CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

COMUNICAÇÃO BREVE
BRIEF COMMUNICATION

Predictive Mean Matching como método de imputação alternativo ao *hot deck* no Vigitel

Predictive Mean Matching as an alternative imputation method to hot deck in Vigitel

Predictive Mean Matching como método de imputación alternativo al *hot deck* en Vigitel

Iolanda Karla Santana dos Santos [1,2]
Wolney Lisbôa Conde [1]

doi: 10.1590/0102-311X00167219

Resumo

O objetivo deste estudo foi descrever a estimativa das médias de peso, altura e índice de massa corporal (IMC) segundo dois métodos de imputação, usando dados do Vigitel (Vigilância de Fatores de Risco e Proteção para Doenças Crônicas por Inquérito Telefônico)*. O delineamento do estudo é transversal e utilizaram-se dados secundários do Vigitel do período de 2006 a 2017. Os dois métodos para imputação utilizados no estudo foram* hot deck *e* Predictive Mean Matching (*PMM). As variáveis peso e altura imputadas por* hot deck *foram disponibilizadas pelo Vigitel. Dois modelos foram conduzidos com a utilização da PMM: (i) variáveis explicativas – cidade, sexo, idade em anos, raça/cor e escolaridade; (ii) variáveis explicativas – cidade, sexo e idade em anos. Nos dois modelos, as variáveis peso e altura foram as variáveis de desfecho. Na PMM, combinam-se regressão linear e seleção aleatória de valor para imputação. A predição linear é usada como medida de distância entre o valor faltante e os seus possíveis doadores e, com isso, se cria o espaço virtual com os casos candidatos a ceder o valor para imputação. Um dos candidatos do* pool *é aleatoriamente selecionado, e o seu valor é atribuído à unidade faltante. O IMC foi calculado por meio da divisão do peso em quilogramas pela altura ao quadrado. Nos resultados, apresentamos as médias e erros-padrão de peso, altura e IMC, segundo método de imputação e ano de monitoramento. Nas estimativas, utilizou-se o módulo* survey *do Stata, que considera os efeitos da amostragem. Observou-se que os valores médios de peso, altura e IMC estimados por* hot deck *e PMM são similares. Os resultados com os dados do Vigitel sugerem a aplicabilidade do PMM ao conjunto dos inquéritos de saúde.*

Inquéritos Nutricionais; Vigilância; Estado Nutricional; Epidemiologia

Correspondência
I. K. S. Santos
Faculdade de Saúde Pública, Universidade de São Paulo.
Av. Dr. Arnaldo 715, São Paulo, SP 02361-100, Brasil.
iolanda.santos@usp.br

[1] Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, Brasil.
[2] Fundação Universidade Federal do ABC, Santo André, Brasil.

Introdução

A ocorrência de dados classificados como faltantes é um problema comum em inquéritos populacionais. Diversas situações podem levar a ausência de informação como, por exemplo, a recusa do indivíduo em fornecer respostas a determinadas questões ou em participar de alguma etapa do estudo. A ausência de dados pode levar a redução do poder estatístico e da precisão das estimativas. Por outro lado, o manejo inadequado de dados faltantes pode enviesar a estimativa de parâmetros (média, desvio padrão, coeficientes de modelos de regressão). Métodos de imputação têm sido amplamente adotados nos últimos anos para lidar com dados faltantes, visto que as técnicas de imputação se propõem a substituir esses valores por outros plausíveis [1,2,3]. De acordo com van Buuren [4], na imputação, o objetivo é aumentar os dados e preservar as relações observadas entre eles.

Nesse sentido, o sistema de *Vigilância de Fatores de Risco e Proteção para Doenças Crônicas por Inquérito Telefônico* (Vigitel), que tem como objetivo monitorar a frequência e distribuição dos principais determinantes das doenças crônicas não transmissíveis, utiliza método de imputação denominado *hot deck* [5] como tratamento para não resposta das variáveis peso e altura. Nesse método, os valores faltantes de uma ou mais variáveis para indivíduos não respondentes são substituídos por valores observados entre indivíduos respondentes que apresentam características similares aos não respondentes.

