificando-se pela importância de levar em consideração a amostragem complexa utilizada no desenho da pesquisa no processo de modelagem dos dados".
Na metodologia, a pesquisadora informa que "serão simulados três bancos de dados com 20% de perda, cada um com um mecanismo de falta de dados, sempre respeitando a complexidade dos dados. Para o primeiro banco, utilizando o mecanismo PCA, será realizada uma amostragem aleatória simples, todas as adolescentes têm a mesma probabilidade de inclusão, levando em consideração as estratificações da pesquisa. O segundo, utilizando o mecanismo PNCA, será realizada uma amostragem aleatória simples, dentro de um grupo específico, ou seja, será realizada a amostragem levando em consideração as estratificações da pesquisa e uma estratificação extra, classe social. Para o terceiro banco, utilizando o método de PNA, será realizada uma amostragem aleatória simples, dentro de um grupo específico cuja estratificação pertença a uma variável não presente ao banco de dados, se fumou ou bebeu durante a gravidez. Serão testados três métodos de tratamento de dados, análise de casos completos (aac), a imputação por escore de propensão e a imputação múltipla por média preditiva, a partir de regressões. Serão utilizadas as estimativas observadas para o conjunto de dados completo a partir de análises estatísticas tradicionais e comparadas às dos métodos escolhidos".

Tamanho da amostra: 4.571

Objetivo da Pesquisa:
Segundo a pesquisadora, o objetivo geral da pesquisa é "Avaliar, a partir de análises estatísticas, métodos de tratamento de dados faltantes que permitam estimações pouco viesadas em inquérito epidemiológico com amostragem complexa".

A pesquisadora informa como objetivos específicos "Descrever as principais estratégias de imputação de dados faltantes" e "Comparar os métodos de imputação descritos a partir de uma aplicação a dados reais de pesquisa epidemiológica com plano amostral complexo - Nascer no Brasil".

Avaliação dos Riscos e Benefícios:
Em relação aos potenciais riscos aos participantes, a pesquisadora esclarece que "O estudo será realizado utilizando dados secundários da base de dados do Nascer no Brasil,

sem uso de informações identificadoras tais como nome, endereço, nome da mãe".

Apresenta como benefício "Proporcionar métodos de ajustes para banco de dados provenientes de amostragem complexa, com perdas, para inquéritos epidemiológicos, seja por meio de imputação ou ponderação".

Comentários e Considerações sobre a Pesquisa:
O projeto está bem estruturado, com referencial teórico e considerações metodológicas adequadas.

A pesquisadora propôs dispensa do TCLE uma vez que irá utilizar parte do banco de dados de uma pesquisa já realizada, previamente aprovada pelo CEP/ENSP. O presente projeto utilizará uma parcela dos dados coletados, sem as informações que possibilitariam identificação pessoal dos participantes da pesquisa original.

Considerações sobre os Termos de apresentação obrigatória:
Foram apresentados os seguintes documentos na Plataforma Brasil:

- Folha de Rosto gerada pela Plataforma Brasil assinada pela pesquisadora responsável;
- Projeto de Pesquisa na íntegra, nomeado "projeto_mestrado.pdf", postado em 26/06/2017;
- Formulário de Encaminhamento nomeado "formulario_encaminhamento_orientador.pdf", postado em 26/06/2017;
- Cronograma, nomeado "Cronograma.docx", postado em 26/06/2017;
- PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_942055.pdf, postado em 26/06/2017;
- Planilha de orçamento, nomeado "orcamento.docx", postado em 26/06/2017;
- TCUD assinado pelo pesquisador responsável, nomeado "TCUD.pdf", postado em 26/06/2017;
- Parecer de aprovação pelo CEP/ENSP do projeto Nascer no Brasil, nomeado "aprovacao_uso_banco.docx", postado em 26/06/2017;
- Termo de autorização assinado e datado para fornecimento de banco de dados para uso na pesquisa em questão; nomeado "Autorizacao_uso_banco.docx", postado em 26/06/2017.

A pesquisadora solicitou dispensa do TCLE.

Recomendações:

Não há.

Conclusões ou Pendências e Lista de Inadequações:

O CEP/ENSP considera que o protocolo do projeto de pesquisa ora apresentado contempla os quesitos éticos necessários, estando apto a ser iniciado a partir da presente data de emissão deste parecer.

