DOI http://dx.doi.org/10.20947/S0102-309820160008

Comparação de métodos para o tratamento das medidas antropométricas da POF 2008-2009

Mariana Vieira Martins de Matos*
Pedro Luis do Nascimento Silva**

A Pesquisa de Orçamentos Familiares 2008-2009, feita por amostragem em nível nacional, coletou informações antropométricas de peso e estatura dos indivíduos no Brasil. Numa pesquisa desse porte, o processo de coleta produz dados que estão sujeitos a contaminações por erros de medição e de não resposta. Tais erros podem afetar os cálculos de indicadores de prevalência de desnutrição, sobrepeso ou obesidade e impactar de forma distinta em diferentes segmentos populacionais. No presente artigo, comparou-se o desempenho do método Cidaq, que foi empregado na POF 2008-2009 para tratar os dados antropométricos, ao de outros dois métodos: os algoritmos de detecção de *outliers* TRC e Bacon, ambos associados ao algoritmo de imputação Poem. Essa comparação é fundamental para assegurar que o melhor método seja utilizado em pesquisas futuras, buscando assegurar a confiabilidade dos dados para os estudos que subsidiam o planejamento de políticas públicas nas áreas de saúde, nutrição, assistência social e outras. Os métodos foram comparados via simulação, considerando o impacto sobre as estimativas de média, desvio padrão e correlação entre peso e estatura. O método Cidaq apresentou uma pequena vantagem sobre os demais nos resultados da simulação paramétrica, enquanto para simulação não paramétrica destacou-se o método Bacon.

Palavras-chave: Crítica. Imputação. Medidas antropométricas. *Outliers*.

* Escola Nacional de Ciências Estatísticas (Ence), Instituto Brasileiro de Geografia e Estatística (IBGE), Rio de Janeiro-RJ, Brasil (marianavmaraujo@gmail.com).
** Escola Nacional de Ciências Estatísticas (Ence), Instituto Brasileiro de Geografia e Estatística (IBGE), Rio de Janeiro-RJ, Brasil (pedro-luis.silva@ibge.gov.br).

Introdução

Medidas antropométricas são relativas a traços físicos do corpo humano. O peso e a altura são duas medidas básicas consideradas no Relatório Técnico da Organização Mundial de Saúde sobre padrões das medidas antropométricas (WHO, 1995). A combinação dessas duas medidas compõe índices importantes para avaliação do estado nutricional dos indivíduos, sendo que a interpretação de uma está sempre associada à outra, ou à idade do indivíduo (WHO, 1995).

O Índice de Massa Corporal (IMC), calculado como a razão do peso e o quadrado da estatura de um indivíduo, é utilizado como base para definição de obesidade para adultos (WHO, 1995). Para crianças, nas quais as medidas do corpo estão em constante mudança, os padrões são definidos a partir das curvas de crescimento por idade e sexo, que são apresentadas no Relatório de Padrão de Crescimento das Crianças da Organização Mundial de Saúde (WHO, 2006).

Esses indicadores utilizados para avaliação do estado nutricional dos indivíduos são importantes no contexto do estudo da saúde das populações, com a finalidade de embasar políticas públicas de prevenção e tratamento de problemas. A epidemiologia é área que discute os problemas de saúde e sua distribuição na população por meio dos indicadores. Em artigo de 1992, republicado em 2012, Araújo já abordava o panorama diversificado do Brasil caracterizado pela incidência de doenças de fases distintas da teoria de transição demográfica e reconhecia a importância das informações para o planejamento da saúde:

> A informação epidemiológica é a base do planejamento de saúde. O processo decisório, a definição das prioridades, em um contexto tão complexo quanto o da saúde no Brasil, tem que se fundamentar em dados confiáveis e atualizados não só de mortalidade, mas também de morbidade (ARAÚJO, 2012, p. 537)

O sobrepeso e a obesidade, por exemplo, são tratados no Relatório Técnico Série 894 da OMS (2000) como um problema de saúde que pode ser prevenido, não sendo caracterizado como individual, mas sim relacionado ao modo de vida da população. Nesse sentido, a percepção desse problema na população é de extrema importância para direcionar as políticas de prevenção.

Batista e Rissin (2003) apresentam dados sobre a diversidade de cenários no Brasil no quesito transição nutricional, mas apontam o declínio da desnutrição e o aumento da prevalência de sobrepeso e obesidade como uma questão de abrangência nacional. Segundo os autores, os resultados dos estudos indicavam um comportamento epidêmico deste problema.

Esses diagnósticos e suas mudanças ao longo do tempo são de extrema importância para os cuidados com a saúde da população e dependem de pesquisas que disponham da coleta de dados antropométricos. A Pesquisa de Orçamentos Familiares (POF) 2008-2009, realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), é um levantamento amostral domiciliar em nível nacional, que contemplou a obtenção das medidas antropo-

métricas básicas (peso e estatura) dos indivíduos pesquisados, constituindo-se, assim, em uma possível fonte de dados para análises sobre o estado nutricional da população do país.

O processo de coleta das medidas antropométricas da POF 2008-2009 pela equipe de entrevistadores não foi simples, sobretudo pela necessidade do uso de equipamentos portáteis e exigência de procedimentos específicos no processo de medição para padronizar a coleta. Por exemplo, para a medição do comprimento das crianças menores de dois anos de idade, era necessário que a criança (ou bebê) fosse deitada e estendida de costas, sobre a superfície plana, dura e lisa de um antropômetro[1] (IBGE, 2010). Esse padrão de procedimentos de coleta, somado à carga elevada de informações a ser obtida em toda a pesquisa, em alguns casos, causou recusa do entrevistado.

Existem cuidados importantes na coleta para minimizar os problemas de omissão de dados e erros de resposta por meio do treinamento dos entrevistadores. Na Pesquisa Nacional de Demografia e Saúde da Criança e da Mulher – PNDS 2006, por exemplo, que também coletou as medidas antropométricas, foram realizados treinamento e estudos amostrais para aferir a qualidade das medidas obtidas pelos entrevistadores no treinamento. Além disso, houve acompanhamento da coleta e avaliação das medidas antropométricas durante todo o processo de pesquisa. Em casos de falta de coerência dos dados, foram feitas remensurações com a volta ao campo (ABEP et al., 2008).

O IBGE também realiza treinamento com a equipe de entrevistadores em todas as suas pesquisas, além de disponibilizar as especificações da coleta, como já mencionado. No entanto, mesmo adotando os devidos cuidados, em uma pesquisa de grande porte, como foi a POF 2008-2009 – em que o número esperado de domicílios entrevistados era de 59.548 (IBGE, 2010) –, a ocorrência de erros de mensuração e de omissão é inerente ao processo de coleta. Nesse tipo de pesquisa o retorno ao campo para conferências é um processo inviável. Como consequência desses eventuais erros, podem ocorrer vieses nas estimativas calculadas, tais como no percentual de subnutridos e/ou obesos, nas médias de estatura e de peso dos indivíduos, etc.