Na literatura são descritas diferentes maneiras de aplicação do *hot deck*. Considerando a seleção do doador, podemos dividi-las em: (i) doador selecionado aleatoriamente a partir de um conjunto de potenciais doadores; (ii) um único doador identificado baseado em método determinístico [6]. A maneira mais simples de executar essa técnica é classificar as unidades em respondentes e não respondentes dentro de classes de imputação baseadas em covariáveis. A imputação é realizada por meio da seleção aleatória de um valor observado entre os indivíduos respondentes pertencentes a uma mesma classe para cada não respondente [6]. Em síntese, a principal desvantagem para aplicação desse método é a ausência de critérios claros para guiar a seleção do doador a partir dos casos completos [7].

Entre os métodos de imputação disponíveis em programas estatísticos, destaca-se a *Predictive Mean Matching* (PMM) [8], que é uma técnica de imputação múltipla semiparamétrica na qual se combinam regressão linear e seleção aleatória de valor para imputação. Assim, a variável de interesse é a variável a ser imputada, e as demais variáveis são as explicativas. A predição linear é utilizada como medida de distância entre o valor faltante e os seus possíveis doadores e, com isso, se cria o espaço virtual com os casos candidatos a ceder o valor para imputação. Valores observados próximos ao valor predito são selecionados como o *pool* doador que, frequentemente, é fixado contendo *k* candidatos doadores. Um desses candidatos é aleatoriamente selecionado, e o seu valor é atribuído à unidade faltante. Por utilizar dados observados, a PMM preserva a distribuição dos valores observados entre os valores imputados. Ao realizar a imputação, é aconselhável incluir no modelo alguns tipos de variáveis explicativas: (i) variáveis que predizem a ausência de valores; (ii) variáveis associadas à variável que está sendo imputada; (iii) variável de desfecho do modelo [2,3,9]. Allison [10] indica duas grandes vantagens dos métodos de imputação múltipla: (i) teoricamente, podem ser aplicados a qualquer tipo de dado ou modelo, e (ii) estão disponíveis em programas estatísticos convencionais.

Considerando a ausência de clareza metodológica em relação à imputação por *hot deck* na publicação do Vigitel e a disponibilidade do método PMM em diversos pacotes estatísticos, o objetivo deste estudo foi descrever a estimativa das médias de peso, altura e índice de massa corporal (IMC) segundo dois métodos de imputação, usando dados do Vigitel.

Método

O delineamento do estudo é transversal e utilizaram-se dados secundários do Vigitel do período de 2006 a 2017. Os aspectos relativos à metodologia de pesquisa do Vigitel estão disponíveis em publicação oficial [5], porém não localizamos na publicação o mecanismo de dados faltantes utilizado como pressuposto para imputação. O método *hot deck* utilizado pelos pesquisadores do Vigitel requer como premissa que o mecanismo de dados faltantes ocorra de maneira totalmente aleatória (*missing completely at random* – MCAR), o que é um pressuposto muito rigoroso [6,7]. Na prática, o mecanismo de dados faltantes raramente é totalmente conhecido [2,11]. Utilizamos como premissa que o mecanismo

de dados faltantes é aleatório (*missing at random* – MAR), ou seja, é condicional aos dados observados, mas não depende de dados não observados [2]. MCAR é um tipo específico de MAR [2] para testar indiretamente se os dados são do tipo MCAR ou MAR; inicialmente criamos uma variável indicadora de dado faltante para as variáveis peso e altura e, em seguida, conduzimos análises de regressão logística tendo como covariáveis sexo, idade, raça/cor, cidade e escolaridade [12]. Em geral, observou-se associação significativa entre as variáveis indicadoras de dado faltante e as variáveis explanatórias em todos os anos de inquérito. Portanto, isso favorece nossa hipótese de que o mecanismo seja do tipo MAR. Em nossas análises, não testamos a hipótese de mecanismo MAR *versus* mecanismo de dados faltantes não aleatório (*missing not at random* – MNAR), pois necessitaríamos de informações adicionais que não estão disponíveis [12].

