Vitor Passos Camargos

IMPUTAÇÃO MÚLTIPLA E ANÁLISE DE CASOS COMPLETOS NO CONTEXTO DA SAÚDE PÚBLICA: UMA AVALIAÇÃO PRÁTICA DO IMPACTO DAS PERDAS NAS ANÁLISES

Universidade Federal de Minas Gerais
Programa de Pós-Graduação em Saúde Pública
Belo Horizonte - MG
2011

Vitor Passos Camargos

IMPUTAÇÃO MÚLTIPLA E ANÁLISE DE CASOS COMPLETOS NO CONTEXTO DA SAÚDE PÚBLICA: UMA AVALIAÇÃO PRÁTICA DO IMPACTO DAS PERDAS NAS ANÁLISES

Dissertação apresentada ao Programa de Pós-Graduação em Saúde Pública da Universidade Federal de Minas Gerais, como requisito parcial para obtenção de título de Mestre em Saúde Pública - área de concentração em Epidemiologia.

Orientador: Fernando Augusto Proietti

Co-orientadora: Cibele Comini César

Belo Horizonte - MG
2011

UNIVERSIDADE FEDERAL DE MINAS GERAIS

Reitor

Prof. Clélio Campolina Diniz

Vice-Reitora

Profª. Rocksane de Carvalho Norton

Pró-Reitora de Pós-Graduação

Profa. Antônia Vitória Soares Aranha

Pró-Reitor de Pesquisa

Prof. Renato de Lima Santos

FACULDADE DE MEDICINA

Diretor

Prof. Francisco José Penna

Chefe do Departamento de Medicina Preventiva e Social

Profª. Maria da Conceição Juste Werneck Côrtes

PROGRAMA DE PÓS-GRADUAÇÃO EM SAÚDE PÚBLICA

Coordenadora

Profª Mariângela Leal Cherchiglia

Sub-Coordenador

Prof. Mark Drew Crosland Guimarães

Colegiado do Programa de Pós-Graduação em Saúde Pública

Profª. Ada Ávila Assunção

Profª.Eli Iola Gurgel Andrade

Prof. Fernando Augusto Proietti

Prof. Francisco de Assis Acúrcio

Profª. Maria Fernanda Furtado de Lima e Costa

Profª. Soraya Almeida Belisário

Prof. Tarcísio Márcio Magalhães Pinheiro

Profª. Waleska Teixeira Caiaffa

Aline Dayrell Ferreira (Representante Titular - Doutorado)

Graziella Lage Oliveira (Representante Suplente - Doutorado)

Orozimbo Henriques Campos Neto (Representante Titular - Mestrado)

Gustavo Laine Araujo de Oliveira (Representante Suplente - Mestrado)

FACULDADE DE MEDICINA
CENTRO DE PÓS-GRADUAÇÃO
Av. Prof. Alfredo Balena 190 / sala 533
Belo Horizonte - MG - CEP 30.130-100
Fone: (031) 3409.9641 FAX: (31) 3409.9640

UFMG

DECLARAÇÃO

A Comissão Examinadora abaixo assinada, composta pelos Professores Doutores: Fernando Augusto Proietti, Cibele Comini César, Ilka Afonso Reis e Waleska Teixeira Caiaffa, aprovou a defesa da dissertação intitulada **"IMPUTAÇÃO MÚLTIPLA E ANÁLISE DE CASOS COMPLETOS NO CONTEXTO DA SAÚDE PÚBLICA: UMA AVALIAÇÃO PRÁTICA DO IMPACTO DAS PERDAS NAS ANÁLISES"** apresentada pelo aluno **VITOR PASSOS CAMARGOS,** para obtenção do título de Mestre em Saúde Pública, pelo Programa de Pós-Graduação em Saúde Pública - Área de Concentração em Epidemilogia, da Faculdade de Medicina da Universidade Federal de Minas Gerais, realizada em 18 de fevereiro de 2011.

Prof. Fernando Augusto Proietti
Orientador

Profa. Cibele Comini César
Coorientadora

Profa. Ilka Afonso Reis

Profa. Waleska Teixeira Caiaffa

Dedico essa dissertação a todos os professores que participaram da minha formação e em especial à minha família.

Agradecimentos

Aos meus orientadores, Fernando e Cibele, que desde a graduação me incentivaram na academia, pelas aulas, exemplo ético, serenidade e paciência. Especialmente à amizade e confiança depositada desde a iniciação científica.

A professora Waleska, sempre presente na minha formação, nas infindáveis reuniões do Grupo de Pesquisa em Epidemiologia e outras, que sempre foram uma sala de aula muito especial, onde todos têm voz e contando sempre com momentos divertidos.

Ao Professor César, que assim como a Waleska sempre me acompanhou, pelo companheirismo, pela atenção e disposição para ajudar em qualquer problema.

Aos colegas que tive o prazer de conhecer no Grupo de Pesquisa, Eulilian, Adriana, Aline Dayrell, Mery, Fabiane, Marcela, Amanda, Grazi, Michelle Ralil, Elaine, Janaína, Guta, Clareci, Lete, Rosana e todos os bolsistas de iniciação científica. À nossa amizade e convivência sempre animadora nesse grupo muito especial.

À Jussara, que me indicou para a bolsa de iniciação científica e à Marcelina, minha professora de matemática do colégio Arthur Versiane Veloso.

Ao Programa da Pós-Graduação em Saúde Pública.

À Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), que me concedeu a bolsa de mestrado.

Aos amigos,

Aos acasos da vida, que de alguma forma sempre me deram uma mãozinha.

Em especial, ao meu pai, Regina, Dora, e aos meus irmãos, Maíra, Arthur e Otávio.

RESUMO

Pesquisadores da área da saúde lidam frequentemente com o problema das bases de dados incompletas. A Análise de Casos Completos (ACC), que restringe as análises aos indivíduos com dados completos, reduz o tamanho da amostra e pode produzir estimativas viciadas. Baseado em fundamentos estatísticos, o método de Imputação Múltipla (IM) utiliza todos os dados coletados e é recomendado como alternativa a ACC. Dados do estudo Saúde em Beagá, inquérito domiciliar em que participaram 4048 adultos de dois dos nove distritos sanitários da cidade de Belo Horizonte no biênio 2008-2009, foram utilizados para avaliar a ACC e diferentes abordagens de IM no contexto de modelos logísticos com covariáveis incompletas. O Índice de Massa Corporal (IMC), um indicador de grande relevância na saúde pública, foi obtido por meio de medidas auto-referidas e posteriormente também por aferições diretas do peso e altura dos participantes. As medidas auto-referidas apresentaram um elevado percentual de perdas, que se propagaram para o IMC baseado nas mesmas. No entanto, perdas mínimas nas medidas aferidas permitiram o calculo do IMC para praticamente toda a amostra. Dada essa particularidade do estudo, a situação hipotética em que os dados ausentes do IMC, baseado nas medidas auto-referidas, são recuperados pôde ser aproximada por um procedimento simples. Os métodos de ACC e diferentes abordagens de IM foram aplicados no contexto em que o IMC, ainda com perdas, é uma das covariáveis de uma regressão logística. Os resultados desses métodos foram então comparados com os resultados posteriores à recuperação dos dados ausentes. Verificou-se que mesmo a abordagem mais simplista de IM obteve melhor desempenho que a ACC, já que se aproximou mais aos resultados pós-recuperação.

Palavras-chave: métodos, análise estatística, modelos logísticos, índice de massa corporal

ABSTRACT

Researchers in the health field often deal with the problem of incomplete databases. Complete Case Analysis (CCA), which restricts the analysis to subjects with complete data, reduces the sample size and may result in biased estimates. Based on statistical grounds, the Multiple Imputation (MI) method uses all collected data and is recommended as an alternative to CCA. Data from the study "Saúde em Beagá", attended by 4048 adults from two of nine health districts in the city of Belo Horizonte in 2008-2009, were used to evaluate CCA and different MI approaches in the context of logistic models with incomplete covariates. The Body Mass Index (BMI), an indicator of high public health relevance, was obtained through self-reported measures, and subsequently by direct measurements of height and weight of participants. The self-reported measures showed high percentage of missing data, which have spread to BMI based on them. However the minimum losses in direct measurements allowed BMI calculation for virtually entire sample. Given this peculiarity of the study, a hypothetical situation in which the missing data for BMI based on self-reported measures are recovered could be approached by a simple procedure. The methods of ACC and different approaches to IM were applied in a context where the BMI, with missing data, is one of the covariates in a logistic regression. The results of these methods were then compared with the results after the missing data recovery. It was found that even the more simplistic MI approach performed better than CCA since it was closer to the post-recovery results.

Key words: methods, statistical analysis, logistic models, body mass index

SUMÁRIO

1 CONSIDERAÇÕES INICIAIS

Graduado em Estatística e membro do Grupo de Pesquisa em Epidemiologia (GPE/UFMG) desde 2006, e posteriormente do Observatório de Saúde Urbana de Belo Horizonte (OSUBH/GPE/UFMG), sempre estive em contato com bases de dados de grandes inquéritos de Saúde. Um de meus primeiros estudos no grupo, como parte da monografia do curso de graduação, envolveu os dados do Inquérito de Saúde dos Adultos na Região Metropolitana de Belo Horizonte[1] (suplementar à Pesquisa de Emprego e Desemprego na Região Metropolitana de Belo Horizonte) cuja coleta foi realizada no ano de 2003. A seção metodológica dessa monografia registra "A restrição de idade (≥ 20 anos) e perdas em algumas das variáveis limitaram a amostra final à 13058 indivíduos...". Essa redução da amostra ocasionada por perdas nas variáveis do estudo, que foi uma constante nos trabalhos que pude acompanhar no grupo, é um fato comum nas pesquisas da área da saúde e de outras áreas humanas.

Os possíveis problemas da restrição das análises aos indivíduos com dados completos, método denominado Análise de Casos Completos (ACC), foram-me apontados pela primeira vez pelos professores Fernando Augusto Proietti e Cibele Comini César, meus atuais orientadores. Nos estudos do GPE, como procedimento padrão anterior às análises, quando um grupo numeroso de indivíduos era ignorado devido a perdas nas variáveis do estudo, esse grupo era comparado ao daqueles indivíduos com dados completos por meio de características sócio-demográficas observadas. Diferenças entre esses grupos indicam que a sub-amostra de indivíduos com dados completos não é representativa da amostra total do estudo, e assim os resultados das análises baseadas nessa sub-amostra podem estar comprometidos. Sob o ponto de vista estatístico, pode-se dizer nesse caso que a perda não ocorre de forma completamente aleatória na amostra, uma suposição geralmente necessária à ACC.

---

[1] Lima-Costa MF. A escolaridade afeta, igualmente, comportamentos prejudiciais à saúde de idosos e adultos mais jovens?: Inquérito de Saúde da Região Metropolitana de Belo Horizonte, Minas Gerais, Brasil. Epidemiol. Serv. Saúde [periódico na Internet]. 2004 Dez [citado 2011 Jan 20] ; 13(4): 201-208. Disponível em: http://scielo.iec.pa.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742004000400002&lng=pt.

Foi nesse contexto que naturalmente surgiu, em 2008, o interesse pelo tema do presente estudo e a proposta de revisão das metodologias de análise de dados incompletos. No ano de 2009, me tornei aluno regular do mestrado do Programa de Pós-Graduação em Saúde Pública e iniciei a pesquisa de revisão da literatura. Os resultados dessa revisão não foram publicados, mas serviram como base para o artigo apresentado nessa dissertação.

Paralelamente à revisão dos métodos, realizava-se a coleta de dados do inquérito domicíliar denominado "Saúde em Beagá" que foi coordenada pelo OSUBH. A pesquisa foi realizada em dois distritos sanitários contíguos da cidade de Belo Horizonte, denominados Barreiro e Oeste, de um total de 9 distritos (ver seção metodológica para maiores detalhes) e contemplou 4048 adultos e 1042 adolescentes. Como membro do OSUBH, participei das várias etapas desse inquérito: construção dos questionários, delineamento da amostra, sensibilização dos moradores, acompanhamento da coleta de dados, entrada de dados e crítica das informações. Durante esse processo, amadurecia-se a idéia de uma avaliação prática do impacto das perdas nas análises utilizando variáveis do inquérito que apresentaram elevado percentual de dados ausentes.

Dentre as diversas metodologias para a análise de dados incompletos abordadas na revisão da literatura, destacam-se a ACC, que, por ser procedimento padrão na maioria dos softwares estatísticos, é ainda muito utilizada, os métodos de ponderação dos casos completos, Máxima Verossimilhança (MV), imputação simples e Imputação Múltipla (IM). Litle & Rubin[2] apresentam uma boa revisão desses métodos apontando as condições em que os mesmos devem produzir resultados válidos, isto é, estimativas não viciadas e intervalos de confiança em que o nível de significância verdadeiro coincide com aquele pré-especificado (geralmente de 5%).

[2] Little RJ, Rubin DB. Statistical analysis with missing data. 2nd ed. New York: John Wiley & Sons; 2002. 408 p.

Estudos comparativos apontam os métodos de MV e IM como referenciais para a análise de bases incompletas, já que ambos, além de utilizar todas as informações coletadas, devem produzir resultados válidos sob condições menos restritas que os demais [3,4]. Esses estudos discutem ainda vantagens da IM sobre o método de MV em relação à praticidade de aplicação e disponibilidade, já que apenas o primeiro encontra-se implementado na maioria dos softwares de análise tradicionais [5].