Depois de coletados os dados, uma forma de contornar esse tipo de problema é adotar um procedimento adequado e eficiente para crítica e imputação de dados, evitando que erros grosseiros possam distorcer os resultados.

As diferenças nas estimativas de média, devido ao impacto de erros e omissões no conjunto de dados, podem representar a variação de estatura e de peso de uma população no período de uma década. Além disso, os erros podem impactar na inclinação da reta que relaciona as variáveis de peso e estatura ou mesmo comprometer o cômputo tanto do percentual de desnutrição quanto o da obesidade, ambos baseados no IMC.

Segundo o Relatório de Antropometria e Estado Nutricional de Crianças, Adolescentes e Adultos no Brasil da POF 2008-2009, a prevalência de déficit de peso em adultos declinou continuamente desde o primeiro inquérito em 1974-1975, passando, nesse período,

[1] Para mais detalhes dos procedimentos de medição do peso e estatura na coleta da Pesquisa de Orçamentos Familiares, ver IBGE (2010, p. 28-29).

de 8,0% para 1,8%, no caso dos homens, e de 11,8% para 3,6%, no caso das mulheres. Ao mesmo tempo, as prevalências de excesso de peso e de obesidade aumentaram neste intervalo de tempo. A obesidade aumentou de 2,8% para 12,4%, em homens, e de 8,0% para 16,9%, em mulheres.

Levando em conta as variações nos padrões de peso e estatura da população ao longo dos anos e o fato de que essas variáveis compõem indicadores relevantes para o planejamento das políticas públicas de saúde para a população, é importante empregar um método para tratar os dados que tenham sido coletados sem a possibilidade de retorno ao campo para conferência.

Este artigo buscou avaliar o método crítica e imputação de dados quantitativos, abreviado por Cidaq, que foi utilizado para o tratamento dos dados antropométricos na POF 2008-2009, comparando-o com outros métodos disponíveis na literatura. O Cidaq foi desenvolvido por Silva (1989) com base na proposta de Little e Smith (1987) para lidar com problemas de respostas espúrias e/ou faltantes em conjuntos de dados multivariados com variáveis contínuas, considerando aspectos de estimação robusta e a geração de valores para imputação.

Foram estudados, também, o algoritmo TRC apresentado por Béguin e Hulliger (2003), o algoritmo Bacon, proposto por Billor et al. (2000) e adaptado por Béguin e Hulliger (2003), e o algoritmo Poem desenvolvido por Béguin e Hulliger (2003). Os dois primeiros são destinados apenas à etapa de detecção de *outliers* e o terceiro apenas à etapa de imputação dos dados (após alguma crítica já aplicada).

Dados e métodos de análise

Os métodos de tratamento de dados contaminados por erros não amostrais ou incompletos foram avaliados comparativamente por meio de simulação com base na distribuição de um subconjunto dos dados da POF 2008-2009. Nessa avaliação, considerou-se o tratamento completo dos dados (as etapas de crítica e imputação) e para tal, os algoritmos de detecção de *outliers*, TRC e Bacon, foram associados ao algoritmo de imputação Poem.

Os dados utilizados para a análise são as medidas antropométricas de peso e estatura coletadas pela POF 2008-2009. Na pesquisa, o método para tratamento dos dados (Cidaq) foi aplicado para grupos de pessoas definidos segundo a idade e o sexo, de acordo com o padrão de crescimento da vida de um indivíduo. Considerou-se também o rendimento familiar *per capita*, além das variáveis peso e estatura.

As variáveis empregadas nesse estudo foram as medidas antropométricas (peso e estatura) e as características individuais de idade e sexo para a definição dos grupos (utilizou-se a mesma divisão considerada na POF 2008-2009) de aplicação dos métodos, mas não foi considerada a variável de rendimento.

A composição dos grupos de idade foi a seguinte: para as crianças de menos de 5 anos de idade, para as quais o processo de crescimento varia de forma rápida, foram definidos

grupos de mês em mês. Para as crianças de 5 até menos de 10 anos foram considerados grupos em idades de seis em seis meses. A população de 10 a menos de 20 anos foi dividida em subgrupos de um ano de idade. Os indivíduos de 20 a menos de 75 anos foram agrupados em intervalos quinquenais. E um último grupo incluiu a população de 75 anos ou mais. Para cada faixa etária foram analisados separadamente indivíduos do sexo feminino e masculino.

Um mecanismo preciso para avaliação de métodos para crítica e imputação de dados é comparar os dados brutos (obtidos da coleta) com aqueles validados (verificação direta, voltando-se ao campo). Esse procedimento é dispendioso e, muitas vezes, inviável quando se trata de uma pesquisa de grande porte como a POF 2008-2009.

Já que a validação dos dados diretamente do sistema real de coleta não é uma prática viável neste caso, a simulação foi utilizada neste trabalho para avaliar os métodos de tratamento de dados em discussão. A simulação é um instrumento que permite a replicação de um processo para estudo e avaliação, quando não se pode intervir no sistema real. Nesse contexto, a simulação possibilitou a geração de dados em um número de 3.000 réplicas para que os métodos de tratamento de dados fossem avaliados. Dois tipos de simulação foram considerados: a não paramétrica, pela qual as réplicas de dados foram geradas diretamente (replicação de 3.000 cópias idênticas ao conjunto de dados considerado); e a paramétrica, pela qual as réplicas foram geradas a partir de um modelo normal bivariado ajustado ao conjunto de dados considerado.

Embora o modelo normal seja adequado para representar a distribuição das variáveis peso e estatura para a faixa etária considerada, avaliar os métodos unicamente sob a ótica da simulação paramétrica baseada em um modelo normal poderia mostrar vantagem para algum dos métodos que trazem em sua formulação o pressuposto de aplicação a dados normais. Já a simulação por replicação direta representa os dados em sua distribuição real da população pesquisada; assim, os dois tipos de simulação são complementares para avaliar os métodos.

O conjunto de dados utilizado para a simulação correspondeu ao subgrupo de dados brutos da POF 2008-2009 referentes a homens com idade entre 19 e 20 anos. A escolha deste grupo levou em conta o tamanho da amostra, já que os grupos de crianças, devido ao intervalo de idade considerado, possuíam tamanhos de amostra pequenos demais para aplicação com segurança. Além disso, a faixa etária de 19 a 20 anos é limítrofe no quesito de mudança das medidas antropométricas. Nesta fase a estatura dos indivíduos começa a se estabilizar, enquanto o peso ainda é bastante variável e, dessa forma, os diagnósticos de obesidade passam a não ter como aliado as mudanças na estatura.

Das réplicas calcularam-se algumas estatísticas de interesse, a média, o desvio padrão e a correlação das variáveis em três momentos: antes de serem introduzidos erros e omissões; após a inclusão dessas contaminações; e depois da aplicação dos métodos para tratamento. Isso permitiu comparar o desempenho de cada um dos métodos quanto ao tratamento dos dados.