No Vigitel, a imputação das variáveis peso e altura é realizada pelo método *hot deck*. Conforme divulgado na publicação oficial, inicialmente, investigou-se a associação entre a ausência de resposta e as variáveis idade, sexo, escolaridade e raça/cor. Grupos de respondentes e não respondentes com características semelhantes para as variáveis preditoras da condição de não resposta foram criados. Em seguida, em cada capital selecionou-se aleatoriamente, dentro de cada grupo, um indivíduo com valores conhecidos que doou seus valores de peso e altura para o não respondente pertencente ao mesmo grupo [5]. As variáveis de peso e altura imputadas pelo método *hot deck* foram disponibilizadas nas bases de dados do Vigitel. As mulheres grávidas e as que não sabiam se estavam grávidas foram excluídas de nossa análise.

O segundo método de imputação utilizado e aplicado em dois modelos foi o PMM. Nos dois modelos, peso e altura foram as variáveis de desfecho. O primeiro modelo consistiu em replicar a imputação a partir do mesmo conjunto de covariáveis mencionadas na publicação oficial do Vigitel [5]: cidade, sexo, idade em anos, raça/cor e escolaridade. A publicação oficial do Vigitel não informa como eles lidaram com os valores faltantes dessas covariáveis. Em nossa análise, os valores faltantes dessas covariáveis foram considerados como uma categoria. O segundo modelo foi elaborado considerando a perspectiva de que, para a imputação de uma variável, as covariáveis não devem possuir valores faltantes e, portanto, foram selecionadas apenas com resposta para todos os indivíduos: cidade, sexo e idade em anos. O *pool* doador foi fixado contendo cinco candidatos. Um dos candidatos do *pool* foi aleatoriamente selecionado, e o seu valor foi atribuído à unidade com dado faltante. De acordo com Morris et al. [2], a partir do estudo de Heitjan & Little [13], no qual se utilizou *pool* doador $k = 5$, os pesquisadores passaram largamente a fixar o *pool* doador em $k > 1$. Em estudo de simulação com diferentes cenários, Kleinke [9] observou que a PMM produz melhores resultados quando se utiliza $k = 5$, exceto em amostras pequenas.

Após a imputação das variáveis peso e altura, o IMC foi calculado por meio da divisão do peso em quilogramas pela altura ao quadrado [14]. Quatro variáveis de IMC foram calculadas: (i) com as variáveis originais peso e altura; (ii) com as variáveis peso e altura imputadas por *hot deck*; (iii) com as variáveis peso e altura imputadas por PMM no primeiro modelo; e (iv) com as variáveis peso e altura imputadas por PMM no segundo modelo.

Em seguida, as médias e erros-padrão (EP) das variáveis peso, altura e IMC foram estimadas segundo método de imputação e ano de monitoramento do Vigitel. Nas estimativas, utilizou-se o módulo *survey* do Stata (https://www.stata.com), que considera os efeitos da amostragem. Nesse módulo, os EP são estimados de maneira linearizada [15]. Além disso, também foi calculado o percentual de dados imputados. As análises foram conduzidas no software Stata, versão 14.

Por fim, quanto aos procedimentos éticos, o Vigitel foi aprovado pela Comissão Nacional de Ética em Pesquisa para Seres Humanos do Ministério da Saúde [5]. No Vigitel, o consentimento livre e esclarecido foi obtido oralmente no momento do contato telefônico com os entrevistados. O presente estudo foi apreciado e aprovado pelo Comitê de Ética em Pesquisa da Faculdade de Saúde Pública da Universidade de São Paulo, sob Parecer nº 1.885.826, de 5 de janeiro de 2017.