Embora existam muitas avaliações do impacto das perdas nas análises por meio de estudos de simulação (e.g., Little [6] e Collins et al. [7]), verificou-se uma lacuna na literatura em relação à mensuração desse impacto em situações práticas. Graham [8] argumenta que isso seria possível por uma tentativa de recuperação, mesmo que parcial, dos dados ausentes. Para acessar o impacto das perdas, os resultados anteriores à recuperação dos dados, produzidos pelos métodos de análise de dados incompletos, poderiam ser comparados aos resultados posteriores. Entretanto, a recuperação desses valores ausentes pode ser difícil ou mesmo inviável, por exemplo, no caso de recusas ou simplesmente desconhecimento por parte do entrevistado.

Assim, com enfoque nos métodos de IM e ACC, propõe-se no presente estudo uma abordagem similar à recuperação dos dados ausentes para avaliar o impacto das perdas em covariáveis de modelos logísticos. Esse enfoque justifica-se pelo fato da IM se tratar de um método de referência e também para ilustrar possíveis problemas da ACC em situações nas quais a proporção de indivíduos com dados incompletos é elevada. O artigo intitulado "Imputação múltipla e análise de casos completos em modelos de regressão logística: Uma avaliação prática do impacto das perdas em covariáveis" foi submetido à revista Cadernos de Saúde Pública e faz parte do corpo dessa dissertação, conforme as exigências do Programa de Pós-graduação em Saúde Pública [9].

---

[3] Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002;7(2):147-77.
[4] Raghunathan TE. What do we do with missing data? some options for analysis of incomplete data. Annu Rev Public Health. 2004;25:99-117
[5] Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. Am Stat. 2007 Feb;61(1):79-90.
[6] Little RJ. Regression with missing X's: a review. J Am Stat Assoc. 1992;87(420):1227-37.
[7] Collins LM, Schafer JL, Kam CM. A comparison of inclusive and 24 restrictive strategies in modern missing data procedures. Psychol Methods. 2001;6(4):330-51.
[8] Graham JW. Missing data analysis: making It work in the real world. Annu Rev Psychol. 2009;60:549-76.
[9] Manual de Orientação 2007. Programa de Pós-graduação em Saúde Pública, Departamento de Medicina Preventiva e Social, Faculdade de Medicina – UFMG

2 OBJETIVOS

Objetivo geral

Discutir os métodos de Análise de Casos Completos e de Imputação Múltipla no contexto da análise de regressão logística com covariáveis incompletas.

Objetivos específicos

Discutir o impacto das perdas nas medidas auto-referidas de peso e altura na distribuição do Índice de Massa Corporal.

Comparar os resultados dos métodos de Análise de Casos Completos e Imputação Múltipla na situação em que o Índice de Massa Corporal é covariável de modelos logísticos.

3 ARTIGO

Imputação múltipla e análise de casos completos em modelos de regressão logística: Uma avaliação prática do impacto das perdas em covariáveis*

Multiple imputation and complete case analysis in logistic regression models: A practical assessment of the impact of incomplete covariate data

Vitor Passos Camargos[1,2], Cibele Comini César[3], Waleska Teixeira Caiaffa[4,2], Cesar Coelho Xavier[5,2], Fernando Augusto Proietti[4,2]

[1] *Programa de Pós-graduação em Saúde Pública, Faculdade de Medicina, UFMG*

[2] *Observatório de Saúde Urbana, Faculdade de Medicina, UFMG*

[3] *Departamento de Estatística, Instituto de Ciências Exatas, UFMG*

[4] *Departamento de Medicina Preventiva e Social, Faculdade de Medicina, UFMG*

[5] *Departamento de Pediatria, Faculdade de Medicina, UFMG*

* Financiamento e coordenadores/participantes do estudo citados no final do texto

Resumo

Pesquisadores da área da saúde lidam frequentemente com o problema das bases de dados incompletas. A Análise de Casos Completos (ACC), que restringe as análises aos indivíduos com dados completos, reduz o tamanho da amostra e pode produzir estimativas viciadas. Baseado em fundamentos estatísticos, o método de Imputação Múltipla (IM) utiliza todos os dados coletados e é recomendado como alternativa a ACC. Dados do estudo Saúde em Beagá, inquérito domiciliar em que participaram 4048 adultos de dois dos nove distritos sanitários da cidade de Belo Horizonte no biênio 2008-2009, foram utilizados para avaliar a ACC e diferentes abordagens de IM no contexto de modelos logísticos com covariáveis incompletas. Peculiaridades de algumas variáveis desse estudo permitiram aproximar uma situação em que os dados ausentes de uma covariável são recuperados e assim os resultados anteriores e posteriores à recuperação são comparados. Verificou-se que mesmo a abordagem mais simplista de IM obteve melhor desempenho que a ACC já que se aproximou mais aos resultados pós-recuperação.

Palavras-chave: métodos, análise estatística, modelos logísticos, índice de massa corporal

Abstract

Researchers in the health field often deal with the problem of incomplete databases. Complete Case Analysis (CCA), which restricts the analysis to subjects with complete data, reduces the sample size and may result in biased estimates. Based on statistical grounds, the Multiple Imputation (MI) method uses all collected data and is recommended as an alternative to CCA. Data from the study "Saúde em Beagá", attended by 4048 adults from two of nine health districts in the city of Belo Horizonte in 2008-2009, were used to evaluate CCA and different MI approaches in the context of logistic models with incomplete covariates. Peculiarities of some variables in this study allowed to approach a situation in which the covariate missing data is recovered and thus the results before and after recovery are compared. It was found that even the more simplistic MI approach performed better than CCA since it was closer to the post-recovery results.

Key words: methods, statistical analysis, logistic models, body mass index

3.1 INTRODUÇÃO

Um problema freqüente nos inquéritos populacionais de saúde é a incompletude das bases de dados. Como passo anterior às análises, a maioria dos softwares estatísticos adota como procedimento padrão a exclusão dos indivíduos com uma ou mais informações ausentes. Esse método, conhecido como Análise de Casos Completos (ACC), pode produzir estimativas viciadas quando suposições necessárias à sua aplicação são violadas [1]. Mesmo quando essas suposições são válidas, a perda de poder devido à exclusão dos indivíduos, que na prática reduz o tamanho da amostra, é inevitável. Embora existam outras metodologias de análise de dados incompletos, a Imputação Múltipla (IM) é recomendada como alternativa à ACC. A IM produz resultados válidos sob suposições menos restritas que a ACC e também ameniza a perda de poder, já que utiliza todos os dados coletados [2].

A ausência de informações é recorrente em temas sensíveis (e.g. renda) e também em indicadores compostos, nos quais as perdas nas variáveis de composição tendem a se somar. O Índice de Massa Corporal (IMC), determinado pela divisão do peso (em quilos) do indivíduo pelo quadrado de sua altura (em metros), é um indicador de grande relevância na saúde pública. Utilizado nas pesquisas como marcador de sobrepeso e obesidade, o IMC é preditor de diversas doenças não transmissíveis como diabetes, doenças cardiovasculares e câncer [9]. Ele pode ser obtido por medidas autorreferidas de peso e altura, que aqui será denominado **IMC_referido**, ou, por medidas diretamente aferidas, denominado **IMC_aferido**. Motivos diversos como constrangimento, desconhecimento ou receio de relatar informações incorretas podem explicar a ausência das medidas autorreferidas de peso e altura, tornando o **IMC_referido** especialmente vulnerável às perdas. Aspectos práticos das análises de uma base de dados na qual o **IMC_referido** apresentou elevado percentual de perdas serão discutidos aqui.

A validade dos resultados dos diferentes métodos de análise de dados incompletos depende de suposições sobre os mecanismos ou fatores associados às perdas. Rubin [3] introduziu três mecanismos teóricos gerais que regulam a ocorrência das perdas e que são extensamente utilizados na literatura: Perda Completamente Aleatória (PCA), Perda Aleatória (PA) e Perda Não Aleatória (PNA). O termo mecanismo é utilizado como sinônimo da função de distribuição dos dados ausentes, que define a probabilidade de cada valor ser ou não observado e os fatores associados a essa probabilidade. O Quadro 1 introduz esses

mecanismos e apresenta exemplos adaptados de Sterne et al. [4], aqui relacionados à ausência de valores do IMC.

Quadro 1 | Tipos de perda em uma variável de interesse (demais variáveis são completamente observadas)
Perda Completamente Aleatória – A probabilidade de perda não depende das variáveis presentes no estudo. Por exemplo, valores ausentes do IMC resultantes de erros de digitação durante a entrada dos questionários.
Perda Aleatória – A probabilidade de perda está relacionada com outras variáveis do estudo, mas não com a variável de interesse, ou seja, a probabilidade de perda está relacionada com um subconjunto conhecido dos dados. Por exemplo, valores ausentes do IMC referido podem ser maiores do que os observados se pessoas fisicamente inativas tiverem maior proporção de perda devido ao acesso menos regular a seu peso.
Perda Não Aleatória – Ocorre quando a probabilidade de perda está relacionada com os valores da própria variável de interesse, que não foram observados – assim, essa relação é desconhecida. Por exemplo, pessoas com valores extremos do IMC podem se sentir menos a vontade para relatar seu peso do que as demais.

Na prática, embora seja plausível para questões como uso de drogas ou renda, a PNA é uma hipótese que não pode ser verificada apenas com os dados observados. O impacto da PNA nas análises produzidas pelos métodos de ACC, IM e outros tem sido majoritariamente avaliado por estudos de simulação [2,5,6]. Apesar da importância desses estudos, Graham [7] argumenta a favor dessa mesma avaliação em situações práticas. Isso seria possível pela recuperação parcial dos dados através de uma amostra aleatória daqueles inicialmente ausentes [8].

Neste estudo dados de um inquérito de saúde domiciliar, denominado Saúde em Beagá (SBH), são analisados e validados por uma abordagem similar à recuperação de dados ausentes, com dois objetivos principais: (a) checar a suposição de PNA na variável **IMC_referido**; (b) avaliar a possibilidade de viés dos coeficientes de modelos de regressão logística estimados pela ACC e por diferentes abordagens de IM na situação em que a perda ocorre predominantemente em uma única covariável (IMC). A seguir são revisados conceitos básicos do método de IM e ACC com enfoque nos modelos de regressão logística com covariáveis incompletas. Posteriormente é apresentado o método de validação das análises, que buscou aproximar a situação de recuperação dos dados ausentes do **IMC_referido**. Finalmente são descritas as variáveis utilizadas nos modelos de regressão logística e as diferentes abordagens de imputação avaliadas.

3.1.1 IMPUTAÇÃO MÚLTIPLA

Proposta por Rubin [10] em 1978, a IM é um método para lidar com o problema da análise de bases incompletas. A IM está disponível nos principais softwares estatísticos comerciais e gratuitos. Horton & Kleinman [11] apresentam uma boa revisão dessas implementações, comparando resultados e fornecendo também para cada software, instruções e sintaxes utilizadas nas análises. O método de IM, se corretamente aplicado, deve produzir estimativas não viciadas mesmo sob os mecanismos de PCA e PA [1,4]. Para ilustrar o método de IM suponha uma base de dados em que valores da variável X_1 estejam ausentes e que outras variáveis sejam completamente observadas. O processo de imputação e análise dos dados pode ser dividido em quatro etapas principais:

Passo 1 (modelo de imputação): Após selecionar as variáveis que serão utilizadas no processo de imputação busca-se, para cada dado ausente, valores plausíveis para preencher essas lacunas. A idéia básica é avaliar a distribuição dos valores observados de X_1 para indivíduos com mesmo perfil (valores idênticos nas variáveis selecionadas) daquele com dado ausente. Para isso usualmente estima-se a distribuição preditiva de X_1 condicionada nas variáveis selecionadas, seguindo uma abordagem bayesiana.

Passo 2: Cada valor ausente é substituído por M valores aleatoriamente amostrados da distribuição condicional preditiva. Esse processo produz ao final M versões completas do banco de dados.

Passo 3 (modelo de análise): Cada banco de dados é analisado pelos métodos tradicionais

Passo 4: Os resultados das M análises são combinados de modo a produzir estimativas que levam em conta a incerteza dos valores imputados.

Os passos 2, 3 e 4 demandam pouco trabalho adicional para o pesquisador quando comparado àquele exigido na ACC. No entanto o primeiro passo é mais trabalhoso e decisivo para a validade dos resultados produzidos pelas análises subseqüentes. Nesse, o pesquisador deve definir as variáveis que farão parte do modelo de predição, também conhecido como modelo de imputação, e o tipo de modelo (por exemplo, linear, logístico ou multinomial) que melhor se ajusta à distribuição de X_1. van Buuren et. al. [12] propõe uma estratégia geral e bem

fundamentada para a seleção de variáveis: (1) Incluir todas as variáveis que serão utilizadas em análises conjuntas com X_1, que é frequentemente denominado modelo de análise; (2) Incluir variáveis associadas às perdas; (3) incluir variáveis preditoras de X_1; (4) Excluir das etapas 2 e 3, aquelas variáveis que apresentam uma elevada proporção de perdas onde X_1 é ausente.

Para ilustrar alguns cuidados necessários ao processo de imputação, considere o contexto no qual o modelo de análise é um modelo logístico no qual X_1 (e.g. IMC) é uma covariável sujeita a perdas e que as demais covariáveis $X_{2:n}$ (e.g. idade e escolaridade) e a variável dependente Y (e.g. diagnóstico de diabetes) são completamente observadas. Considere ainda que o modelo de imputação de X_1 inclui como preditoras apenas as variáveis do modelo logístico (Y e $X_{2:n}$).