Considerações Finais a critério do CEP:

ATENÇÃO: ***CASO OCORRA ALGUMA ALTERAÇÃO NO FINANCIAMENTO DO PROJETO ORA APRESENTADO (ALTERAÇÃO DE PATROCINADOR, COPATROCÍNIO, MODIFICAÇÃO NO ORÇAMENTO), O PESQUISADOR TEM A RESPONSABILIDADE DE SUBMETER UMA EMENDA AO CEP SOLICITANDO AS ALTERAÇÕES NECESSÁRIAS. A NOVA FOLHA DE ROSTO A SER GERADA DEVERÁ SER ASSINADA NOS CAMPOS PERTINENTES E A VIA ORIGINAL DEVERÁ SER ENTREGUE NO CEP. ATENTAR PARA A NECESSIDADE DE ATUALIZAÇÃO DO CRONOGRAMA DA PESQUISA. CASO O PROJETO SEJA CONCORRENTE DE EDITAL, SOLICITA-SE ENCAMINHAR AO CEP, PELA PLATAFORMA BRASIL, COMO NOTIFICAÇÃO, O COMPROVANTE DE APROVAÇÃO. PARA ESTES CASOS, A LIBERAÇÃO PARA O INÍCIO DO TRABALHO DE CAMPO (COLETA DE DADOS, ABORDAGEM DE POSSÍVEIS PARTICIPANTES ETC.) ESTÁ CONDICIONADA À APRESENTAÇÃO DA FOLHA DE ROSTO, ASSINADA PELO PATROCINADOR, EM ATÉ 15 (QUINZE) DIAS APÓS A DIVULGAÇÃO DO RESULTADO DO EDITAL AO QUAL O PROJETO FOI SUBMETIDO.***

****

Verifique o cumprimento das observações a seguir:

1* Em atendimento a Resolução CNS nº 466/2012, cabe ao pesquisador responsável pelo presente estudo elaborar e apresentar ao CEP RELATÓRIOS PARCIAIS (semestrais) e FINAL. Os relatórios compreendem meio de acompanhamento pelos CEP, assim como outras estratégias de monitoramento, de acordo com o risco inerente à pesquisa. O relatório deve ser enviado pela Plataforma Brasil em forma de "notificação". Os modelos de relatórios (parciais e final) que devem ser utilizados encontram-se disponíveis na homepage do CEP/ENSP (www.ensp.fiocruz.br/etica).

2* Qualquer necessidade de modificação no curso do projeto deverá ser submetida à apreciação do CEP, como EMENDA. Deve-se aguardar parecer favorável do CEP antes de efetuar a/s modificação/ões.

3* Justificar fundamentadamente, caso haja necessidade de interrupção do projeto ou a não publicação dos resultados.

4* O Comitê de Ética em Pesquisa não analisa aspectos referentes a direitos de propriedade intelectual e ao uso de criações protegidas por esses direitos. Recomenda-se que qualquer consulta que envolva matéria de propriedade intelectual seja encaminhada diretamente pelo pesquisador ao Núcleo de Inovação Tecnológica da Unidade.

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_942055.pdf	26/06/2017 13:34:44		Aceito
Cronograma	Cronograma.docx	26/06/2017 13:34:25	VANESSA EUFRAUZINO PACHECO	Aceito
Orçamento	orcamento.docx	26/06/2017 13:26:00	VANESSA EUFRAUZINO PACHECO	Aceito
Outros	aprovacao_cep_nascer.pdf	26/06/2017 13:18:01	VANESSA EUFRAUZINO PACHECO	Aceito
Folha de Rosto	folha_rosto_assinada.pdf	26/06/2017 13:16:12	VANESSA EUFRAUZINO PACHECO	Aceito
Outros	TCUD.pdf	26/06/2017 13:11:02	VANESSA EUFRAUZINO PACHECO	Aceito
Declaração de Pesquisadores	formulario_encaminhamento_orientador.pdf	26/06/2017 12:03:37	VANESSA EUFRAUZINO PACHECO	Aceito
Outros	Autorizacao_uso_banco.docx	26/06/2017 12:02:24	VANESSA EUFRAUZINO PACHECO	Aceito
Projeto Detalhado	projeto_mestrado.pdf	26/06/2017	VANESSA	Aceito

Endereço: Rua Leopoldo Bulhões, 1480 - Térreo
Bairro: Manguinhos **CEP:** 21.041-210
UF: RJ **Município:** RIO DE JANEIRO
Telefone: (21)2598-2863 **Fax:** (21)2598-2863 **E-mail:** cep@ensp.fiocruz.br

/ Brochura Investigador	projeto_mestrado.pdf	10:07:42	EUFRAUZINO PACHECO	Aceito
Outros	Folhaderosto_VanessaEufrazino.pdf	03/08/2017 16:24:03	Jennifer Braathen Salgueiro	Aceito

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

RIO DE JANEIRO, 03 de Agosto de 2017

Assinado por:
Jennifer Braathen Salgueiro
(Coordenador)