Resultados

Na Tabela 1, são apresentadas as médias e EP das variáveis peso, altura e IMC, segundo método de imputação. É possível observar que os valores médios de peso, altura e IMC estimados com utilização da imputação por *hot deck* e por PMM são similares em todos os anos de monitoramento. A frequência de valores imputados para a variável peso variou de 2,76% em 2008 a 5,55% em 2015. A frequência de valores imputados para a variável altura variou de 5,74% em 2008 a 7,16% em 2007.

Tabela 1

Percentual de dados imputados, média e erro-padrão (EP) das variáveis peso, altura e índice de massa corporal (IMC) segundo os métodos de imputação e ano de monitoramento. *Vigilância de Fatores de Risco e Proteção para Doenças Crônicas por Inquérito Telefônico* (Vigitel), 2006 a 2017.

Ano	n	Original		*Hot deck*		PMM *		PMM **		Dados imputados (%)
		Média	EP	Média	EP	Média	EP	Média	EP	
Peso (kg)										
2006	53.861	68,61	0,41	68,50	0,40	68,56	0,41	68,56	0,40	2,99
2007	53.754	68,93	0,36	68,82	0,35	68,84	0,36	68,87	0,34	3,71
2008	53.778	69,42	0,40	69,33	0,39	69,34	0,40	69,36	0,41	2,76
2009	53.864	69,82	0,40	69,73	0,39	69,77	0,40	69,76	0,40	3,17
2010	53.848	70,34	0,33	70,24	0,31	70,26	0,32	70,27	0,31	3,87
2011	53.711	70,95	0,34	70,86	0,32	70,87	0,33	70,90	0,32	4,11
2012	45.089	71,65	0,37	71,59	0,38	71,59	0,37	71,57	0,38	3,89
2013	52.437	71,74	0,36	71,58	0,34	71,63	0,34	71,62	0,33	4,36
2014	40.940	71,81	0,28	71,70	0,26	71,72	0,26	71,76	0,27	4,36
2015	53.653	72,52	0,37	72,44	0,38	72,45	0,35	72,44	0,36	5,55
2016	52.945	72,55	0,33	72,45	0,32	72,49	0,31	72,52	0,32	3,81
2017	52.763	72,75	0,30	72,60	0,30	72,69	0,28	72,71	0,28	4,58
Altura (cm)										
2006	53.861	166,15	0,25	165,76	0,27	165,74	0,25	165,81	0,25	6,41
2007	53.754	166,34	0,24	165,88	0,26	165,86	0,25	165,96	0,25	7,16
2008	53.778	166,36	0,28	166,04	0,31	166,02	0,30	166,08	0,29	5,74
2009	53.864	166,43	0,27	166,04	0,30	166,04	0,29	166,12	0,29	6,69
2010	53.848	166,48	0,25	166,10	0,25	166,10	0,24	166,20	0,24	7,24
2011	53.711	166,53	0,25	166,21	0,28	166,17	0,26	166,28	0,26	6,48
2012	45.089	166,81	0,23	166,50	0,25	166,47	0,23	166,49	0,22	6,63
2013	52.437	166,82	0,24	166,46	0,24	166,46	0,24	166,52	0,23	6,65
2014	40.940	166,50	0,26	166,20	0,28	166,14	0,26	166,24	0,27	6,32
2015	53.653	166,69	0,26	166,34	0,26	166,33	0,26	166,42	0,25	6,99
2016	52.945	166,70	0,24	166,40	0,27	166,41	0,25	166,52	0,25	5,89
2017	52.763	166,94	0,22	166,63	0,23	166,63	0,22	166,65	0,22	5,85

(continua)

Tabela 1 (continuação)