Se a perda em X_1 está associada apenas às variáveis utilizadas no modelo de imputação, as estimativas resultantes da IM são válidas (mecanismo de PA). Entretanto, se o mecanismo de PNA atua em X_1, mesmo quando consideramos indivíduos com mesmo perfil em Y e $X_{2:n}$, as estimativas dos coeficientes dos modelos logísticos baseadas na IM podem estar viciadas. O mecanismo de PNA pode ser resultante da omissão de variáveis no modelo de imputação que estão associadas às perdas. Assim, além de incluir as variáveis presentes no modelo logístico, deve-se incluir ainda variáveis potencialmente associadas às perdas, principalmente se essas também estiverem associadas aos valores de X_1 [6].

Como regra geral, todas as variáveis (Y e $X_{2:n}$) e complexidades (estratificação ou efeitos de interação) presentes no modelo logístico devem ser incluídas no modelo de imputação. Por exemplo, a omissão de Y no modelo imputação de X_1 pressupõe a inexistência de uma associação direta entre ambas, e assim, a *Odds Ratio* (OR) de X_1 estimada pela IM é viciada em direção a 1 [13]. Se modelos logísticos distintos serão avaliados em estratos da população de estudo (e.g. sexo), para que possíveis interações entre os estratos e a distribuição de X_1 sejam preservadas, o modelo de imputação também deve ser aplicado independentemente em cada estrato [2].

O modelo de imputação, assim como outros modelos de predição, envolve preocupações relacionadas à preservação de características importantes da distribuição de X_1. Se X_1 é uma variável contínua com distribuição claramente não normal, recomenda-se a aplicação de

alguma transformação (e.g. logarítmica) antes do processo de imputação, já que o modelo preditivo linear usualmente envolve a suposição de normalidade. Nesse caso, após a imputação, aplica-se a transformação inversa para que a variável retorne à sua escala original[2].

Outro aspecto importante a ser considerado na IM é a escolha do número de imputações (M). A escolha de um M pequeno pode inflacionar o intervalo de confiança das estimativas e consequentemente reduzir o poder das análises. Rubin [10] quantifica essa inflação para diferentes frações de informação ausente e escolhas de *M*. O conceito de fração de informação ausente não é o mesmo da proporção de dados ausentes, mas seu valor tende a ser igual ou inferior à proporção de dados ausentes (Rubin [14], p.114). Por exemplo, para frações de informação ausentes de 20%, que pode advir de proporções ainda superiores de dados ausentes, valores de *M*=3 ou *M*=5, produzem respectivamente intervalos apenas 3% e 1% maiores do que os ideais (intervalos mínimos que seriam alcançados quando *M* tende ao infinito). Graham et al. [15] sugerem que a escolha do número de imputações deve ser definida com base no poder (capacidade de detectar efeitos verdadeiros) associado a esta escolha. A queda de poder é avaliada também para diferentes frações de informação ausente e escolhas de *M*.

Recomenda-se que os modelos de imputação incorporem ainda características de desenhos amostrais complexos como, por exemplo, pesos amostrais, estratificação, por meio da inclusão de variáveis indicadoras dos estratos como covariáveis e conglomerados, pela utilização de um modelo de imputação com efeitos aleatórios no nível do conglomerado (Little & Rubin [16], p. 90; Rubin [13]).

Quando as perdas ocorrem em múltiplas variáveis, o processo de imputação torna-se mais complexo, já que as covariáveis utilizadas no modelo de imputação podem também apresentar perdas. Um dos métodos utilizados nesse caso é a imputação pela especificação condicional completa, que realiza as imputações por um processo iterativo de regressões. Uma discussão completa do método é apresentada em van Buuren et al. [17].

3.1.2 ANÁLISE DE CASOS COMPLETOS: REGRESSÃO LOGÍSTICA COM COVARIÁVEL INCOMPLETA

A validade do ajuste do modelo de regressão logística com dados ausentes nas covariáveis pelo método de ACC depende de suposições distintas da IM. Considere novamente um modelo logístico no qual a perda ocorre na covariável X_1 e que as demais covariáveis $X_{2:n}$ e a variável dependente Y são completamente observadas. Em termos práticos, a restrição das análises a indivíduos com dados completos (ACC) seleciona uma sub-amostra dos participantes do estudo. No mecanismo de PCA, essa sub-amostra é representativa da população e assim as estimativas da ACC são válidas.

A ACC pode selecionar uma sub-amostra com características distintas da amostra total do estudo quando as perdas em X_1 dependem dos valores de Y . Por exemplo, indivíduos com dados completos podem apresentar estimativas da OR distintas daqueles omitidos das análises e, portanto, distintas das estimativas que seriam obtidas caso não houvesse perdas. Isso pode ser visto como uma interação entre as perdas e o efeito das covariáveis. A ACC pode ainda, como apontado por Vach & Illi [18], criar interações entre as próprias covariáveis do modelo logístico, anteriormente inexistentes. Como conseqüência, um modelo que seria bem ajustado sem essas interações caso as perdas não ocorressem pode deixar de sê-lo na ACC.

Ainda com relação aos aspectos teóricos, a ACC apresentará resultados válidos quando a probabilidade de perda em X_1 para indivíduos com mesmo perfil nas covariáveis (X_1 e $X_{2:n}$) não está associada a Y. Isso ocorre mesmo que a probabilidade de perda dependa dos próprios valores de X_1 (mecanismo de PNA) e de $X_{2:n}$. Existem ainda condições específicas de dependência entre a probabilidade de perda e Y em que ACC resulta em viés apenas para o intercepto do modelo [18].

3.2 MÉTODOS

3.2.1 O INQUÉRITO SAÚDE EM BEAGÁ (SBH)

Dados deste estudo são provenientes do projeto SBH, um inquérito de saúde domiciliar realizado pelo Observatório de Saúde Urbana de Belo Horizonte (OSUBH) da Universidade Federal de Minas Gerais (UFMG) em dois dos nove distritos sanitários de Belo Horizonte no

biênio 2008-2009. Foi adotada uma amostra estratificada por conglomerados em três estágios: (a) setor censitário, selecionado proporcionalmente ao total de setores de um dos três estratos definidos pelo Índice de Vulnerabilidade à Saúde [19]; (b) domicílio, selecionado por meio de amostra aleatória simples dos domicílios cadastrados na base de dados da Prefeitura Municipal de Belo Horizonte; (c) 1 morador adulto (18 anos ou mais) e 1 morador na faixa de 11 a 17 anos, ambos selecionados aleatoriamente no domicílio.

Os adultos, foco deste estudo, após a entrevista face a face, tiveram aferidas suas medidas de peso (balança Tanita BC-553®) e altura (estadiômetro móvel), que permitiram o cálculo do **IMC_aferido**. O **IMC_referido** foi calculado pelas medidas de peso e altura obtidas respectivamente pelas perguntas "O(A) Sr.(a) sabe seu peso (mesmo que seja valor aproximado)?" e "O(A) Sr.(a) sabe sua altura?". Da amostra final de 4.048 adultos, foram excluídas previamente às análises as mulheres grávidas (n=47) para evitar distorções nas associações do IMC com outras variáveis e os indivíduos com o **IMC_aferido** ausente (n=11). O resultado é uma base de 3.990 adultos (1653 homens e 2337 mulheres), todos com medidas do **IMC_aferido**.

3.2.2 METODOLOGIA DE VALIDAÇÃO DAS ANÁLISES

A variável **IMC_referido** está ausente para 789 indivíduos (21%), o que ocorreu principalmente devido a não resposta da medida de altura (15%). Caso os dados ausentes do **IMC_referido** fossem substituídos pelos valores do **IMC_aferido,** teríamos uma situação muito similar à recuperação desses dados. Entretanto, devido às distorções existentes entre as medidas referidas e aferidas do IMC [20]**,** optou-se por utilizar nas análises apenas a variável **IMC_aferido**. Os valores do **IMC_aferido** foram excluídos para os mesmos 789 indivíduos com dados referidos ausentes, criando uma nova variável pós-exclusão, que será diferenciada por um asterisco (**IMC_aferido***). Em seguida, os dados do **IMC_aferido*** foram analisados pelos métodos de ACC e IM e os resultados foram comparados com aqueles obtidos com o **IMC_aferido** pré-exclusão. Com isso, o mecanismo de perda original é mantido e a validação dos resultados não é comprometida.

3.2.3 MODELOS LOGÍSTICOS AVALIADOS

Para avaliar um possível viés na estimativa dos coeficientes de regressão logística quando o **IMC_aferido*** é uma das covariáveis, foram testados três modelos que se diferenciam pela variável dependente: (1) **Diabetes** (1.sim, 0.não); (2) **Hipertensão** (1.sim, 0.não); (3) **Peso_acima** (1.acima do peso, 0.satisfeito ou abaixo). As duas primeiras variáveis dependentes foram obtidas de uma questão sobre doenças crônicas ("Alguma vez, um médico ou outro profissional de saúde já disse que o(a) Sr.(a) tem alguma dessas doenças crônicas listadas abaixo") e a última, da pergunta "Com relação a seu peso, o(a) Sr.(a) está:", que foi tratada de forma dicotômica. Além do **IMC_aferido*** as covariáveis **Idade**, **Cor** e **Escolaridade,** categorizadas em três níveis (Tabela 1), compõem os modelos logísticos. As análises foram estratificadas por sexo, resultando, assim, em 6 modelos logísticos.

As estimativas de cada modelo logístico com a covariável **IMC_aferido** servirão de referência para a comparação com estimativas obtidas com a mesma variável pós-exclusão. O modelo de referência pode ser visto como uma ACC com perdas mínimas, já que no máximo 18 mulheres e 8 homens são ignorados por não apresentarem dados completos nas variáveis de cada modelo.

3.2.4 MÉTODOS DE ANÁLISE DE DADOS INCOMPLETOS

Para o ajuste dos modelos logísticos com a covariável **IMC_aferido***, além da ACC, foram avaliadas quatro abordagens de IM (IM0, IM1, IM2 e IM3), que se diferenciam unicamente pelas variáveis utilizadas no modelo de imputação. A ACC reduz sensivelmente a amostra, ignorando os 789 indivíduos (586 mulheres e 203 homens) com valores ausentes no **IMC_aferido***. Nas abordagens de IM, as variáveis incluídas no modelo de imputação foram selecionadas de forma cumulativa. Assim, o modelo de IM1 contém todas as variáveis utilizadas no IM0 e assim sucessivamente.

O modelo de imputação de IM0 inclui apenas as covariáveis dos modelos logísticos (**Idade, Cor** e **Escolaridade**) e, assim, omite a variável dependente. Espera-se aqui ilustrar o potencial viés dessa abordagem. O IM1 contém as mesmas variáveis utilizadas em cada modelo de análise, acrescentado comparativamente a IM0, a variável dependente de cada modelo logístico. As variáveis adicionadas aos modelos IM2 e IM3 foram selecionadas

seguindo a estratégia de Van Burren [12] descrita anteriormente. A lista de covariáveis de cada modelo de imputação é apresentada no Quadro 2 com uma breve descrição das mesmas entre parênteses.

Quadro 2 | Lista de covariáveis utilizadas em cada modelo de imputação (IM0, IM1, IM2 e IM3)

IM0: **Idade** (em anos:18-39; 40-59; 60 ou mais) **+ Cor** (Branca; Negra; Outra) + **Escolaridade** (em anos: 0-3; 4-10; 11 ou mais)

IM1: *IM0* **+ Var_dependente** (**Diabetes** ou **Hipertenção** ou **Peso_Acima)**

IM2: *IM1* **+ Silhueta[1] + Silhueta2[1] + Satisfação_corporal[1] + Estado_civil** (solteiro; casado/amigado; separado/desquitado; viúvo) **+ Renda_familiar** (em salários mínimos: menos de 2; 2 |- 5; 5 |- 10; 10 ou mais) **+ Doença_mental[2] + Mudança_peso** (indicadora: está tentando alterar o peso) **+ Inativo** (indicadora: não praticou atividade física nos últimos 3 meses)

IM3: *IM2* + **Colesterol_alto[2] + Artrite[2] + Epilepsia[2] + Chefe_família** (indicadora: é chefe da família) **+ Posse_veículo** (indicadora: possui carro ou moto) **+ Restrição_alimentar** (indicadora: alguma vez na vida reduziu ou deixou de fazer refeições devido a problemas financeiros**) + Consumo_fruta** (consumo de frutas semanal codificado em 3 categorias) **+ Fumo** (indicadora: fuma diariamente)

[1] variáveis que atuam como proxy do **IMC** (mais detalhes no texto); [2] Indicadora de doença crônica autorreferida.

O modelo IM2 inclui, além das variáveis do modelo de análise, as 8 principais variáveis associadas ao **IMC_aferido*** ou à variável indicadora das perdas do **IMC_aferido*** (1. ausente, 0. observado) e pode ser visto, segundo discutido por Collins et al. [6], como uma abordagem restritiva de seleção de variáveis. O modelo IM3 adiciona ainda 8 variáveis e pode ser classificada como uma abordagem inclusiva, já que as variáveis incluídas têm menos poder de predição ou associação mais fraca com as perdas. Todas as variáveis incluídas em IM2 e IM3 tiveram associação estatisticamente significativa ou com o **IMC_aferido*** ou com a variável indicadora de perdas.