Às variáveis de estatura e peso do banco de dados simulados foram introduzidas omissões e contaminações (possíveis erros não amostrais) conhecidas, cujas características basearam-se no padrão observado no conjunto de dados original (dado bruto, também disponibilizado na base da POF 2008-2009), conforme descrição a seguir. Foram utilizadas as quatro variáveis disponibilizadas pelo IBGE na base da POF 2008-2009: peso original; estatura original; peso após a crítica e imputação de dados; e estatura após a crítica e imputação de dados.

Sob a hipótese de que os dados tratados fossem os corretos, um banco de fatores multiplicativos foi criado a partir de toda a amostra da POF 2008-2009, dividindo-se o valor original (dado bruto da amostra) pelo valor corrigido (dado após o tratamento feito pelo IBGE), para aqueles casos "corrigidos" pela pesquisa. As contaminações foram introduzidas nas réplicas simuladas por meio desses fatores multiplicativos, selecionados aleatoriamente do banco criado, que alteraram ("contaminaram") alguns dos valores simulados.

Importante explicitar que os fatores multiplicativos utilizados para introduzir erros nos dados simulados foram calculados a partir de toda a amostra da POF 2008-2009, ou seja, incluindo todos os grupos de idade, e foram aplicados ao subconjunto de dados selecionado para simulação (homens de 19 a 20 anos) para evitar viés vantajoso na avaliação comparativa dos métodos, já que a POF empregou o método Cidaq para tratar os dados. Além disso, como observado no início dessa seção, para a aplicação do Cidaq na POF 2008-2009, foi considerada também a variável rendimento familiar *per capita*, enquanto neste artigo foram incluídas apenas as variáveis antropométricas peso e estatura.

Para replicar as omissões, alguns dados das réplicas foram suprimidos de forma determinística, reproduzindo a quantidade observada nos dados originais (do conjunto selecionado – homens com 19 a 20 anos), considerando as omissões apenas na variável estatura, apenas na variável peso e em ambas.

As omissões foram feitas em mesma quantidade para todas as réplicas simuladas, enquanto as contaminações foram introduzidas em três níveis distintos (baixo, médio e alto) para cada grupo de 1.000 réplicas. Os resultados apresentados referem-se aos dados simulados com alto nível de contaminação, que correspondem a aproximadamente 7% do total.

Os métodos de crítica e imputação foram aplicados aos dados em duas condições, transformados previamente (utilizando a transformação Box-Cox) e em sua escala original (não transformados). O Cidaq, que contempla os passos de crítica e imputação, foi aplicado isoladamente a cada réplica do conjunto de dados, enquanto os métodos Bacon e TRC foram associados ao algoritmo Poem para imputação em uma etapa subsequente à detecção de *outliers*.

Avaliaram-se a eficiência dos métodos na detecção de *outliers* e a influência da imputação nas estimativas de média, desvio padrão e correlação. O primeiro quesito foi avaliado com base no percentual de detecções acertadas e falhas de cada método e, o segundo,

por meio da comparação das estimativas antes da contaminação, após a contaminação e depois da imputação. As seguintes medidas foram utilizadas:

$$Impacto\ da\ contaminação = Y_i^k - N_i^k \quad (1)$$

Onde: Y_i^k representa a estimativa k da réplica i após a contaminação; e N_i^k refere-se ao valor da estimativa k da réplica i antes da contaminação, com k podendo ser estimativa de média, desvio padrão ou correlação.

$$Impacto\ da\ imputação = I_i^{km} - N_i^k \quad (2)$$

Onde: I_i^{km} representa a estimativa k da réplica i após o tratamento de dados pelo método m.

O impacto da imputação é a diferença entre a estimativa após a imputação e antes da contaminação e, portanto, valores mais próximos de zero indicam melhor correção do viés causado pela contaminação. As medidas que compõem as tabelas apresentadas na seção de resultados referem-se aos impactos médios, ou seja, considerou-se a média dos impactos para cada grupo de 1.000 réplicas.

Outra medida utilizada para avaliação dessas diferenças foi a raiz quadrada do erro quadrático médio relativa (REQMR), que resume aspectos de variabilidade relacionando-a com a média, podendo ser interpretada como o percentual que o desvio padrão de uma estimativa representa em relação à sua média. Matematicamente a REQMR é definida como:

$$REQMR = \frac{\sqrt{\frac{1}{R}\sum_{i=1}^{R}(Y_i^k - N_i^k)^2}}{\bar{N}^k} \quad (3)$$

Onde: R representa o número de réplicas simuladas; N_i^k é o valor esperado para cada réplica i; e $\bar{N}^k$ corresponde à média das estimativas das R réplicas. Uma medida equivalente foi calculada também com as estimativas obtidas após a detecção e imputação (I_i^{km}).

Essas medidas permitiram avaliar o nível de variação dos impactos da contaminação e da imputação entre as réplicas em relação aos impactos médios, verificando aspectos de estabilidade dos resultados apresentados pelos métodos empregados para o tratamento dos dados.

Os métodos de tratamento comparados

O Cidaq (SILVA, 1989) combina técnicas de três diferentes áreas da estatística: estimação robusta, detecção de *outliers* e análise e inferência estatística com dados ausentes, todas aplicadas a problemas multivariados. O método é composto por seis passos: organização e transformação dos dados para aproximar a normalidade multivariada; estimação robusta do vetor de médias e da matriz de covariâncias; identificação dos casos (questionários) com problemas; identificação e descarte de valores suspeitos em cada caso com problemas; imputação dos dados faltantes ou descartados; e transformação inversa dos valores imputados para escala usual.

A identificação dos casos suspeitos é baseada na distância de *Mahalanobis* (equação 4), que mede o afastamento das observações em relação à média dos dados, levando em conta a estrutura de covariância dos dados.

$$D_i^2 = \left[X_{ip_i} - \hat{\mu}_{p_i}\right]^T . \left[\hat{V}_{p_i p_i}\right]^{-1} . \left[X_{ip_i} - \hat{\mu}_{p_i}\right] \quad (4)$$

Onde: X_{ip_i} é o vetor dos valores presentes da observação *i*; $\hat{\mu}_{p_i}$ e $\hat{V}_{p_i p_i}$ são, respectivamente, o vetor das médias e a matrix de variâncias e covariâncias, todas calculadas a partir dos valores presentes.

Sob o modelo normal multivariado proposto para descrever os dados no Cidaq, utiliza-se o algoritmo "ER" para a estimação dos parâmetros: o vetor de médias (μ) e a matriz de variâncias e covariâncias (V). Esse algoritmo é aplicável a conjuntos de dados com valores faltantes para produzir estimativas robustas de média e variância. No Cidaq as estimativas resultantes do algoritmo "ER" são utilizadas na equação (4) para detecção de *outliers*.

Os valores descartados na etapa de detecção e os valores faltantes são imputados em etapa seguinte por meio da regressão sobre os valores presentes. Os parâmetros desse modelo são extraídos de nova aplicação do algoritmo "ER". Caso tenha sido usada a transformação dos dados inicialmente para garantir a hipótese de normalidade dos dados, o procedimento de crítica e imputação Cidaq é finalizado com a transformação inversa dos dados imputados para a escala usual.