Ano	n	Original		*Hot deck*		PMM *		PMM **		Dados imputados (%)
		Média	EP	Média	EP	Média	EP	Média	EP	
IMC (kg/m²)										
2006	53.861	24,95	0,10	24,91	0,09	24,92	0,10	24,90	0,10	8,37
2007	53.754	25,02	0,07	25,01	0,07	25,00	0,07	24,98	0,07	9,40
2008	53.778	25,15	0,07	25,13	0,06	25,12	0,07	25,11	0,08	7,60
2009	53.864	25,30	0,08	25,29	0,07	25,28	0,08	25,25	0,08	8,86
2010	53.848	25,48	0,08	25,45	0,07	25,44	0,07	25,41	0,08	9,93
2011	53.711	25,66	0,06	25,63	0,06	25,63	0,07	25,61	0,06	9,23
2012	45.089	25,84	0,08	25,82	0,08	25,81	0,08	25,79	0,09	9,39
2013	52.437	25,87	0,08	25,84	0,07	25,83	0,07	25,81	0,07	9,32
2014	40.940	25,99	0,10	25,97	0,08	25,98	0,08	25,96	0,09	8,92
2015	53.653	26,15	0,10	26,18	0,13	26,16	0,11	26,12	0,11	10,14
2016	52.945	26,16	0,07	26,16	0,07	26,15	0,07	26,13	0,07	8,17
2017	52.763	26,19	0,08	26,14	0,07	26,15	0,07	26,15	0,07	8,59

PMM: *Predictive Mean Matching*.

* Variáveis explicativas: sexo, idade, cidade, raça/cor e escolaridade;

** Variáveis explicativas: sexo, idade e cidade.

Discussão

Os resultados indicam que, após imputação, os valores da tendência central e EP das variáveis peso, altura e IMC são similares ao uso de dois métodos: *hot deck* e PMM. Nossos resultados sugerem que o PMM pode ser vantajosamente utilizado no Vigitel, dada a sua facilidade operacional, aplicação a maiores espectros em relação a tipos de *missing* e ampla disponibilidade nos principais pacotes estatísticos.

Na literatura, é descrita mais de uma maneira de conduzir o *hot deck* [6]. Porém, na metodologia descrita em publicação oficial [5], não está claro qual método *hot deck* foi utilizado no Vigitel. Outro aspecto que necessita de esclarecimentos pelo Vigitel é o manejo de covariáveis que originalmente possuem valores faltantes para a imputação, como escolaridade e raça/cor.

Conforme discutido por Andridge & Little [6], métodos como o PMM em relação ao *hot deck* têm como vantagem a possibilidade de inclusão de variáveis contínuas e também de maior número de variáveis nos modelos. Por outro lado, Morris et al. [2] reforçam a importância da especificação correta do modelo de imputação ao utilizar métodos como o PMM. Abayomi et al. [16], Marchenko & Eddings [12] e Bondarenko & Raghunathan [17] sugerem que seja conduzido diagnóstico das imputações com a utilização de gráficos de *kernel* e gráficos de dispersão com curva de *lowess*. Nós conduzimos as análises de diagnóstico para as imputações realizadas por PMM nos dois modelos para todos os anos de inquérito, e observamos que o método produziu imputações razoáveis (resultados disponíveis mediante comunicação com os autores).

Em análises adicionais, observou-se que, nos percentis extremos da distribuição, a densidade não apresenta a mesma uniformidade, o que impacta a estimativa da prevalência de obesidade. Por exemplo, em 2017, a prevalência de obesidade estimada com os dados imputados por *hot deck* foi de 19%. Com os dados imputados por PMM no primeiro modelo, foi de 18,7%. Com os dados imputados por PMM no segundo modelo, foi de 18,8%.