Dentre as 8 variáveis incluídas no IM2 destacam-se variáveis que atuam como *proxy* do IMC: **Silhueta**, **Silhueta2** e **Satisfação_corporal**. A variável **Silhueta** tem valores de 1 a 9 e corresponde à escala de silhuetas de Stunkard et al. [21], discutida e ilustrada mais recentemente por Gardner [22]. A escala é composta por 9 figuras masculinas e femininas variando nos extremos, da magreza à representação de um indivíduo obeso. A variável **Silhueta** corresponde à escolha do entrevistado sobre a "figura que se parece mais com" ele hoje e a variável **Silhueta2** corresponde àquela variável elevada ao quadrado. A variável **Satisfação_corporal** representa a diferença entre a **Silhueta** e a escolha da figura que o indivíduo gostaria de se parecer.

Embora, no modelo logístico, o IMC seja tratado como variável categórica, a imputação foi realizada considerando sua forma contínua por meio de um modelo linear. Após avaliações, verificou-se que a transformação logarítmica da variável **IMC_aferido*** foi a que apresentou melhor ajuste ao modelo linear e, portanto, a transformação foi aplicada previamente às imputações. Após as imputações, aplicou-se a transformação inversa e a variável foi categorizada. Todas as imputações levaram em consideração os pesos amostrais da pesquisa e incluíram ainda uma variável identificadora dos estratos amostrais. Em todas as abordagens ,foram realizadas 20 imputações independentemente para cada sexo, acompanhando a estratificação das análises.

O teste Kolmogorov-Smirnov foi utilizado para avaliar a igualdade da distribuição da variável **IMC_aferido** em diferentes grupos e o teste qui-quadrado para verificar associações entre as perdas e variáveis utilizadas nos modelos logísticos. Com objetivo de verificar problemas resultantes da seleção de indivíduos da ACC, o teste Wald para múltiplos coeficientes foi utilizado para testar a significância conjunta das interações entre as perdas do **IMC_referido** e os coeficientes de cada modelo logístico de referência. Avaliou-se ainda a qualidade do ajuste dos modelos logísticos de referência (pré-exclusão) e dos modelos da ACC (pós-exclusão) por meio do teste de Pearson. Foi adotado o nível de significância de 5% em todos os testes.

Todas as análises foram realizadas no software STATA 11 e levaram em consideração o desenho amostral. As imputações desse artigo foram realizadas utilizando o comando “mi ice” (ver Royston [23] para detalhes da última atualização da implementação) que utiliza como referência o método descrito por van Buuren et al. [12].

3.3 RESULTADOS

O nível de perdas na variável **IMC_referido** foi quase duas vezes superior no sexo feminino (25,1% contra 12,3% do sexo masculino) e assim, espera-se observar para esse grupo um maior impacto nas análises. A distribuição do **IMC_aferido** das mulheres com **IMC_referido** ausente apresentou caudas mais largas do que daquelas com **IMC_referido** observado ($p<0,01$; Figura 1). Assim, mulheres com valores extremos do **IMC_aferido** têm maior propensão a perdas, i.e. não responder as medidas referidas de peso e/ou altura.

Homens com **IMC_referido** ausente tiveram menores valores do **IMC_aferido**, dado o desclocamento da distribuição para a esquerda (p<0,01). Isso indica que controlando apenas pela variável **Sexo,** o mecanismo de PNA não pode ser descartado.

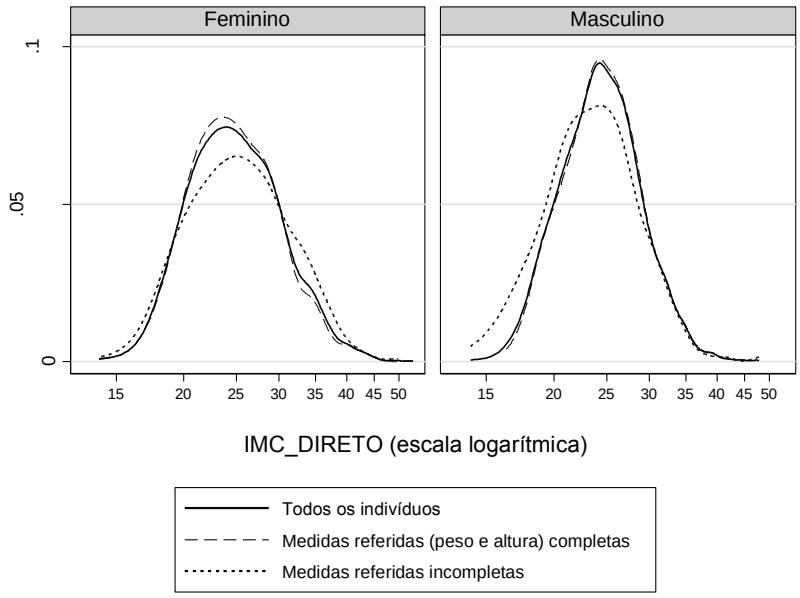

Figura 1. Distribuição do **IMC_aferido** para todos os indivíduos, para aqueles com medidas referidas de peso e altura completas e para aqueles com medidas incompletas, segundo sexo.

Os dados da Tabela 1 apresentam maiores detalhes da distribuição das perdas na variável **IMC_referido**. Indivíduos mais escolarizados apresentaram menor proporção de perdas com forte gradação para ambos os sexos. A perda, diferentemente do que foi verificado para a variável contínua, está associada ao **IMC_aferido** na sua forma categorizada apenas para o sexo feminino. Entre as mulheres, a perda está associada a todas as variáveis dependentes dos modelos logísticos. Para os homens, a proporção de perdas é praticamente constante nos níveis das variáveis dependentes **Diabetes** e **Hipertensão**, mas está associada à variável **Peso_acima**.

Os resultados dos modelos logísticos utilizando a variável **IMC_aferido** (pré-exclusão), que servem como referência para a avaliação dos métodos de ACC e IM, são apresentados na Tabela 2. Considerando os 6 modelos ajustados, temos um total de 48 coeficientes estimados (sem levar em conta os 6 interceptos), dos quais 32 são estatisticamente significativos.

Testes indicaram bom ajuste de todos os modelos logísticos de referência (Tabela 2) com exceção do modelo **Hipertensão** para o sexo feminino (p=0,01). A inclusão de interações da covariável **IMC_aferido** com as demais melhoraram o ajuste desse modelo (p=0,37; resultados não publicados, ver discussão), o mesmo foi verificado na ACC. O modelo **Diabetes** para o sexo feminino, que na análise de referência apresentou bom ajuste (p= 0,11), foi o único que na ACC perdeu essa qualidade (p= 0,03).

Tabela 1. Distribuição das variáveis utilizadas nos modelos de regressão, proporção de indivíduos com **IMC_referido** ausente e teste para diferença nessas proporções segundo sexo

	Feminino			Masculino		
	N	IMC_referido ausente (%)	Valor de p	N	IMC_referido ausente (%)	Valor de p
IMC_aferido						
<25	1043	22,8	(0,01)	817	13,7	(0,16)
≥ 25 e <30	759	23,7		592	10,3	
≥ 30	535	31,4		244	12,3	
Cor						
Branca	933	21,3	(<0,01)	593	11,1	(0,39)
Negra	274	36,5		241	14,5	
Outra	1117	25,2		814	12,4	
Escolaridade						
0-3 anos	395	49,9	(<0,01)	195	22,6	(<0,01)
4-10 anos	913	29,0		692	14,7	
11+ anos	1027	12,1		766	7,4	
Idade						
18-39 anos	960	22,9	(<0,01)	750	14,1	(0,01)
40-59 anos	875	23,5		578	9,0	
60+ anos	502	31,9		325	13,8	
Variáveis dependentes dos modelos logísticos						
Peso_acima						
Satisfeito ou abaixo	1021	29,3	(<0,01)	1088	13,7	(0,01)
Acima	1313	21,7		565	9,6	
Diabetes						
Não	2133	24,6	(0,05)	1533	12,3	(0,93)
Sim	202	30,7		117	12,0	
Hipertensão						
Não	1566	22,3	(<0,01)	1186	12,2	(0,90)
Sim	769	30,7		466	12,4	

Para cada modelo de referência (tabela 2), foi avaliada a possibilidade de interação entre os efeitos das covariáveis e uma variável indicadora das perdas no **IMC_referido**. Apenas no modelo Peso_acima (sexo masculino e feminino), a inclusão de interações para todas as

covariáveis (além da própria variável indicadora) foi estatisticamente significativa, indicando que os indivíduos com dados completos apresentam associações (entre covariáveis e variável dependente) distintas dos demais (**IMC_referido** ausente).

Tabela 2. Resultados dos modelos logísticos com a covariável **IMC_aferido** (pré-exclusão) segundo variável dependente e sexo

	Diabetes				**Hipertensão**				**Peso_acima**			
	Fem. (n= 2320)		Mas. (n= 1645)		Fem. (n= 2320)		Mas. (n=1647)		Fem. (n=2319)		Mas. (n= 1648)	
Covariável (ref.)	OR	IC (95%)	OR	IC (95%)	OR	IC (95%)	OR	IC (95%)	OR	IC (95%)	OR	IC (95%)
***IMC_aferido** (<25)*	1		1		1		1		1		1	
≥ 25 e <30	**1,7**	[1,0-2,8]	1,8	[1,0-3,5]	**2,0**	[1,5-2,8]	**1,7**	[1,1-2,4]	**13,6**	[9,9-18,6]	**16,7**	[11,2-25,0]
≥ 30	**2,8**	[1,7-4,8]	**2,8**	[1,3-5,7]	**4,9**	[3,5-7,0]	**3,8**	[2,4-6,2]	**40,3**	[24,8-65,3]	**88,1**	[48,9-158,5]
***Cor** (Branca)*	1		1		1		1		1		1	
Negra	1,6	[0,9-3,0]	0,8	[0,3-2,2]	**2,6**	[1,7-3,9]	**2,4**	[1,4-4,1]	0,7	[0,5-1,1T	**0,5**	[0,3-0,8]
Outra	**1,8**	[1,2-2,7]	1,2	[0,7-2,0]	1,3	[1,0-1,7]	1,1	[0,7-1,6]	0,9	[0,6-1,2]	**0,6**	[0,4-0,9]
***Escolaridade** (0-3)*	1		1		1		1		1		1	
4-10 anos	1,0	[0,7-1,6]	1,3	[0,5-3,2]	**0,6**	[0,4-0,9]	**0,5**	[0,3-0,7]	**1,8**	[1,2-2,7]	1,3	[0,7-2,6]
11+ anos	0,6	[0,3-1,0]	1,5	[0,6-3,3]	**0,4**	[0,3-0,6]	**0,4**	[0,3-0,7]	**3,6**	[2,4-5,4]	**2,6**	[1,3-5,2]
***Idade** (18-39)*	1		1		1		1		1		1	
40-59 anos	**3,7**	[2,0-6,9]	**8,7**	[2,9-26,3]	**4,8**	[3,4-6,9]	**2,7**	[1,7-4,2]	0,7	[0,5-1,0]	1,0	[0,7-1,4]
60+ anos	**12,0**	[6,7-21,7]	**26,6**	[9,1-77,6]	**22,5**	[15,4-32,8]	**9,8**	[5,9-16,4]	**0,2**	[0,1-0,3]	0,6	[0,4-1,1]

Valores da OR em negrito indicam p<0,05

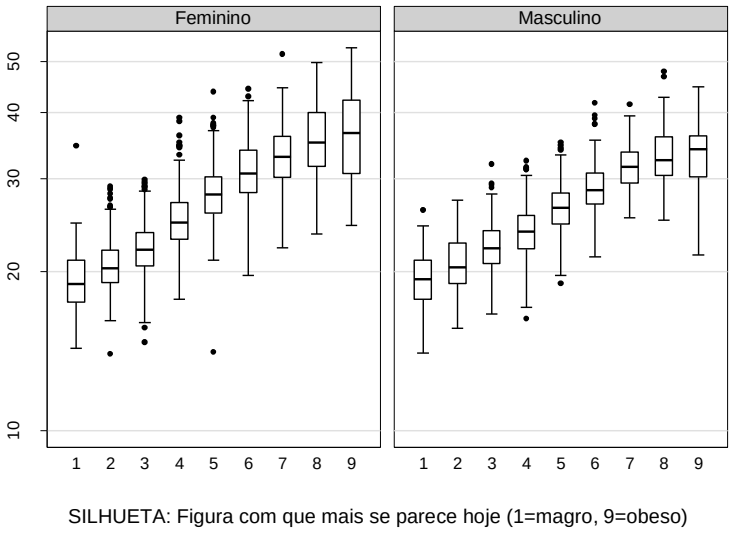

Figura 2. Box-plot da variável **IMC_aferido** segundo a variável **Silhueta** e sexo.

No processo de seleção de variáveis para os modelos de imputação de IM2 e IM3, a variável **Silhueta** foi a mais importante preditora do IMC. O ajuste de um modelo linear do logaritmo

do **IMC_aferido*** tendo como covariáveis apenas a variável **Silhueta** e a interação da mesma com a variável **Sexo**, apresentou 60% da variância explicada pelo modelo (R^2). Esse percentual aumentou ainda para 62% quando a variável **Satisfação_corporal** (categorizada) interada com a variável **Sexo** foi incluída no modelo. A Figura 2 ilustra a relação quase linear entre a variável **Silhueta** e a mediana do **IMC_aferido** (na escala logarítmica).