O algoritmo TRC (BÉGUIN; HULLIGER, 2003), traduzido como correlação de postos transformada, também identifica os casos como *outliers* a partir da distância de *Mahalanobis*, com uma modificação do que foi apresentado na equação (4). Enquanto o Cidaq busca um método para obter estimativas de média e variância robustas, o método TRC propõe o uso de estimativas alternativas para as medidas de locação e dispersão.

Nessa proposta, a correlação entre as variáveis é estimada pela correlação de postos de *Spearman* (equação 5), que é utilizada, juntamente com o desvio absoluto da mediana, para obter uma estimativa robusta da matriz de covariância dos dados, a saber:

$$\tilde{\rho}_{jk} = 2sen\left(\frac{\pi}{6} R\ (x^j\, x^k)\right),$$
$$com\ R(x^j\, x^k) = 1 - 6\ \frac{\sum_{i=1}^{n} \delta_i^{\,2}}{n\,(n^2-1)} \quad (5)$$

Onde: $r(x_{ij})$ é o posto da observação *i* para a variável *j* na amostra, $\delta_i = r(x_{ij}) - r(x_{ik})$; e *n* é o número de observações da amostra.

O método inclui uma conversão da matriz de covariâncias, utilizando suas componentes principais para garantir uma matriz de covariâncias definida positiva e que medidas de locação e dispersão, formuladas a partir dela, sejam robustas. Os estimadores TRC são finalmente obtidos por uma transformação dessas medidas robustas de volta à base original. Esse método foi adaptado para lidar com valores faltantes e incluir a informação do desenho amostral no processo de detecção de *outliers*.

O algoritmo Bacon (BILLOR et al., 2001) é um método de detecção iterativo de "busca para frente". Seu ponto de partida é a identificação de um subconjunto inicial dos dados, suposto livre de observações *outliers*, o qual vai aumentando gradualmente a cada passo do algoritmo com a inclusão de pontos também declarados não *outliers*, de modo que fiquem excluídos desse conjunto apenas os pontos a serem descartados. Esse algoritmo também utiliza a distância de *Mahalanobis* com medidas de dispersão e locação robustas na detecção de *outliers*.

Para compor o subconjunto inicial de dados, suposto livre de *outliers*, selecionam-se as observações com menores valores de distância em relação à mediana, definida na equação (6):

$$d_i^{med} = \left[\sum_k (X_{ik} - med_k)^2 \right]^{1/2} \quad (6)$$

Onde: med_k é a mediana referente à *k-ésima* variável.

As observações selecionadas nesse primeiro passo são utilizadas na segunda etapa para estimar os parâmetros, média e variância, para o ajuste do modelo aos dados, a partir do qual se inicia a busca por *outliers*. Sob a suposição de que o subconjunto de dados inicialmente selecionado é livre de *outliers*, as estimativas calculadas a partir dele são robustas. Esse algoritmo foi adaptado para lidar com dados faltantes, utilizando de forma integrada o algoritmo "EM", e também para incluir a informação do desenho amostral da pesquisa da qual os dados são oriundos (BÉGUIN; HULLIGER, 2003). O algoritmo de imputação Poem foi associado aos algoritmos TRC e Bacon para repor os valores faltantes e substituir os valores rejeitados por esses sistemas de crítica. O Poem (BÉGUIN; HULLIGER, 2003) é um mecanismo automático de imputação de dados baseado no método do vizinho mais próximo e permite levar em conta regras de edição e a informação do peso amostral. A métrica utilizada para selecionar o vizinho mais próximo é a distância de *Mahalanobis*, sendo sua aplicação condicionada à hipótese de que a massa de dados seja aproximadamente elíptica.

Resultados

Simulação paramétrica

Na simulação paramétrica as réplicas dos dados "limpos" foram construídas gerando observações de uma distribuição normal bivariada, tendo como valores dos parâmetros as estimativas obtidas da POF 2008-2009 para o subgrupo de homens com idade entre 19 e 20 anos. Estes dados "limpos" foram então contaminados conforme descrito anteriormente.

Na Tabela 1, que traz os resultados para as detecções de *outliers*, observa-se que o maior percentual de detecções corretas em relação ao total é obtido com a aplicação do Cidaq, que apresentou em média 97,8% de acertos na condição em que não há transformação prévia dos dados e 93,7% quando os dados são transformados previamente.

O desempenho do algoritmo TRC é o pior entre os métodos avaliados, com percentuais de detecções corretas de 68,1% e 63,4%, respectivamente, quando não há e quando há transformação prévia dos dados.

Os resultados para as diferenças das estimativas antes e após o tratamento dos dados pelos métodos descritos (Tabela 2) mostram que o impacto da alta contaminação na média das variáveis antropométricas causou redução média de 2,07 centímetros na estatura e aumento médio de 1,69 quilograma no peso. Esse viés médio causado pela alta contaminação pode representar a variação da média de estatura e da média de peso de uma população no período de uma década, fenômeno importante, o qual se tem interesse em identificar. Nas estimativas de desvio padrão o impacto é de mais de dez unidades na estatura e aproximadamente 14 no peso. E na correlação a diferença média, antes e após a contaminação dos dados, é de 0,53, levando à severa subestimação se a contaminação não for detectada e tratada.

TABELA 1
Quantidade média de observações detectadas como suspeitas nos dados de peso e estatura simulados (1) a partir da POF 2008-2009, com nível alto de contaminação (2), segundo os métodos aplicados

Transformação nos dados	Métodos	Número médio de detecções			Percentual de detecções	
		Total	Acertadas	Falsas	Acertadas	Falsas
Não	TRC	166,5	113,5	53,0	68,1	31,9
	Bacon	131,9	110,6	21,2	83,9	16,1
	Cidaq	107,0	104,7	2,4	97,8	2,2
Sim	TRC	180,6	114,5	66,1	63,4	36,6
	Bacon	143,5	112,6	30,9	78,4	21,6
	Cidaq	114,7	107,5	7,3	93,7	6,3

Fonte: Dados simulados a partir da Pesquisa de Orçamentos Familiares 2008-2009, IBGE.
(1) Os dados simulados referem-se a homens com idade entre 19 e 20 anos.
(2) Nível em que 7% das observações simuladas tiveram seus valores alterados por um multiplicador, indicando contaminação (120 observações).