Ao comparar PMM com métodos baseados em regressão linear e distribuição normal, PMM imputa valores mais plausíveis e com maior probabilidade de serem reais. Isso ocorre porque, se a distribuição da variável for assimétrica, ou seja, quando a distribuição dos valores não está repartida exatamente em torno do ponto central, a PMM manterá essa distribuição entre os dados imputados. Além disso, a PMM utilizará apenas valores observados para realizar a imputação, mantendo os valores mínimos e máximos da distribuição. Portanto, o método não produz valores novos e não

observados [18]. Nosso estudo apresenta limitações que devem ser reportadas. Entre elas, destaca-se o número limitado de metodologias de imputação aplicadas na análise de dados, além da ausência de análise de sensibilidade, conforme proposto por Rezvan et al. [1]. Outro ponto é a possibilidade de os nossos resultados se mostrarem consistentes, devido ao baixo percentual de valores faltantes em todos os anos de monitoramento do Vigitel [9]. Para uma discussão mais ampla sobre o impacto da utilização da PMM em bases de dados com percentual elevado de valores faltantes, sugerimos a consulta aos estudos de Marshall et al. [11] e Kleinke [9].

A utilização do método de imputação PMM produziu estimativas das médias de peso, altura e IMC similares às estimativas produzidas com a utilização do método *hot deck* em 12 anos de inquéritos transversais. Em geral, não se recomenda realizar imputação de dados com covariáveis previamente imputadas ou com valores faltantes. Por isso, sugerimos a utilização do método PMM no modelo com as variáveis explicativas completas. O PMM é uma alternativa de imputação viável e amplamente disponível nos principais pacotes estatísticos usados na área da saúde. Finalmente, os resultados que obtivemos com os dados do Vigitel sugerem sua aplicabilidade ao conjunto dos inquéritos de saúde.

Colaboradores

I. K. S. Santos contribuiu com a concepção do estudo, revisão da literatura, análise e interpretação dos dados, redação, revisão crítica do texto e aprovação da versão final. W. L. Conde contribuiu com a concepção do estudo, análise e interpretação dos dados, revisão crítica do texto e aprovação da versão final.

Informações adicionais

ORCID: Iolanda Karla Santana dos Santos (0000-0003-3347-8396); Wolney Lisbôa Conde (0000-0003-0493-134X).

Referências

1. Rezvan PH, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. BMC Med Res Methodol 2015; 15:30.
2. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. BMC Med Res Methodol 2014; 14:75.
3. Miri HH, Hassanzadeh J, Rajaeefard A, Mirmohammadkhani M, Angali KA. Multiple imputation to correct for nonresponse bias: application in non-communicable disease risk factors survey. Glob J Health Sci 2016; 8:133-58.
4. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 2007; 16:219-42.
5. Ministério da Saúde. VIGITEL Brasil 2016. Vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico. Estimativas sobre frequência e distribuição sociodemográfica de fatores de risco e proteção para doenças crônicas nas capitais dos 26 estados brasileiros e no Distrito Federal em 2016. Brasília: Ministério da Saúde; 2017.
6. Andridge RR, Little RJA. A review of hot deck imputation for survey non-response. Int Stat Rev 2010; 78:40-64.
7. Pérez A, Dennis RJ, Gil JFA, Rondón MA, López A. Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia. Stat Med 2002; 21:3885-96.

8. Rubin DB. Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons; 1987. (Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics).
9. Kleinke K. Multiple imputation under violated distributional assumptions: a systematic evaluation of the assumed robustness of predictive mean matching. J Educ Behav Stat 2017; 42:371-404.
10. Allison PD. Missing data. In: Millsap RE, Maydeu-Olivares A, editors. The SAGE handbook of quantitative methods in psychology. London: SAGE Publications; 2009. p. 72-89.
11. Marshall A, Altman DG, Holder RL. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. BMC Med Res Methodol 2010; 10:112.
12. Marchenko YV, Eddings W. A note on how to perform multiple-imputation diagnostics in Stata. http://www.stata.com/users/ymarchenko/midiagnote.pdf (acessado em 12/Abr/2020).
13. Heitjan DF, Little RJA. Multiple imputation for the fatal accident reporting system. J R Stat Soc Ser C Appl Stat 1991; 40:13-29.
14. World Health Organization. Obesity: preventing and managing the global epidemic. Geneva: World Health Organization; 2000.
15. StataCorp. Stata survey data reference manual. Release 13. College Station: StataCorp; 2013.
16. Abayomi K, Gelman A, Levy M. Diagnostics for multivariate imputations. J R Stat Soc Ser C Appl Stat 2008; 57:273-91.
17. Bondarenko I, Raghunathan T. Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. Stat Med 2016; 35:3007-20.
18. Allison P. Imputation by predictive mean matching: promise & peril. https://statisticalhorizons.com/predictive-mean-matching (acessado em 12/Jan/2020).