Após a exclusão dos valores do **IMC_aferido** e aplicação dos métodos de ACC e de IM, os resultados de cada modelo logístico foram comparados ao modelo de referência. Como esperado, na abordagem IM0, que omite a variável dependente do modelo de imputação, a OR da variável **IMC_aferido*** foi subestimada em todos os modelos, caindo para menos da metade da OR de referência no modelo **Peso_acima** para o sexo feminino (Figura 3). De modo geral, o desvio de cada estimativa da OR reduz de forma consistente quando se caminha, na Figura 3, da ACC para a abordagem IM2 e se mantém constante na abordagem IM3.

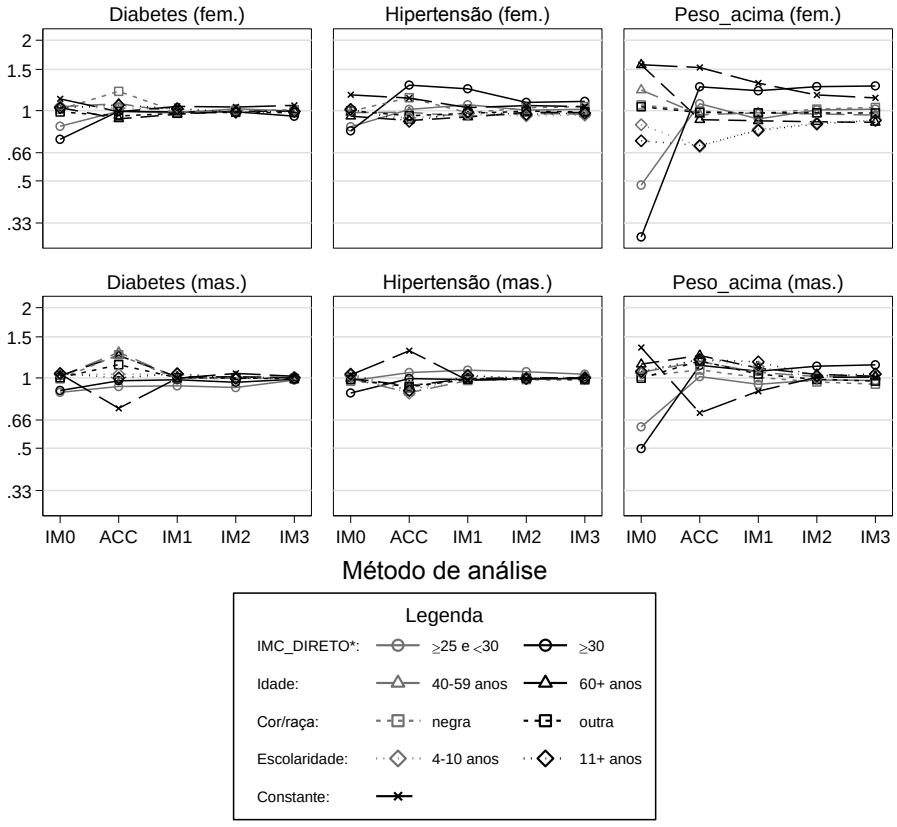

Figura 3. Razão (em escala logarítmica) entre a OR estimada pelos métodos de análise de dados incompletos e a OR do modelo de referência (pré-exclusão) segundo variável dependente e sexo.

Similarmente à Figura 3, a Figura 4 apresenta a distribuição (média e amplitude) da razão entre a OR estimada pelos métodos e a OR de referência. Entretanto, nessa figura, as razões menores do que 1 são invertidas e os desvios da OR podem ser avaliados independente de se tratarem de sub ou sobre-estimações. Também são omitidos todos os valores referentes às estimativas do intercepto do modelo logístico, que geralmente não são de interesse do pesquisador. À direita da Figura 3, apresenta-se um resumo dos resultados por sexo.

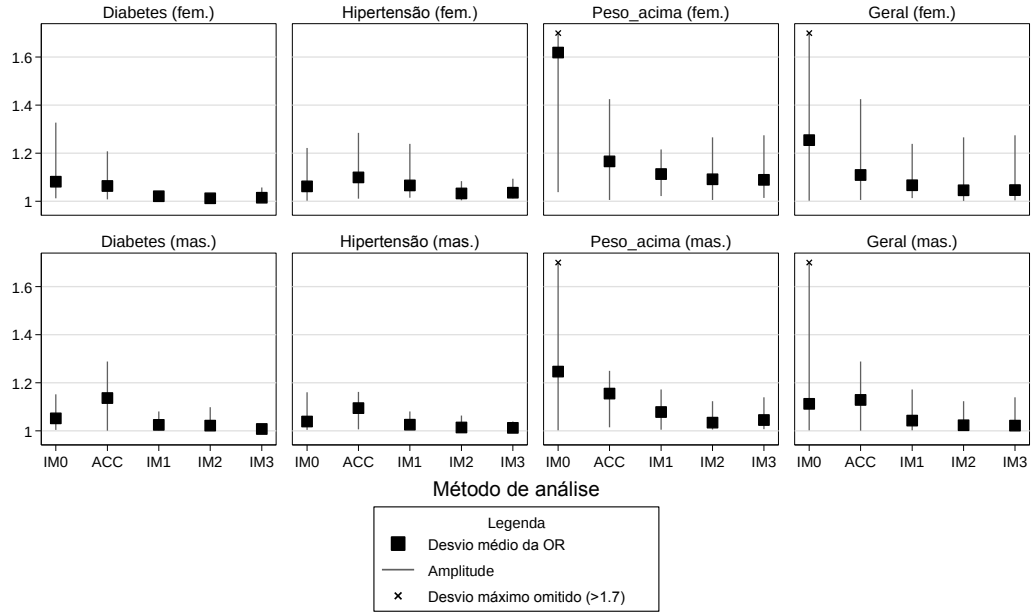

Figura 4. Distribuição (média e amplitude) do desvio entre a OR estimada pelos métodos de análise de dados incompletos e a OR do modelo de referência (pré-exclusão) segundo variável dependente e sexo.

Nos resultados gerais de ambos os sexo,s a ACC apresentou em média o maior desvio na OR dentre as abordagens consideradas próprias (feminino: ACC com desvio médio de 11,0%, IM1 6,7%, IM2 4,6% e IM3 4,7%; masculino: ACC 12,9%, IM1 4,3%, IM2 2,3%, IM3 2,2%). Esse pior desempenho da ACC também é verificado em cada um dos modelos logísticos avaliados. Observa-se em geral um melhor desempenho das abordagens de IM entre os homens, nos quais houve uma menor proporção de perdas do **IMC_referido**. O modelo IM0, sem surpresas, foi a abordagem que apresentou o maior desvio médio da OR entre as mulheres e também a maior variação (amplitude) para ambos os sexos.

É natural que o erro padrão de um coeficiente estimado pelos métodos de análise de dados incompletos seja superior àquele que idealmente seria obtido se todos os dados fossem observados. Entretanto, métodos mais eficientes tendem a produzir erros padrão próximos desse ideal. Os erros padrão dos coeficientes estimados pelos métodos de ACC, IM1, IM2 e

IM3 apresentaram respectivamente inflação média de 18,9%, 5,6%, 3,2% e 3,5%, para o sexo feminino e de 7,0%, 2,7%, 1,4% e 1,9% para o sexo masculino quando comparados aos erros padrão respectivos do modelo de referência (Figura 5 – resultados gerais). O elevado percentual de perdas do **IMC_referido** no sexo feminino teve impacto sensível na inflação dos erros padrão desse grupo.

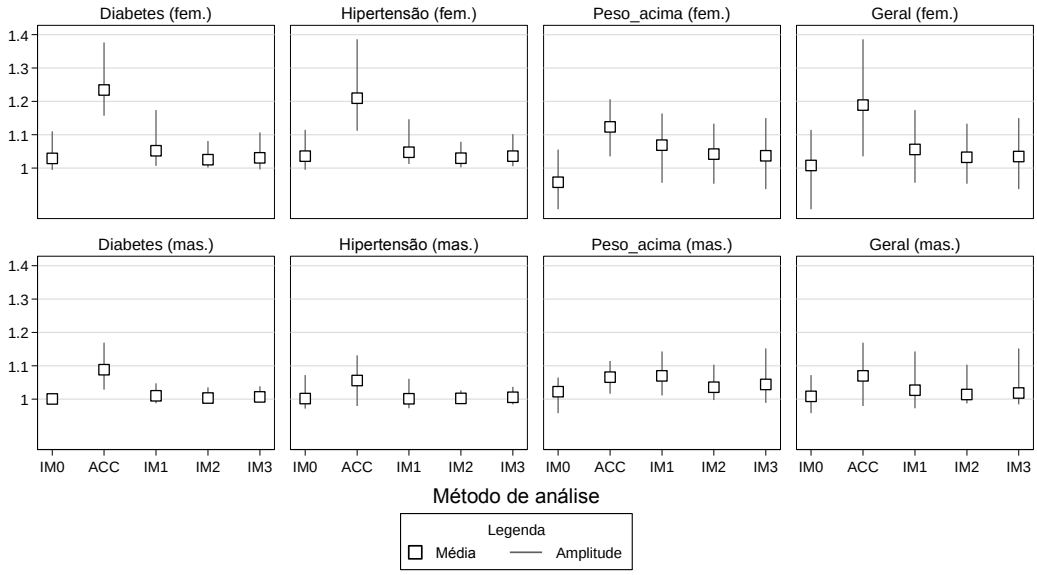

Figura 5. Distribuição (média e amplitude) da razão entre o erro padrão dos coeficientes estimados pelos métodos de análise de dados incompletos e o erro padrão estimado pelo modelo de referência (pré-exclusão) segundo variável dependente e sexo.

A abordagem IM0, que claramente vicia a OR da variável **IMC_aferido***, não incorpora essa imprecisão no erro padrão de seus coeficientes. Isso indica que os intervalos de confiança produzidos por essa abordagem são inconsistentes, i.e. não devem refletir a confiança nominal (95%).

A inflação do erro padrão reflete diretamente na significância dos coeficientes. Dos 32 coeficientes significantes dos modelos de referência (Tabela 2), deixaram de ser significativos a OR do **IMC_aferido** (≥ 25 e <30) do modelo **Diabetes** (fem.) em todos os métodos, a OR da **Escolaridade** (4-10 anos) do modelo **Peso_acima** (fem.) na ACC e IM1, e a OR da **Cor** (outra) do modelo Peso_acima (mas.) apenas na ACC. Dos 16 coeficientes não significativos, apenas a abordagem IM3 apresentou divergência (OR da categoria "40-59 anos" do modelo modelo **Peso_acima**, sexo feminino). Assim, a ACC foi o método com maior discrepância em relação aos modelos de referência (3 resultados divergentes). As divergências, de um modo

geral, ocorreram quando a significância é limítrofe no modelo de referência (valor p próximo de 0,05).

3.4 DISCUSSÃO

Em todos os modelos logísticos avaliados, as abordagens de imputação IM1, IM2 e IM3 tiveram, em média, menor desvio/viés da OR em relação às estimativas de referência (pré-exclusão) quando comparadas com o desvio da ACC. Assim, mesmo a abordagem de imputação mais simples (IM1), que inclui no modelo de imputação apenas as variáveis do modelo de análise, obteve melhor desempenho que a ACC. Além disso, os intervalos de confiança da ACC apresentaram uma inflação consideravelmente maior que todas as abordagens de IM, levando a não detecção de efeitos que, no modelo de referência, eram significativos. As abordagens IM2 e IM3, que utilizam a estratégia de seleção de variáveis proposta por Van Burren et al.[12], tiveram menor desvio da OR e inflação do erro padrão que a abordagem IM1.

Os piores resultados da ACC indicam a existência de algum nível de associação entre as perdas no **IMC_referido** e as variáveis dependentes avaliadas, mesmo para indivíduos com mesmo perfil nas covariáveis [5,18]. Isso fez com que indivíduos com **IMC_referido** observado (dados completos) tivessem relações entre covariáveis e variável dependente distintas dos demais (modelo Peso_acima, sexo masculino e feminino). No modelo **Diabetes** para o sexo feminino, observou-se ainda que a restrição das análises aos indivíduos com dados completos levou a um modelo mal ajustado. Esse resultado corrobora com a possibilidade de surgimento de falsas interações entre as covariáveis na ACC apontada por Vach & Illi [18].

Nesse estudo a abordagem inclusiva de seleção de variáveis para o modelo de imputação, representada por IM3, teve resultados muito similares ao obtido pela abordagem restritiva (IM2). Segundo Collins et al. [6], que avalia por meio de simulações diferentes abordagens de imputação em uma única variável (X_1), estratégias inclusivas são preferíveis, pois reduzem o risco de omissão de variáveis importantes no modelo de imputação e, portanto, o risco de viés. Verificou-se, por exemplo, que a não inclusão de uma variável sabidamente associada a essas perdas no modelo de imputação levou a viés de estimativas posteriores e ainda, que a magnitude do viés depende do grau de correlação dessa variável omitida com X_1. Em outra circunstância, na qual foi imposta uma PNA em X_1, verificou-se que a inclusão de variáveis

fortemente relacionadas à X_1 no modelo de imputação resultou em redução considerável no viés das análises posteriores.

Os modelos apresentados nesse estudo têm como único objetivo ilustrar e comparar os métodos de análises de dados incompletos. Portanto, não houve qualquer intenção de ajustar modelos baseados em marcos teóricos, ou, por exemplo, preocupações relativas a efeitos de confusão. Considerou-se um mínimo de plausibilidade na composição dos modelos para que parte dos coeficientes tivesse significância estatística e ainda foi avaliada a qualidade do ajuste dos mesmos. O único modelo mal ajustado, **Hipertensão** para o sexo feminino, produziu resultados muito similares àquele que incluía as interações necessárias. Assim, para evitar uma apresentação demasiadamente complexa nos gráficos e tabelas, optou-se pela publicação dos resultados sem interações.