TABELA 2
Diferenças médias das estimativas das variáveis antropométricas (1) antes e após a contaminação (2) e depois do tratamento dos dados com os métodos TRC, Bacon e Cidaq

Estimativa	Variável	Impacto médio da contaminação	Impacto médio da imputação					
			Sem transformação			Com transformação		
			TRC + Poem	Bacon + Poem	Cidaq	TRC + Poem	Bacon + Poem	Cidaq
Média	Estatura	-2,07	-0,41	-0,43	-0,01	-0,43	-0,45	-0,03
	Peso	1,69	0,01	0,00	-0,17	0,20	0,15	0,09
Desvio padrão	Estatura	10,61	0,34	0,65	-0,04	0,36	0,66	-0,03
	Peso	13,75	-0,01	0,30	0,11	-0,12	0,23	-0,06
Correlação	Est. x peso	-0,53	-0,07	-0,07	-0,01	-0,09	-0,08	-0,01

Fonte: Dados simulados a partir da Pesquisa de Orçamentos Familiares 2008-2009, IBGE.
(1) Variáveis peso e estatura referentes a dados de homens com idade entre 19 e 20 anos.
(2) Nível alto de contaminação (7% das observações simuladas tiveram seus valores alterados por um multiplicador).

Os resultados de REQMR, apresentados na Tabela 3, indicam que a variabilidade das diferenças de estimativa de desvio padrão devido à contaminação representa 138,90% do valor de sua média para a variável estatura e 121,07% de sua média para o peso, ou seja, a variabilidade das distribuições dos impactos da contaminação ultrapassa o valor de suas médias para as estimativas de desvio padrão e correlação. Após o tratamento dos dados, os percentuais de variabilidade das diferenças calculadas se reduzem bastante para quaisquer dos métodos aplicados. O maior percentual de variabilidade ainda observado após o tratamento dos dados refere-se à estimativa de correlação após o tratamento dos dados pelo TRC com transformação prévia, mas não ultrapassa 18%.

TABELA 3
Raiz quadrada do erro quadrático médio relativa das diferenças médias das estimativas das variáveis antropométricas (1) antes e após a contaminação (2) e depois do tratamento dos dados com os métodos TRC, Bacon e Cidaq

Estimativa	Variável	REQMR do impacto da contaminação	REQMR do impacto da imputação (%)					
			Sem transformação			Com transformação		
			TRC + Poem	Bacon + Poem	Cidaq	TRC + Poem	Bacon + Poem	Cidaq
Média	Estatura	1,21	0,25	0,26	0,07	0,26	0,27	0,07
	Peso	2,59	0,29	0,31	0,36	0,42	0,38	0,29
Desvio padrão	Estatura	138,90	4,68	8,59	1,31	4,95	8,78	1,34
	Peso	121,07	1,30	2,87	1,55	1,70	2,45	1,42
Correlação	Est. x peso	101,90	13,23	14,68	3,04	17,54	15,80	3,52

Fonte: Dados simulados a partir da Pesquisa de Orçamentos Familiares 2008-2009, IBGE.
(1) Variáveis peso e estatura referentes a dados de homens com idade entre 19 e 20 anos.
(2) Nível alto de contaminação (7% das observações simuladas tiveram seus valores alterados por um multiplicador).

O Cidaq apresenta melhores resultados do que os demais métodos tanto na detecção de *outliers* quanto na correção do viés causado pela contaminação nas estimativas calculadas.

Simulação não paramétrica

Nesse cenário de simulação, as réplicas dos dados foram construídas com base na distribuição empírica dos dados da POF 2008-2009. Foram utilizados os dados antropométricos já tratados, também disponibilizados por essa pesquisa, do mesmo grupo já mencionado (homens com idade entre 19 e 20 anos). Nesse caso em particular, não foi feita reamostragem dos dados de origem (o subgrupo selecionado), apenas sua replicação, à qual foram introduzidas as omissões e contaminações. Sob esse tipo de simulação, pode-se dizer que os dados são mais condizentes com a realidade, já que preservam aspectos distribucionais não contemplados por um modelo estatístico. Os resultados para essa simulação são apresentados nas Tabelas 4, 5 e 6.

Os resultados para as detecções de *outliers* (Tabela 4) mostram que o percentual de acertos em relação ao total detectado diminui para todos os métodos aplicados para a simulação não paramétrica se comparado à simulação paramétrica, sendo o Cidaq com a transformação prévia nos dados o método de melhor desempenho quanto ao percentual de acertos (79,4%). Embora consiga detectar, em média, apenas 112 dos 120 casos

contaminados, esse método destaca-se por apresentar menor percentual de falsas detecções e, portanto, tem menor risco de classificar um caso compatível com a massa de dados como um caso *outlier*. Outro aspecto importante é que, nesse tipo de simulação, os métodos apresentam melhor resultado com a transformação prévia dos dados, ao contrário da simulação paramétrica.

Para essa aplicação, os dados contaminados também revelaram como os impactos da contaminação podem ser prejudiciais à estimação. Na Tabela 5, que traz os resultados para as diferenças das estimativas antes e após o tratamento dos dados pelos métodos descritos, observa-se que, em média, a estatura média é subestimada em 1,71 centímetro e o peso médio é superestimado em 1,87 quilograma. Também se nota que o desvio padrão, após a contaminação dos dados simulados, aumenta 10,68 centímetros para a estatura, em média, e 12,14 quilogramas para o peso, em média.

TABELA 4
Quantidade média de observações detectadas como suspeitas nos dados de peso e estatura simulados (1) a partir da POF 2008-2009, com nível alto de contaminação (2), segundo os métodos aplicados

Transformação nos dados	Métodos	Número médio de detecções			Percentual de detecções	
		Total	Acertadas	Falsas	Acertadas	Falsas
Não	TRC	232,4	116,4	116,1	50,1	49,9
	Bacon	191,2	114,2	77,0	59,7	40,3
	Cidaq	144,5	109,7	34,8	75,9	24,1
Sim	TRC	214,9	117,2	97,7	54,5	45,5
	Bacon	172,9	116,1	56,9	67,1	32,9
	Cidaq	141,3	112,2	29,1	79,4	20,6

Fonte: Dados simulados a partir da Pesquisa de Orçamentos Familiares 2008-2009, IBGE.
(1) Os dados simulados referem-se a homens com idade entre 19 e 20 anos.
(2) Nível em que 7% das observações simuladas tiveram seus valores alterados por um multiplicador, indicando contaminação (120 observações).

TABELA 5
Diferenças médias das estimativas das variáveis antropométricas (1) antes e após a contaminação (2) e depois do tratamento dos dados com os métodos TRC, Bacon e Cidaq

Estimativa	Variável	Impacto médio da contaminação	Impacto médio da imputação					
			Sem transformação			Com transformação		
			TRC + Poem	Bacon + Poem	Cidaq	TRC + Poem	Bacon + Poem	Cidaq
Média	Estatura	-1,71	-0,08	-0,09	0,30	-0,05	0,03	0,29
	Peso	1,87	-0,48	-0,44	-0,73	-0,20	0,06	-0,46
Desvio padrão	Estatura	10,68	0,08	0,31	-0,22	0,06	0,19	-0,24
	Peso	12,14	-1,54	-0,97	-1,45	-1,23	-0,61	-1,27
Correlação	Est. x peso	-0,53	0,00	0,00	0,05	-0,04	-0,04	0,05

Fonte: Dados simulados a partir da Pesquisa de Orçamentos Familiares 2008-2009, IBGE.
(1) Variáveis peso e estatura referentes a dados de homens com idade entre 19 e 20 anos.
(2) Nível alto de contaminação (7% das observações simuladas tiveram seus valores alterados por um multiplicador).