Abstract

This study aimed to describe the estimated means for weight, height, and body mass index (BMI) according to two imputation methods, using data from Vigitel (Risk and Protective Factors Surveillance System for Chronic Non-Communicable Diseases Through Telephone Interview)*. This was a cross-sectional study that used secondary data from the Vigitel survey from 2006 to 2017. The two imputation methods used in the study were hot deck and Predictive Mean Matching (PMM). The weight and height variables imputed by hot deck were provided by Vigitel. Two models were conducted with PMM: (i) explanatory variables – city, sex, age in years, race/color, and schooling; (ii) explanatory variables – city, sex, and age in years. Weight and height were the outcome variables in the two models. PMM combines linear regression and random selection of the value for imputation. Linear prediction is used as a measure of distance between the missing value and the possible donors, thereby creating the virtual space with the candidate cases for yielding the value for imputation. One of the candidates from the pool is randomly selected, and its value is assigned to the missing unit. BMI was calculated by dividing weight in kilograms by height squared. The result shows the means and standard deviations for weight, height, and BMI according to imputation method and year. The estimates used the survey module from Stata, which considers the sampling effects. The mean values for weight, height, and BMI estimated by hot deck and PMM were similar. The results with the Vigitel data suggest the applicability of PMM to the set of health surveys.*

Nutrition Surveys; Surveillance; Nutritional Status; Epidemiology

Resumen

El objetivo de este estudio fue describir la estimación de medias de peso, altura e índice de masa corporal (IMC), según dos métodos de imputación, usando datos del Vigitel (Vigilancia de Factores de Riesgo y Protección para Enfermedades Crónicas No Transmisibles por Entrevista Telefónica)*. El diseño del estudio es transversal y se utilizaron datos secundarios de Vigitel, durante el período de 2006 a 2017. Los dos métodos para la imputación utilizados en el estudio fueron* hot deck *y* Predictive Mean Matching (*PMM). Las variables peso y altura imputadas por* hot deck *se recabaron de Vigitel. Se realizaron dos modelos con la utilización de la PMM: (i) variables explicativas -ciudad, sexo, edad en años, raza/color y escolaridad; (ii) variables explicativas -ciudad, sexo y edad en años. En los dos modelos, las variables peso y altura fueron las variables de desenlace. En la PMM, se combinan regresión lineal y selección aleatoria de valor para imputación. La predicción lineal es usada como medida de distancia entre el valor faltante y sus posibles donadores y, de este modo, se crea el espacio virtual con los casos candidatos de ceder su valor para la imputación. Uno de los candidatos del* pool *se selecciona aleatoriamente, y su valor es atribuido a la unidad faltante. El IMC se calculó mediante la división del peso en kilogramos por la altura al cuadrado. En los resultados, presentamos las medias y errores-patrón de peso, altura e IMC, según el método de imputación y año de seguimiento. En las estimaciones, se utilizó el módulo de encuesta del Stata, que considera los efectos de la muestra. Se observó que los valores medios de peso, altura e IMC estimados por* hot deck *y PMM son similares. Los resultados con los datos del Vigitel sugieren la aplicabilidad del PMM al conjunto de las investigaciones de salud.*

Encuestas Nutricionales; Vigilancia; Estado Nutricional; Epidemiología

Recebido em 28/Ago/2019
Versão final reapresentada em 14/Mai/2020
Aprovado em 17/Mai/2020