As diferenças observadas entre as estimativas das análises de dados incompletos e aquelas obtidas pelo modelo de referência (pré-exclusão), são diferenças amostrais, no sentido de que as mesmas estão sujeitas a variações caso esse estudo fosse replicado na mesma população. Assim, não se deve interpretar, por exemplo, os desvios da OR observados na ACC como se esses fossem desvios em relação à OR verdadeira (parâmetro populacional) como geralmente é feito em estudos de simulação. No caso do modelo Peso_acima (sexo masculino e feminino), dada a significância estatística das interações incluídas no modelo de referência, os desvios da OR devem refletir desvios populacionais. O mesmo vale para a diferença verificada na distribuição do IMC_aferido de homens e mulheres entre indivíduos com IMC_referido ausente e aqueles em que essa variável foi observada.

É importante ressaltar que embora as perdas no **IMC_aferido** tenham sido geradas artificialmente, reproduzindo as perdas do **IMC_referido**, esse estudo busca aproximar ao máximo a situação em que os dados ausentes do **IMC_referido** são recuperados. Entendemos que os resultados observados aqui são generalizáveis para estudos que trabalham unicamente com a variável **IMC_referido** e recomendamos que o método de IM seja preferido em relação a ACC nas análises que envolvam essa variável.

A recuperação de dados ausentes pode trazer diversas dificuldades práticas, já que os problemas que levaram à ausência de informações podem se repetir (e.g. recusas) ou ainda essa recuperação pode ser inviável, por exemplo, no caso de perda de seguimento em um

estudo de coorte. A abordagem aqui adotada permite, de forma simples e eficaz, a avaliação do impacto real das perdas nas análises. Sugerimos, assim, a condução de outros estudos similares para a acumulação de novas evidências.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Raghunathan TE. What do we do with missing data? some options for analysis of incomplete data. Annu Rev Public Health. 2004;25:99-117.

2. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002;7(2):147-77.

3. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581-90; discussion 590-2.

4. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.

5. Little RJ. Regression with missing X's: a review. J Am Stat Assoc. 1992;87(420):1227-37.

6. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and 24 restrictive strategies in modern missing data procedures. Psychol Methods. 2001;6(4):330-51.

7. Graham JW. Missing data analysis: making It work in the real world. Annu Rev Psychol. 2009;60:549-76.

8. Glynn RJ, Laird NM, Rubin DB. Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. J Am Stat Assoc. 1993;88(423):984-993.

9. World Health Organization. Obesity: preventing and managing the global epidemic. Report of a WHO consultation. World Health Organ Tech Rep Ser. 2000;894:i-xii, 1-253.

10. Rubin DB. Multiple imputations in sample surveys - a phenomenological bayesian approach to nonresponse. In Proceedings of the Survey Research Methods Section, Am Stat Assoc [internet]. 1978;20-8. Disponível em: http://www.amstat.org/sections/srms/proceedings/papers/1978_004.pdf (acessado em 18/jan/2011)

11. Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. Am Stat. 2007 Feb;61(1):79-90.

12. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999 Mar 30;18(6):681-94.

13. Rubin DB. Multiple imputation after 18+ years. J Am Stat Assoc. 1996;91(434):473-89.

14. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987. 288 p.

15. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev Sci. 2007 Sep;8(3):206-13.

16. Little RJ, Rubin DB. Statistical analysis with missing data. 2nd ed. New York: John Wiley & Sons; 2002. 408 p.

17. van Buuren S, Brand JP, Groothius-Oudshoorn CG, Rubin DB. Fully conditional specification in multivariate imputation. J Stat Comput Simul. 2006;76(12):1049-64. Disponível em: http://www.stefvanbuuren.nl/publications/FCS%20in%20multivariate%20imputation%20-%20JSCS%202006.pdf (acessado em 18/jan/2011)

18. Vach W, Illi S. Biased estimation of adjusted odds ratios from incomplete covariate data due to violation of the missing at random assumption. Biometrical Journal. 1997;39:13-28.

19. Secretaria Municipal de Saúde de Belo Horizonte, Gerência de Epidemiologia e Informação. Índice de vulnerabilidade à saúde 2003. Jul 2003. Disponível em: www.pbh.gov.br/smsa/biblioteca/gabinete/risco2003 (acessado em 18/jan/2011)

20. McAdams MA, Van Dam RM, Hu FB. Comparison of self-reported and measured BMI as correlates of disease markers in US adults. Obesity (Silver Spring). 2007 Jan;15(1):188-96.

21. Stunkard AJ, Sorensen T, Schulsinger F. Use of the daniel adoption registry for the study of obesity and thinness. In: Skety S, Rowland LP, Sidman RL; Matthysse SW. Genetics of neurological and psychiatric disorders. New York: Raven Press; 1983. p. 115-20.

22. Gardner RM, Friedman BN, Jackson NA. Methodological concerns when using silhouettes to measure body image. Percept Mot Skills. 1998 Apr;86(2):387-95.

23. Royston P. Multiple imputation of missing values: further update of ice, with an emphasis on categorical variables. Stata Journal. 2009;9(3):466-77.

4 CONSIDERAÇÕES FINAIS

Os resultados desse estudo se alinham aos resultados obtidos por estudos de simulação, que apontam o ganho de eficiência da IM e uma possível redução do viés das estimativas quando comparadas àquelas produzidas pela Análise de Casos Completos (ACC). A aplicação incorreta do método de IM foi ilustrada quando se omitiu do modelo de imputação a variável dependente do modelo logístico. Verificou-se nesse caso que a IM produziu piores resultados que a ACC.

> *"O método de Imputação Múltipla (IM) têm potencial para aumentar a validade das pesquisas médicas. No entanto, o processo de imputação múltipla exige que o pesquisador modele a distribuição de cada variável com valores ausentes, em relação aos valores observados. A validade dos resultados da IM depende da especificação cuidadosa e apropriada desses modelos. A IM não deve ser tratada como uma técnica rotineira aplicada ao apertar de um botão – sempre que possível a ajuda de um especialista em estatística deve ser requerida." Sterne et al.*[1]

A inclusão de bons preditores do IMC no modelo de imputação teve impacto positivo nas análises. Isso pôde ser verificado na melhora dos resultados do modelo IM1 para o IM2 quando também são incluídas no modelo de imputação variáveis *proxy* do IMC (relacionadas à escala de silhuetas). A abordagem inclusiva de seleção de variáveis para o modelo de imputação, IM3 (ver APÊNDICE A, para sintaxe do STATA 11), incluiu um total de 22 variáveis preditoras no modelo de imputação do **IMC_aferido*** e obteve resultados similares à abordagem restritiva (IM2), que utilizou 14 variáveis. Esse resultado coincide com achados de Collins et al. [2], que recomenda a adoção da abordagem inclusiva, já que a mesma reduz o risco da omissão de variáveis importantes no modelo de imputação.

[1] Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.

[2] Collins LM, Schafer JL, Kam CM. A comparison of inclusive and 24 restrictive strategies in modern missing data procedures. Psychol Methods. 2001;6(4):330-51.

Merecem ainda destaque as considerações de Sterne et al.[1] sobre a publicação de estudos baseados no método de IM. Eles revisaram publicações que utilizaram a IM em quatro periódicos biomédicos (*New England Journal of Medicine*, Lancet, BMJ e JAMA de 2002 a 2007) registrando os principais aspectos metodológicos da IM reportados nos estudos. Observou-se, por exemplo, que, dos 59 artigos encontrados, 53 não reportaram as variáveis utilizadas no modelo de imputação. Isso é um problema, já que a escolha das variáveis é um aspecto que pode determinar a validade da metodologia. Os mesmos autores propõem um guia esquemático com orientações para a publicação de resultados em estudos que utilizam o método de IM.

A restrição desse estudo às perdas no Índice de Massa Corporal baseado nas medidas auto-referidas de peso e altura limita em algum nível suas conclusões. Outras variáveis terão mecanismos de perda distintos, e assim, mesmo que a proporção de dados ausentes seja similar à encontrada aqui, os resultados serão diferentes. Assim, é importante a realização de outros estudos para o acúmulo de novas evidências.

[1] Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.

APÊNDICES E ANEXOS

APÊNDICE A – Exemplo da sintaxe de comandos utilizados no artigo (STATA 11)

APÊNDICE B – Projeto de Pesquisa

ANEXO A – Recibo de submissão do artigo

ANEXO B – Folha de aprovação do Comitê de Ética

ANEXO C – Certificado de Qualificação

APÊNDICE A - Exemplo da sintaxe de comandos utilizados no artigo (STATA 11)

Sintaxe dos comandos para o modelo de imputação IM3 e análise (STATA 11)

Como ilustração, apresenta-se a sintaxe dos comandos necessários para o modelo de imputação IM3 e, subsequentemente, para o ajuste dos 6 modelos logísticos no banco pós-imputação. No modelo IM3, são incluídas, além do próprio **IMC_aferido***, que é o foco desse processo de imputação, as três variáveis dependentes (**Diabetes**, **Hipertensão** e **Peso_acima**), demais covariáveis dos modelos logísticos (**Cor**, **Escolaridade** e **Idade**) e 16 variáveis preditoras do **IMC_aferido*** ou da variável indicadora de perdas do **IMC_referido**. Apesar do **IMC_aferido*** ser tratado como variável de interesse, todas as demais variáveis com perdas terão seus valores ausentes imputados. Para isso, o comando "mi ice" do STATA 11, que executa as imputações, requer a especificação prévia das variáveis com valores ausentes. O **IMC_aferido*** foi a variável mais impactada pelas perdas, apresentando 789 valores ausentes, seguida da **Renda_familiar** (97 perdas) e **Colesterol_alto** (27 perdas). Abaixo, os comandos necessários para instalação no STATA 11 dos pacotes "mi ice", "mim", que é utilizado nas análises posteriores à imputação, e "mvpatterns", que descreve a distribuição das perdas na base de dados:

```
net install mi_ice, from(http://www.homepages.ucl.ac.uk/~ucakjpr/stata)
net install mim, from(http://www.homepages.ucl.ac.uk/~ucakjpr/stata)
net install dm91, from(http://www.stata.com/stb/stb61)
```

As imputações em múltiplas variáveis são realizadas por um processo iterativo de regressões (ver artigo para referências completas). De modo simplista, o processo de imputação, que envolve apenas as variáveis com dados ausentes, pode ser ilustrado da seguinte forma: (Passo 1) Os valores ausentes de cada variável são substituídos por valores iniciais, que são amostrados aleatóriamente dos valores observados da própria variável; (Passo 2) para cada variável, seguindo uma ordem crescente de acordo com a proporção de perdas, são realizadas as seguintes etapas: (2.1) os valores imputados da variável são excluídos; (2.2) ajusta-se a distribuição preditiva dos valores ausentes, condicionada nos valores de todas as demais variáveis (todas as variáveis selecionadas para o processo de imputação, inclusive as que não tinham perdas inicialmente) por meio de modelos de regressão (linear, logístico, multinomial ou ordinal); (2.3) cada valor ausente é substituído por um valor amostrado da distribuição preditiva. O Passo 2 é repetido até que a distribuição dos valores imputados se estabilize (o comando "mi ice" do STATA 11 adota como padrão 10 repetições) e, assim, a primeira base de dados completa é salva. Todo esse processo é repetido M (número de imputações) vezes, gerando M bancos completos.