Após a imputação, todas as estimativas ficam bem próximas das estimativas dos dados antes de serem contaminados, para todos os métodos aplicados, assim como no cenário

de simulação paramétrica. Nesse cenário de simulação destaca-se o método Bacon com a transformação prévia dos dados, do qual resultam as menores diferenças em todas as estimativas calculadas, exceto para a de desvio padrão da estatura (oitava coluna da Tabela 5). As maiores diferenças após a imputação (impacto da imputação) referem-se à estimativa de desvio padrão do peso, enquanto as menores correspondem à estimativa de correlação entre as variáveis.

Observa-se, pela Tabela 6 de resultados de REQMR, que a variabilidade relativa das diferenças das estimativas após a contaminação dos dados extrapola 100% para as estimativas de desvio padrão e correlação, assim como mostram os resultados na simulação paramétrica. Com o tratamento dos dados, a variabilidade para as diferenças de desvio padrão do peso fica 10% superior aos seus valores médios para os métodos TRC e Cidaq nas duas condições de aplicação (dados transformados e não transformados) e 10% inferior para o método Bacon, nos dois casos. Destacam-se também os diferentes efeitos de variabilidade gerados na estimativa de correlação após a aplicação dos diferentes métodos para tratar os dados (sem a transformação prévia dos dados). O resultado apresentado após a aplicação do Cidaq sem transformação prévia nos dados produz uma variabilidade relativa pouco superior a 10%, que é a maior entre os métodos, enquanto a aplicação do TRC produz uma variabilidade relativa aproximadamente igual a 1,45%. Com a transformação dos dados os resultados são mais homogêneos entre os métodos, que apresentam percentuais inferiores a 10%.

Para esse cenário de simulação (não paramétrica) o método Bacon apresenta melhores resultados na estimação do que os demais métodos. Já na detecção de *outliers* destaca-se o método Cidaq.

TABELA 6
Raiz quadrada do erro quadrático médio relativa das diferenças médias das estimativas das variáveis antropométricas (1) antes e após a contaminação (2) e depois do tratamento dos dados com os métodos TRC, Bacon e Cidaq

Estimativa	Variável	REQMR do impacto da contaminação	REQMR do impacto da imputação (alta contaminação)					
			Sem transformação			Com transformação		
			TRC + Poem	Bacon + Poem	Cidaq	TRC + Poem	Bacon + Poem	Cidaq
Média	Estatura	1,01	0,07	0,07	0,18	0,05	0,06	0,17
	Peso	2,84	0,74	0,67	1,09	0,41	0,21	0,71
Desvio padrão	Estatura	141,49	1,23	4,23	3,06	1,44	2,70	3,25
	Peso	107,28	13,22	8,33	12,49	10,60	5,28	10,89
Correlação	Est. x peso	104,40	1,45	2,01	10,48	9,23	8,76	9,67

Fonte: Dados simulados a partir da Pesquisa de Orçamentos Familiares 2008-2009, IBGE.
(1) Variáveis peso e estatura referentes a dados de homens com idade entre 19 e 20 anos.
(2) Nível alto de contaminação (7% das observações simuladas tiveram seus valores alterados por um multiplicador).

Ilustração do efeito de não se utilizar métodos para tratar os dados

Para ilustrar o efeito da aplicação dos métodos de crítica e imputação na estimação da média, desvio padrão e correlação do conjunto de dados, os métodos descritos foram empregados em dois grupos (segundo sexo e idade) selecionados da POF 2008-2009, considerando os dados brutos (não tratados) disponibilizados na base de dados da pesquisa. Nessa aplicação não há como verificar a eficácia do efeito do tratamento de cada método aplicado, pois, como já mencionado, validar esses dados é um processo custoso. O intuito é mostrar a importância do tratamento dos dados na estimação, independentemente do método usado, observando como esses métodos "alteram" (ou melhoram) as estimativas de interesse.

É claro que o nível de mudança nas estimativas finais para determinado grupo de indivíduos é influenciado pelo nível de possíveis erros que existem no conjunto de dados. Como verificado no estudo, o impacto do tratamento nas estimativas é maior quanto maior for o nível de contaminação existente no conjunto de dados.

Os resultados apresentados referem-se aos grupos de meninos com idade entre 7 e 7anos e meio e de meninas com idade entre 24 e 25 meses, excluindo-se o limite superior das idades. A escolha dos grupos foi motivada pela diferença nos níveis de impacto do tratamento de dados nas estimativas calculadas. Quaisquer outros grupos poderiam ter sido escolhidos. No grupo de meninas é possível observar as maiores diferenças entre os valores estimados antes e após o tratamento dos dados para a variável estatura e o contrário ocorre para a variável peso (as maiores diferenças acontecem no grupo de meninos). Essa diferença depende da quantidade de dados omissos e com erros de informação, que, por sua vez, é influenciada pela faixa etária.

TABELA 7
Estimativas da estatura (cm) e do peso (kg) para o grupo masculino de 7 a 7,5 anos de idade, antes e após o tratamento dos dados com os métodos TRC, Bacon e Cidaq

Variável	Estimativas	Tratamento						
		Ausente	Sem transformação			Com transformação		
			TRC+ Poem	Bacon+ Poem	Cidaq	TRC+ Poem	Bacon+ Poem	Cidaq
Estatura	1º decil	112,28	114,07	114,36	112,60	112,28	114,00	113,81
	9º decil	132,80	132,00	132,00	132,80	132,80	132,78	132,30
	Mediana	123,40	123,30	123,50	123,30	123,40	123,60	123,50
	Média	122,91	123,04	123,18	123,11	122,90	123,41	123,26
	Desvio padrão	7,49	6,72	6,62	7,31	7,43	6,77	6,93
Peso	1º decil	20,00	19,70	20,00	19,60	20,00	20,00	20,00
	9º decil	32,40	31,80	31,90	31,20	32,10	32,00	32,07
	Mediana	24,70	24,70	24,60	24,30	24,60	24,74	24,92
	Média	25,68	25,24	25,30	24,97	25,62	25,66	25,69
	Desvio padrão	5,13	4,35	4,51	4,50	4,99	4,94	4,96
Estatura x peso	Correlação	0,43	0,56	0,56	0,53	0,45	0,56	0,51

Fonte: Pesquisa de Orçamentos Familiares 2008-2009, IBGE.
Nota: Estimativas calculadas considerando o plano amostral e a pós-estratificação.

O grupo masculino era composto por 881 indivíduos na amostra, sendo que a maioria das observações possuía informação completa das medidas antropométricas (88,8%), enquanto 7,4% estavam sem nenhuma informação e os 3,9% restantes tinham informação parcial.

Na Tabela 7 observa-se, quanto à variável estatura, que para a medida do 1º decil há aumento da estimativa após o tratamento por todos os métodos, com exceção do TRC com transformação, chegando a ser 2,08 centímetros maior após o tratamento pelo método Cidaq sem transformação. A estimativa do 9º decil diminui em até 0,80 centímetro, considerando os tratamentos pelos métodos TRC e Bacon sem transformação prévia. Para a variável peso, a medida de 9º decil também diminui após o tratamento de dados, em até 1,20 quilograma, considerando o método Cidaq sem transformação prévia dos dados.

As estimativas de média e mediana, que são medidas centrais da distribuição, são menos impactadas pelos tratamentos, até porque esses buscam preservar a estrutura da distribuição dos dados, procurando focar os valores atípicos, destoantes da massa de dados considerada. Na variável estatura a estimativa aumenta em 0,5 centímetro considerando o tratamento pelo método Bacon com transformação prévia e na variável peso há redução da estimativa em até 0,71 quilograma pelo tratamento com Cidaq sem transformação prévia dos dados.

Esses valores podem parecer pequenos para análises univariadas, mas note-se que os vícios se manifestam em direções opostas nas variáveis peso e estatura. Assim, o efeito combinado dos erros pode ser maior em indicadores que dependam de combinar as duas medidas. Por exemplo, o valor do índice de massa corporal (IMC), calculado a partir das estimativas de média das variáveis antropométricas, seria bastante semelhante antes e após o tratamento dos dados por quaisquer dos métodos, com diferença máxima de 0,52 unidade se considerado o método Cidaq sem transformação prévia dos dados. No entanto, para análises mais detalhadas, com base em toda a estrutura da distribuição (os decis e variância, por exemplo), o impacto do tratamento pode ser maior. Vale destacar as diferenças para as estimativas de desvio padrão, que diminuem considerando quaisquer dos métodos tanto para a variável estatura (em até 0,87 centímetro) quanto para o peso (em até 0,78 quilograma). Isso quer dizer que a distribuição dos dados fica mais concentrada e, além disso, a estrutura de correlação entre as variáveis aumenta.

Do total de observações do grupo feminino (121), 10,7% não apresentaram valores para as duas variáveis e 7,4% não registravam informação apenas na variável estatura. Na Tabela 8 é possível observar as diferenças nas estimativas quando é adotado algum dos métodos para tratar os dados e na ausência de tratamento para esse grupo (meninas com idade entre 24 e 25 meses).

Para a variável estatura observa-se[2] que a estimativa de 1º decil aumenta 2,23 centímetros pelo método Cidaq sem transformação prévia. Pelo método Bacon a estimativa de 9º decil diminui em até 2,20 centímetros. As estimativas de mediana e média aumentam,

[2] Vide dados destacados na Tabela 8.

respectivamente, em 0,99 cm (pelo método TRC sem transformação) e 1,11 cm (pelo método Cidaq sem transformação). As variações para peso são menores, destacando-se a estimativa de média que diminui 0,67 quilograma com a utilização do método Cidaq sem transformação prévia.

TABELA 8
Estimativas da estatura (cm) e do peso (kg) para o grupo feminino de 24 a 25 meses de idade, antes e após o tratamento dos dados com os métodos TRC, Bacon e Cidaq

Variável	Estimativas	Tratamento						
		Ausente	Sem transformação			Com transformação		
			TRC+ Poem	Bacon+ Poem	Cidaq	TRC+ Poem	Bacon+ Poem	Cidaq
Estatura	1º decil	78,27	79,02	78,37	80,50	78,40	78,41	79,04
	9º decil	95,00	93,88	92,80	95,75	94,96	92,90	94,40
	Mediana	88,00	88,99	88,07	88,37	87,99	88,08	88,30
	Média	87,53	88,29	87,43	88,64	87,51	87,30	88,23
	Desvio padrão	7,23	5,81	4,79	6,00	6,99	4,90	5,71
Peso	1º decil	10,40	10,50	10,43	10,24	10,40	10,40	10,45
	9º decil	15,00	14,80	14,88	14,18	15,00	14,81	14,78
	Mediana	12,70	12,70	12,50	12,40	12,70	12,50	12,50
	Média	12,99	12,61	12,54	12,32	12,94	12,54	12,50
	Desvio padrão	4,66	1,63	1,58	1,65	4,44	1,52	1,60
Estatura x peso	Correlação	0,20	0,41	0,65	0,41	0,21	0,67	0,30

Fonte: Pesquisa de Orçamentos Familiares 2008-2009, IBGE.
Nota: Estimativas calculadas considerando o plano amostral e a pós-estratificação.

Da mesma forma que para o grupo de meninos, registra-se diminuição na medida de desvio padrão para ambas as variáveis. Nesse caso, as diferenças são mais acentuadas, com reduções de 2,44 centímetros na variável estatura pelo método Bacon sem transformação e de 3,14 quilogramas na variável peso pelo método Bacon com transformação prévia dos dados. As estimativas de correlação também aumentam para as análises desse grupo.

Para as análises que consideram os percentis da distribuição por idade, como apresentado no Relatório de Padrão de Crescimento das Crianças da Organização Mundial de Saúde (WHO, 2006) nas faixas de idade iniciais – estudo com o qual a POF 2008-2009 (IBGE, 2010) faz comparações –, essas diferenças observadas nos decis podem ser importantes para os resultados finais.

Considerações finais

Como resultado da avaliação comparativa dos métodos abordados, o Cidaq, que foi aplicado à Pesquisa de Orçamentos Familiares em suas duas edições (2002-2003 e 2008-2009), mostrou, de forma geral, maior eficiência no tratamento dos dados em relação aos métodos TRC e Bacon associados ao algoritmo de imputação Poem.

O Cidaq apresentou maior proporção de detecções de *outliers* feitas corretamente nos três níveis de contaminação estudados e em ambos os cenários de simulação. Quanto à

estimação, verificou-se que o viés causado pela contaminação foi bem corrigido para todos os métodos empregados, em qualquer dos níveis de contaminação. O impacto da imputação foi semelhante entre os métodos. O Cidaq apresentou uma pequena vantagem em relação aos demais nos resultados da simulação paramétrica, enquanto para simulação não paramétrica destacou-se o método Bacon.

Verificou-se, ainda, que na simulação não paramétrica a transformação prévia dos dados favoreceu o desempenho dos métodos de detecção comparando-se à simulação paramétrica; e que o tratamento tem maior efeito quanto maior é o nível de contaminação do conjunto de dados.

Se houver disponibilidade do uso do *software* SAS, linguagem na qual está implementado o método Cidaq, esse é recomendado para o tratamento de dados antropométricos, visto que apresentou desempenho superior na detecção de casos suspeitos se comparado aos algoritmos Bacon e TRC. Caso haja restrição de *software*, o algoritmo Bacon associado ao Poem, que pode ser facilmente implementado no *software* R, é mais recomendado do que o algoritmo TRC associado ao Poem.