Na sintaxe, comandos distintos estão separados por um parágrafo e os comentários (em itálico) são precedidos por um asterisco e descrevem a função de cada comando. Os comandos apresentados a seguir descrevem a preparação do banco para o processo de imputação.

```
*seleciona pasta na qual estão os bancos de dados:
cd "E:\mestrado\bancos\"

*abre banco de dados original (sem imputações):
use "banco_original.dta", clear

*comando que configura o desenho amostral:
svyset GEO_SETOR [pweight=A_PESO_ADULTO], fpc(A_SETOR_ESTR) strata(A_ESTRATO) ||
 _n,  singleunit(scaled)

*criando variável indicadora que será utilizada para restringir as análises aos
*indivíduos com IMC_aferido observado e mulheres não grávidas:
gen SUB=GRAVIDA!=1&IMC_aferido!=.

*criando a mesma indicadora para cada sexo:
gen SUB_FEM=Sexo==0&SUB==1
gen SUB_MAS=Sexo==1&SUB==1

*criando variáveis indicadoras que serão utilizadas para restringir as análises
*aos indivíduos com dados completos nas variáveis do modelo de referência (tabela
*2 do artigo):
egen Peso_acima_REF=rowmiss(Peso_acima IMC_aferido Idade Cor Escolaridade)
egen diabetes_REF=rowmiss(Diabetes IMC_aferido Idade Cor Escolaridade)
egen Hipertensao_REF=rowmiss(Hipertensao IMC_aferido Idade Cor Escolaridade)
recode Peso_acima_REF diabetes_REF Hipertensao_REF (0=1) (1/5=0)

*criando nova variável do IMC_aferido com perdas que refletem as perdas do
*IMC_referido (IMC_aferido_p)
gen IMC_aferido_p=IMC_aferido
replace IMC_aferido_p=. if IMC_referido==.

*criando variável com o logaritmo do IMC_aferido_p (LN_IMC_aferido_p)
gen LN_IMC_aferido_p=ln(IMC_aferido_p)

*criando variável indicadora de sobrepeso:
gen Sobrepeso=(IMC_aferido_p>=25)&(IMC_aferido_p<30)
replace Sobrepeso=. if IMC_aferido_p==.

*criando variável indicadora de obesidade:
gen Obesidade=(IMC_aferido_p>=30)
replace Obesidade=. if IMC_aferido_p==.

*criando variável que nas análises representará o intercepto do modelo logístico:
gen cons=1

*mantém no banco original apenas as variáveis que serão utilizadas nas análises,
*no modelo de imputação e também variáveis relativas ao desenho amostral
*(ID_DOMICILIO GEO_SETOR SETOR A_PESO_ADULTO A_ESTRATO A_SETOR_ESTR). Essa
*redução de variáveis não é necessária, mas acelera o processo de imputação:
keep LN_IMC_aferido_p Sobrepeso Obesidade Peso_acima Diabetes Hipertensao cons
Idade Cor Escolaridade Silhueta Silhueta2 Satisfacao_corporal Estado_civil
Inativo Doenca_mental Renda_familiar Mudanca_peso Colesterol_alto Fumo Artrite
Chefe_familia Posse_veiculo Restricao_alimentar Consumo_fruta Epilepsia Sexo SUB
SUB_FEM SUB_MAS Peso_acima_REF Diabetes_REF Hipertensao_REF ID_DOMICILIO
GEO_SETOR SETOR A_PESO_ADULTO A_ESTRATO A_SETOR_ESTR
```

Configurando a imputação:

```
*descreve o padrão de perdas das variáveis (este comando permite observar quais
*as variáveis com perdas e também os padrões em que essas mesmas ocorrem):
mvpatterns LN_IMC_aferido_p Sobrepeso Obesidade Peso_acima Diabetes Hipertensao
cons Idade Cor Escolaridade Silhueta Silhueta2 Satisfacao_corporal Estado_civil
Inativo Doenca_mental Renda_familiar Mudanca_peso Colesterol_alto Fumo Artrite
Chefe_familia Posse_veiculo Restricao_alimentar Consumo_fruta Epilepsia Sexo if
SUB, sort nodrop

*define o formato do banco de dados após as imputações. O formato flong ("full
*long") gera M (número de imputações) bancos de dados completos e armazena todos
*eles na mesma base, concatenando as bases uma abaixo da outra. Variáveis
*identificadoras permitem localizar cada base de dados. A base de dados original,
*ainda com perdas, permanece como a primeira base do banco seguida de M bases
*completas:
mi set flong

*define as variáveis que possuem dados ausentes (todas elas passarão pelo
*processo de imputação):
mi register imputed LN_IMC_aferido_p Sobrepeso Obesidade Cor Escolaridade
Peso_acima Satisfacao_corporal Silhueta Hipertensao Diabetes Inativo
Renda_familiar Mudanca_peso Colesterol_alto Artrite Chefe_familia Consumo_fruta
```

O processo de imputação é executado pelo comando "mi ice":

```
*Comando de imputação com a opção "dryrun", que não efetiva as imputações, mas
*produz uma saída que permite avaliar como cada variável com dados ausentes está
*sendo modelada:
mi ice LN_IMC_aferido_p Sobrepeso Obesidade Peso_acima i.Idade m.Cor
m.Escolaridade m.Satisfacao_corporal Silhueta Silhueta2 Hipertensao
i.Estado_civil Diabetes Inativo Doenca_mental m.Renda_familiar Mudanca_peso
Colesterol_alto Fumo Artrite Chefe_familia Posse_veiculo Restricao_alimentar
m.Consumo_fruta Epilepsia i.A_ESTRATO if SUB [pweight=A_PESO_ADULTO], by(Sexo)
add(20) passive(Sobrepeso:(LN_IMC_aferido_p<log(30))&(LN_IMC_aferido_p>=log(25))
\ Obesidade:LN_IMC_aferido_p>=log(30) \Silhueta2:Silhueta^2)
substitute(LN_IMC_aferido_p:Sobrepeso Obesidade) conditional(Mudanca_peso:
Peso_acima==1) dryrun

*Realizando as imputações (mesmo comando anterior sem a opção dryrun).
mi ice LN_IMC_aferido_p Sobrepeso Obesidade Peso_acima i.Idade m.Cor
m.Escolaridade m.Satisfacao_corporal Silhueta Silhueta2 Hipertensao
i.Estado_civil Diabetes Inativo Doenca_mental m.Renda_familiar Mudanca_peso
Colesterol_alto Fumo Artrite Chefe_familia Posse_veiculo Restricao_alimentar
m.Consumo_fruta Epilepsia i.A_ESTRATO if SUB [pweight=A_PESO_ADULTO], by(Sexo)
add(20) passive(Sobrepeso:(LN_IMC_aferido_p<log(30))&(LN_IMC_aferido_p>=log(25))
\ Obesidade:LN_IMC_aferido_p>=log(30) \Silhueta2:Silhueta^2)
substitute(LN_IMC_aferido_p:Sobrepeso Obesidade) conditional(Mudanca_peso:
Peso_acima==1)

*Re-configurando a base de dados para as análises:
mi export ice, clear

*Salvando banco final
save "banco_IM3.dta", replace
```

Após o prefixo “mi ice”, são listadas as variáveis que participarão do processo iterativo de imputações. Cada variável com perdas será modelada como variável dependente de um modelo de regressão (linear, logístico, multinomial ou ordinal) no qual as demais variáveis da lista são tratadas como covariáveis. Como padrão, variáveis binárias (0,1) que possuem valores ausentes são modeladas no processo de imputação por meio de regressões logísticas que utilizam todas as demais variáveis listadas como covariáveis. Também como padrão, todas as variáveis não binárias com dados ausentes são tratadas como contínuas e modeladas por regressões lineares. Variáveis categóricas com valores ausentes podem ser modeladas por regressões logísticas multinomiais, adicionando-se o prefixo “m.” ao nome da variável na lista, ou ordinais, adicionando-se o prefixo “o.”. Variáveis que não possuem dados ausentes participarão apenas como covariáveis nesses modelos de regressão. Para que essas últimas sejam tratadas como covariáveis categóricas o prefixo “i.” deve ser adicionado. Variáveis binárias (0,1) sem perdas são automaticamente tratadas como categóricas e não precisam desse prefixo.

O número de imputações é definido pela opção “add” (e.g., “add(20)” para 20 imputações). A opção “by” permite realizar as imputações independentemente para subgrupos da base de dados (e.g. **Sexo**). Quando uma variável com dados ausentes é utilizada como base para criação de outras variáveis (e.g., a variável Silhueta2, que é simplesmente a variável silhueta ao quadrado) a opção “passive” faz com que os valores imputados na variável original sejam propagados para as variáveis derivadas, mantendo assim a consistência do banco de dados. A opção “substitute” pode ser utilizada, por exemplo, para imputar uma variável na sua forma contínua (por um modelo linear) e tratá-la como categórica quando a mesma for utilizada como covariável no modelo de imputação de outras variáveis.

Algumas variáveis podem requerer ainda um processo de imputação condicional. Considere, por exemplo, que as variáveis **Sexo** (1 feminino, 0 masculino) e **Gravidez** (1 sim, 0 não, 99 não se aplica) possuam dados ausentes. A dependência implícita entre essas variáveis determina que quando um valor ausente do **Sexo** for substituído por zero no processo de imputação, o valor da **Gravidez** deverá ser substituído por 99. Essas dependências podem ser especificadas com a opção “conditional”. Deve-se garantir, no entanto, que todo indivíduo com valor ausente na variável **Sexo** tenha valor ausente na variável **Gravidez**. É ainda necessário garantir que sempre que a condição de aplicabilidade imposta seja falsa (**Sexo** masculino) que a segunda variável tenha um único valor (por exemplo, 99) na base de dados.

No exemplo dado, a opção adicionada ao comando seria "conditional(**Gravidez**:**Sexo**==1)", ou seja, a variável **Gravidez** só é aplicável para o sexo feminino (**Sexo**==1).

Na sintaxe apresentada, a opção "if **SUB**" restringe as imputações ao sub-grupo de indivíduos com **IMC_aferido** observado e mulheres não grávidas (**SUB**==1). A ponderação amostral é considerada pela opção "[pweight=**A_PESO_ADULTO**]" e os estratos amostrais são incluídos como variável categórica (i.**A_ESTRATO**).

Comandos das análises:

Na sintaxe dos modelos logísticos apresentada a seguir, o prefixo "xi:" indica a presença de covariáveis categóricas, o prefixo "mim:" que o banco de dados é um banco com imputações múltiplas e o prefixo "svy" que o desenho amostral deve ser considerado nas análises. A opção "subpop" após a vírgula restringe as análises aos indivíduos com dados completos nos modelos de referência (Tabela 2 do artigo) e ainda para o sexo masculino ou feminino (**SUB_FEM** ou **SUB_MAS**). A opção "or" indica que os resultados devem fornecer as estimativas da *Odds Ratio* (como padrão os coeficientes não transformados são apresentados). A adição da covariável **cons** (que possui valor 1 para todos os indivíduos) juntamente com a opção "nocons" garante que a estimativa do intercepto também apareça nos resultados.

```
*Modelo logístico da variável dependente Peso_acima, sexo feminino
xi:mim:svy, subpop(if SUB_FEM&Peso_acima_REF):logit Peso_acima cons Sobrepeso
Obesidade i.Idade i.Cor i.Escolaridade, or nocons

*Modelo logístico da variável dependente Peso_acima, sexo masculino
xi:mim:svy, subpop(if SUB_MAS&Peso_acima_REF):logit Peso_acima cons Sobrepeso
Obesidade i.Idade i.Cor i.Escolaridade, or nocons

*Modelo logístico da variável dependente Diabetes, sexo feminino
xi:mim:svy, subpop(if SUB_FEM&Diabetes_REF):logit Diabetes cons Sobrepeso
Obesidade i.Idade i.Cor i.Escolaridade, or nocons

*Modelo logístico da variável dependente Diabetes, sexo masculino
xi:mim:svy, subpop(if SUB_MAS&Diabetes_REF):logit Diabetes cons Sobrepeso
Obesidade i.Idade i.Cor i.Escolaridade, or nocons

*Modelo logístico da variável dependente Hipertensão, sexo feminino
xi:mim:svy, subpop(if SUB_FEM&Hipertensao_REF):logit Hipertensao cons Sobrepeso
Obesidade i.Idade i.Cor i.Escolaridade, or nocons

*Modelo logístico da variável dependente Hipertensão, sexo masculino
xi:mim:svy, subpop(if SUB_MAS&Hipertensao_REF):logit Hipertensao cons Sobrepeso
Obesidade i.Idade i.Cor i.Escolaridade, or nocons
```

APÊNDICE B - Projeto de Pesquisa

Vitor Passos Camargos

Revisão dos métodos para analise de dados incompletos “missing data”: Projeto MOVE-SE e Determinantes Sociais de BH

Plano de trabalho submetido à seleção de mestrado (2009) do Programa de Pós-graduação em Saúde Pública da Universidade Federal de Minas Gerais

Orientador: Fernando Augusto Proietti
Co-orientadora: Cibele Comini César

BELO HORIZONTE
2008

Revisão dos métodos para analise de dados incompletos “missing data”: Projeto MOVE-SE e Determinantes Sociais de BH

I. Introdução

Um problema inerente aos grandes inquéritos na saúde pública e outras áreas de pesquisas humanas[4] é a não resposta a determinados itens. As características dos itens com não resposta/perda normalmente sugerem hipóteses sobre o processo causal das mesmas. Uma hipótese comum está relacionada a questões consideradas excessivamente delicadas – uso de drogas, comportamento sexual e renda, por exemplo – nas quais as perdas ocorre geralmente em uma população diferenciada do estudo. E em outros casos a perda pode estar ligada a fatores mais aleatórios como: falhas em medições, erros de digitação, perguntas demasiadamente complexas e questões não aplicáveis. O procedimento tradicionalmente utilizado nas análises é a exclusão de indivíduos com dados incompletos, que podemos chamar de análise de dados completos. A aplicação deste procedimento é induzida pelos softwares de análise, já que a maioria define o procedimento como padrão – talvez pela falta de outros procedimentos tão práticos e de uso generalizado. Seu uso é tão banalizado que muitos desconhecem as suposições nas quais o procedimento produz resultados válidos.

Um dos problemas relativos à exclusão dos indivíduos com dados incompletos é obviamente a perda das informações coletadas do mesmo, além da expressiva perda de indivíduos quando muitas variáveis com perdas não coincidentes são incluídas na análise. Isto causa minimamente um problema, a perda de poder. Quando os indivíduos com dados completos não podem ser considerados como uma amostra aleatória de todos os indivíduos do estudo, os resultados das análises utilizando este procedimento podem estar viciados. Este viés aumenta à medida que se reduz a proporção de indivíduos com dados completos ou quando as características dos indivíduos excluídos da análise – dados incompletos – diferem consideravelmente dos demais. Problemas ainda maiores ocorrem quando existe a suspeita de que a não resposta de determinados itens é induzida pela qualidade da própria resposta.[1,2] Sem a aplicação de metodologias mais avançadas a perda sistemática em questões de interesse das pesquisas pode levar a uma redução forçada do número de variáveis a serem consideradas nas análises – para evitar perda excessiva de poder -, a uma mudança no foco do estudo, ou mesmo a um estudo com conclusões limitadas.