Esses resultados são importantes para mostrar como a falta de tratamento dos dados pode comprometer, em diferentes níveis (dependendo da quantidade de erros), a qualidade de uma pesquisa, afetando as análises do estado nutricional dos indivíduos de uma população produzidas com base em suas informações. Entre essas estão a prevalência de obesidade e/ou desnutrição e as curvas de crescimento das populações. Além disso, o estudo valida a qualidade do Cidaq diante de métodos mais recentes para o tratamento de dados antropométricos.

Referências

ABEP et al. **Minicurso**: aspectos metodológicos e operacionais da PNDS 2006. [S.l.], 2008. Disponível em: ‹http://bvsms.saude.gov.br/bvs/pnds/img/Minicurso_PNDS2006_8_Antropometria.pdf›. Acesso em: mar. 2016.

ARAUJO, J. D. de. Polarização epidemiológica no Brasil. **Epidemiologia e Serviços de Saúde**, Brasília, v. 21, n. 4, dez. 2012.

BATISTA, M. F.; RISSIN, A. A transição nutricional no Brasil: tendências regionais e temporais. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 19, sup. 1, p. S181-S191, 2003.

BÉGUIN, C.; HULLIGER, B. **Robust multivariate outlier detection and imputation with incomplete survey data**. Deliverable D4/5.2.1/2 Part C, EUREDIT Project, 2003.

BILLOR, N.; HADI, A. S.; VELLEMANN, P. F. Bacon: Blocked Adaptive Computationally-Efficient Outlier Nominators. **Computational Statistics and Data Analysis**, v. 34, n. 3, p. 279-298, 2000.

IBGE – Instituto Brasileiro de Geografia e Estatística. **Pesquisa de Orçamentos Familiares 2008-2009**: antropometria e estado nutricional de crianças, adolescentes e adultos no Brasil. Rio de Janeiro: IBGE, 2010. Disponível em: ‹http://www.ibge.gov.br/home/estatistica/populacao/condicaodevida/pof/2008_2009_encaa/pof_20082009_encaa.pdf›. Acesso em: out. 2014.

LITTLE, R. J. A.; SMITH, P. J. Editing and imputation for a quantitative survey data. **Journal of the American Statistical Association**, v. 82, n. 397, Mar. 1987.

SILVA, P. L. N. **Crítica e imputação de dados quantitativos utilizando o SAS**. Informes de Matemática. Série D. Rio de Janeiro: Instituto de Matemática Pura e Aplicada, 1989.

WORLD HEALTH ORGANIZATION. **Physical status**: the use and interpretation of anthropometry. Geneva, 1995 (Technical Report Series, 854). Disponível em: ‹http://whqlibdoc.who.int/trs/WHO_TRS_854.pdf›. Acesso em: out. 2014.

_____. **WHO child growth standards**: length/height-for-age, weight-for-age, weight-forlength, weight-for-height and body mass index-for-age: methods and development. Geneva, 2006. Disponível em: ‹http://www.who.int/childgrowth/standards/Technical_report.pdf›. Acesso em: out. 2014.

_____. **Obesity**: preventing and managing the global epidemic. Geneva, 2000 (WHO Technical Report Series, 894). Disponível em: ‹http://www.who.int/nutrition/publications/obesity/WHO_TRS_894/en/› Acesso em: mar. 2016.

Sobre os autores

Mariana Vieira Martins de Matos é bacharel em Estatística pela Universidade de Brasília e mestre em Estudos Populacionais e Pesquisas Sociais pela Escola Nacional de Ciências Estatísticas.

Pedro Luis do Nascimento Silva é bacharel em Estatística pela Escola Nacional de Ciências Estatísticas, mestre em Estatística pelo Instituto de Matemática Pura e Aplicada e doutor em Estatística pela Universidade de Southampton, Inglaterra. Pesquisador titular da Escola Nacional de Ciências Estatísticas.

Endereço para correspondência

Mariana Vieira Martins de Matos
Escola Nacional de Ciências Estatísticas
Rua André Cavalcanti, 106 – Bairro de Fátima
20231-050 – Rio de Janeiro-RJ, Brasil

Pedro Luis do Nascimento Silva
Escola Nacional de Ciências Estatísticas
Rua André Cavalcanti, 106 – Bairro de Fátima
20231-050 – Rio de Janeiro-RJ, Brasil

Abstract

Comparison of methods for the treatment of anthropometric measures of POF 2008-2009

The 2008-2009 Pesquisa de Orçamentos Familiares (Household Budget Survey) is a nationwide sample survey that collected anthropometric data on height and weight of individuals in Brazil. Due to the size of the research, the collection process allows contamination in the collected data by non-sampling errors and non-response. Such errors can affect the indicators of malnutrition, prevalence of overweight persons and obesity, and produce differing effects in different population segments. In this study, the Cidaq approach – the methodology employed in POF 2008-2009 to tackle these problems and preserve the quality of the data – was compared with two other approaches – namely, the TRC algorithm and the Bacon algorithm, both coupled with the Poem algorithm. Such comparisons are essential to ensure the choice of the best method future research efforts in order to ensure reliability of the data used in population studies that

support the planning of public policies in health, nutrition, social assistance and others. The approaches were compared by simulation through impact in mean, standard deviation and correlation estimates of the weight and height. The Cidaq approach proved to be more efficient than the others under the parametric simulation, while the Bacon approach showed to be better under the non parametric simulation.

Keywords: Editing. Imputation. Anthropometric measures. Outliers.

Resumen

Comparación de los métodos para el tratamiento de las medidas antropométricas de la POF 2008-2009

La Pesquisa de Orçamentos Familiares 2008-2009 (Encuesta de Presupuestos Familiares) es una encuesta por muestreo a nivel nacional que contempla los datos antropométricos de peso y talla de las personas en Brasil. Siendo una extensa encuesta, la contaminación por errores ajenos al muestreo y la falta de respuesta son inherentes en el proceso de recolección. Estos tipos de errores pueden cambiar los indicadores de prevalencia de desnutrición, sobrepeso u obesidad y afectar diferencialmente los diferentes segmentos de la población analizados. Este artículo comparó el rendimiento del método Cidaq, utilizado para tratar los datos antropométricos de esta encuesta, a los otros dos métodos: el algoritmo TRC y el algoritmo Bacon, ambos asociados con el algoritmo Poem. Esta comparación es esencial para asegurar que el mejor método pueda ser utilizado en futuras investigaciones para estudios de población con la finalidad de subvencionar la planificación de políticas públicas en los ámbitos de la salud, la nutrición, la asistencia social y otras. Los métodos fueron comparados por medio de la simulación y calculando el impacto en estimaciones en términos de promedios, desviación estándar y correlación entre el peso y la altura. El método Cidaq fue más eficiente que los otros en la simulación paramétrica y el método Bacon presenta ventajas en la simulación no paramétrica.

Palabras-claves: Edición. Imputación. Medidas antropométricas. Valores atípicos.

Recebido para publicação em 13/12/2014
Recomendado para publicação em 18/03/2016
Aceito para publicação em 09/04/2016