A parte de pequenos avanços na aplicação de métodos mais efetivos para dados incompletos em estudos longitudinais, e, de diversas metodologias - com propriedades mais desejáveis que a simples exclusão dos indivíduos - terem sido desenvolvidas há algum tempo, é predominante a aplicação da análise de dados completos. A falta de referências bibliográficas em português e o crescente desenvolvimento de novas metodologias na área acabam por inibir a aplicação desses métodos no Brasil.

As perdas de todos os itens de uma pessoa, que pode ocorrer devido à recusa, por exemplo, é considerado um tipo especial de perda e por ter extensa bibliografia relacionada não será objeto do presente estudo.

II. Objetivos

Revisar as metodologias existentes para lidar com a análise de bancos de dados incompletos exemplificando sua aplicação com dados reais resultantes do projeto Move-se/Determinantes Sociais de Belo Horizonte. A aplicação e comparação das metodologias estarão vinculadas ao modelo de regressão generalizado – a ser definido – abordando variáveis que potencializem os efeitos dos métodos, ou seja, aquelas que apresentam um maior nível de perdas e que também sejam importantes no âmbito da saúde pública. Por fim o estudo se propõe a fazer um analise crítica sobre a aplicabilidade de cada método.

Objetivos Específicos

1-Apresentar aspectos teóricos sobre os tipos de perdas nos estudos e conceitos que definem possíveis processos causais relacionados às perdas – perdas completamente ao acaso, perdas ao acaso e perdas não devidas ao acaso – e suas implicações nas análises.
2-Aplicar em dados reais dos seguintes métodos de análise de dados incompletos: Exclusão de casos incompletos; Ponderação de casos completos; Método da verossimilhança; Imputação Múltipla.
3-Descrever a aplicação dos métodos de forma didática utilizando os softwares R e Stata com referencias às sintaxes de cada um, com vistas a uma disseminação do uso.

III. Proposta metodológica

O estudo terá como base para aplicação das metodologias o inquérito de saúde realizado pelo Observatório de Saúde Urbana de Belo Horizonte (OSUBH) da Universidade Federal de Minas Gerais (UFMG). Este inquérito tem como objetivos principais determinar os modos, estilos e hábitos de vida da população residente em Belo Horizonte (Projeto Move-se BH; CNPq 02/2006 – Universal - #475004/2006-0 e Fundação Nacional de Saúde, Ministério da Saúde) e avaliar os determinantes sociais nesta população (Projeto Determinantes Sociais de BH; Edital MCT- CNPq / MS-SCTIE-DECIT – Nº 26/2006). A Pesquisa está em campo desde o mês de agosto de 2008, e tem previsão para o termino das entrevistas em janeiro de 2009.

O inquérito de saúde abrange os distritos Barreiro e Oeste, dois dos nove distritos sanitários (DS), em que é dividido o município de Belo Horizonte. A base para o sorteio das amostras foram os domicílios cadastrados pela base de informações da SMSA-BH e participam aqueles que tenham pelo menos um morador com mais de 18 anos. A metodologia adotada foi uma amostragem estratificada – proporcional - por conglomerados em três estágios: setor censitário, domicílio, 1 morador adulto e 1 morador na faixa de 11 a 17 anos. Três questionários distintos estão em campo: Um para o sorteado adulto (18 anos ou mais) aplicado por meio de entrevista face a face, um auto-aplicável para adolescentes de 11 a 13 anos e outro confidencial para adolescentes de 14 a 17 anos também auto-aplicável. Medidas diretas como peso, altura e circunferência da cintura também serão avaliadas. Ao fim do projeto são esperadas aproximadamente 4500 entrevistas concluídas para os adultos e 1500 para os adolescentes (11 a 17 anos).

O Inquérito irá produzir uma quantidade enorme de variáveis potenciais para aplicações de metodologias para análise de dados incompletos. Alguns temas são: uso de drogas, violência doméstica e renda. Além desses, o questionário confidencial do adolesceste, que também traz temas muito particulares sobre do ambiente doméstico e experiências de vida, deve ser uma fonte importante para o estudo.

A aplicação das metodologias para análise de dados incompletos - Exclusão de casos incompletos; Ponderação de casos completos; Método da verossimilhança; Imputação Múltipla - se darão em conjunto com modelos de regressão generalizados utilizando os

softwares estatísticos R e STATA. , Apesar da necessidade de pacotes adicionais, as metodologias citadas já estão disponíveis para ambos os softwares.

1º Artigo: Revisão dos métodos para analise de dados incompletos.

2º Artigo: Aplicação e comparação das metodologias de análise de dados incompletos associadas aos modelos de regressão generalizados com dados provenientes do projeto MOVE-SE/DERMINANTES SOCIAIS.

Questões éticas: Esta proposta de estudo foi submetida ao Comitê de Ética em Pesquisa da UFMG, seguido de todos os preceitos éticos e aprovada em 19 de abril de 2007 (parecer nº ETIC 017/07) e 16 de outubro de 2006 (Parecer no ETIC 253/06) (Anexo 1 e 2).

IV. Viabilidade (custo/cronograma)

O Projeto MOVE-SE BH é financiado pelo Ministério da Saúde e pelo CNPq, não havendo custos adicionais que possam impedir ou interromper a execução do mesmo.

Atividades	**Semestre anterior**	**1º ano**		**2º ano**	
		1º	**2º**	**1º**	**2º**
Revisão Bibliográfica	X	X	X	X	X
Disciplinas	X	X	X		
Elaboração dos instrumentos	X				
Coleta de dados		X			
Análise dos dados			X	X	
Elaboração de artigos			X	X	
Defesa de dissertação					X

V. Referências Bibliográficas

1. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002;7(2):147-77.

2. Raghunathan TE. What do we do with missing data? some options for analysis of incomplete data. Annu Rev Public Health. 2004;25:99-117.

3. Schafer JL. Multiple imputation: a primer. Statistical Methods in Medical Research. 1999;8:3-15.

4. Molenberghs G. Editorial: What to do with missing data?. Journal Of The Royal Statistical Society Series A. 2007;170(4):861-863.

ANEXO A – Recibo de submissão artigo


Orkut Gmail Calendar Documents Web more ▾
vitorcamargos@
Gmail by Google
Search Mail
Search the Web
Show search options
Create a filter
Mail
Contacts
Tasks
Compose mail
Inbox (590)
Buzz
Starred
Sent Mail
Drafts (1)
Personal
Travel
6 more ▾
Chat
Search, add, or invite
vitor camargos
Sign into chat
Adriana Meireles
Renato Faria
Ricardo Castro
elci
joaojorge
junior
manutencao
proietti
ricardolanap
vitorpassos
Invite a friend
Give Gmail to:
Send Invite 98 left
Higienização de Dados - www.datawash.com.br - Ferramenta completa e amigável Ambientes Windows, Linux e
« Back to Inbox
Archive
Report spam
Delete
Move to ▾
Labels ▾
More actions ▾
Novo artigo (CSP_0077/11)
Inbox | X
from Cadernos de Saude Publica <cadernos@ensp.fiocruz.br>
hide details Jan 19 (2 days ago)
Reply
to vitorcamargos@gmail.com
date Wed, Jan 19, 2011 at 1:05 PM
subject Novo artigo (CSP_0077/11)
mailed-by ensp.fiocruz.br
Prezado(a) Dr(a). Vitor Passos Camargos:
Confirmamos a submissão do seu artigo "Imputação múltipla e análise de casos completos em modelos de regressão logística: Uma avaliação prática do impacto das perdas em covariáveis" (CSP_0077/11) para Cadernos de Saúde Pública. Agora será possível acompanhar o progresso de seu manuscrito dentro do processo editorial, bastando clicar no link "Sistema de Avaliação e Gerenciamento de Artigos", localizado em nossa página http://www.ensp.fiocruz.br/csp.
Em caso de dúvidas, envie suas questões através do nosso sistema, utilizando sempre o ID do manuscrito informado acima. Agradecemos por considerar nossa revista para a submissão de seu trabalho.
Atenciosamente,
Prof. Carlos E.A. Coimbra Jr.
Prof. Mario Vianna Vettore
Editores
CSP
Cadernos de Saúde Pública / Reports in Public Health
Escola Nacional de Saúde Pública Sergio Arouca
Fundação Oswaldo Cruz
Rua Leopoldo Bulhões 1480
Rio de Janeiro, RJ 21041-210, Brasil
Tel.: +55 (21) 2598-2511, 2508 / Fax: +55 (21) 2598-2737
cadernos@ensp.fiocruz.br
http://www.ensp.fiocruz.br/csp

ANEXO B - Aprovação do Comitê de Ética

Universidade Federal de Minas Gerais
Comitê de Ética em Pesquisa da UFMG - COEP

Parecer nº. ETIC 253/06

Interessado: Profa. Waleska Teixeira Caiaffa
Departamento de Medicina Preventiva e Social
Faculdade de Medicina - UFMG

DECISÃO

O Comitê de Ética em Pesquisa da UFMG – COEP aprovou, *ad referendum*, no dia 16 de outubro de 2006, após atendidas as solicitações de diligência, o projeto de pesquisa intitulado **"Análise dos fatores condicionantes da saúde da população por áreas delimitadas e formulação de propostas de intervenção: Projeto modos de vida, estilos e hábitos saudáveis em BH (Projeto Move-se BH) - Uma avaliação epidemiológica"** bem como o Termo de Consentimento Livre e Esclarecido do referido projeto.

O relatório final ou parcial deverá ser encaminhado ao COEP um ano após o início do projeto.

Profa. Dra. Maria Elena de Lima Perez Garcia
Presidente do COEP/UFMG

Av. Presidente Antônio Carlos, 6627, Unidade Administrativa II – 2º andar sala: 2005 - 31.270-901 - BH - MG
(31) 3499-4592 - FAX: (31) 3499-4516 - coep@prpq.ufmg.br

UFMG | Universidade Federal de Minas Gerais
Comitê de Ética em Pesquisa da UFMG - COEP

Parecer nº. ETIC 017/07

Interessado(a): Prof. Fernando Augusto Proietti
Departamento de Medicina Preventiva e Social
Faculdade de Medicina-UFMG

DECISÃO

O Comitê de Ética em Pesquisa da UFMG – COEP aprovou, no dia 9 de abril de 2007, após atendidas as solicitações de diligência, o rojeto de pesquisa intitulado **"Observatório de saúde urbana de Belo lorizonte – análise dos fatores condicionantes da saúde da opulação por áreas delimitadas e formulação de propostas de ntervenção. Saúde urbana"** bem como o Termo de Consentimento ivre e Esclarecido.

O relatório final ou parcial deverá ser encaminhado ao COEP um no após o início do projeto.

Profa. Dra. Maria Elena de Lima Perez Garcia
Presidente do COEP-UFMG

ANEXO C - Certificado de Qualificação

FACULDADE DE MEDICINA
CENTRO DE PÓS-GRADUAÇÃO
Av. Prof. Alfredo Balena 190 / sala 533
Belo Horizonte - MG - CEP 30.130-100
Fone: (031) 3409.9641 FAX: (31) 3409.9640
cpg@medicina.ufmg.br

UFMG

Ata do exame de qualificação a que se submeteu o Mestrando VITOR PASSOS CAMARGOS.

Aos onze dias do mês de dezembro de dois mil e nove, convocada pelo Colegiado do Programa de Pós-Graduação em Saúde Pública - Área de Concentração em Epidemiologia compareceu o mestrando **VITOR PASSOS CAMARGOS** para submeter-se ao exame de qualificação com o projeto de dissertação intitulada: **"REVISÃO DOS METODOS PARA ANALISE DE DADOS INCOMPLETOS EM ESTUDOS TRANVERSAIS EM SAÚDE"**, perante a Comissão Examinadora composta pelos professores: Fernando Augusto Proietti- UFMG – Cibele Comini César/ Coorientadora - UFMG, Waleska Teixeira Caiaffa – UFMG , Edna Afonso Reis – UFMG, Participaram como ouvintes da sessão, o Prof. Fernando Augusto Proietti/orientador e a Profa. Cibele Comini César/ Coorientadora da dissertação. A sessão iniciou-se às oito horas, na sala 705-, 7° andar da Faculdade de Medicina com a presença dos professores acima citados. Após a exposição da candidata, os professores participantes da Comissão Examinadora fizeram comentários sobre a apresentação oral, do conteúdo, relevância, metodologia e viabilidade do Projeto. Após a argüição a banca examinadora considerou o Projeto coerente e o aluno apto a prosseguir a sua investigação. Para constar, lavrou-se a presente ATA, que segue assinada pela comissão examinadora. Belo Horizonte, 11 de dezembro de 2009.

Profa. Waleska Teixeira Caiaffa _____

Profa. Edna Afondo Reis _____

Profa. Mariângela Leal Cherchiglia (coordenadora) _____

CONFERE COM O ORIGINAL
Centro de Pós-Graduação

Universidade Federal de Minas Gerais
Faculdade de Medicina
Programa de Pós-Graduação em Saúde Pública
Seminários em Saúde Coletiva

Certificado

Certifico que**Vitor Passos Camargos,** participou do *Seminários em Saúde Coletiva* promovido pelo Programa de Pós-Graduação em Saúde Pública, apresentando o projeto de dissertação, **"Revisão Dos Metodos para Analise de Dados Incompletos em Estudos Tranversais em Saúde."**

Belo Horizonte, 11 de Dezembro de 2009

Profª Mariângela Leal Cherchiglia
Coordenadora do Programa de Pós-Graduação em Saúde Pública