http://dx.doi.org/10.18222/eae.v0ix.3916

DADOS AUSENTES EM AVALIAÇÕES EDUCACIONAIS: COMPARAÇÃO DE MÉTODOS DE TRATAMENTO

LUIS GUSTAVO DO AMARAL VINHA
JACOB ARIE LAROS

RESUMO

Os dados ausentes são comuns nas avaliações educacionais. Por isso, o uso de métodos adequados torna-se fundamental para reduzir o impacto da perda de informação. O objetivo deste estudo é comparar o desempenho de quatro métodos de tratamentos de dados ausentes (imputação pela média, listwise deletion*, máxima verossimilhança e imputação múltipla), tendo como base o uso de modelos de regressão aplicados aos dados da avaliação educacional realizada no estado do Ceará. Foram utilizadas informações de 7.000 estudantes, simulando-se diversos cenários de acordo com o percentual e o tipo de ausência. A imputação pela média apresentou o pior desempenho em todos os cenários simulados e os demais métodos mostraram resultados semelhantes entre si. Verificou-se ainda que o uso de variáveis auxiliares na estimação por máxima verossimilhança e imputação múltipla reduziu o viés das estimativas de parâmetros importantes do modelo quando a ausência simulada não é ao acaso.*

PALAVRAS-CHAVE TRATAMENTO DE DADOS AUSENTES • AVALIAÇÃO DA EDUCAÇÃO • DESEMPENHO ACADÊMICO • SIMULAÇÃO.

DATOS AUSENTES EN EVALUACIONES EDUCACIONALES: COMPARACIÓN DE MÉTODOS DE TRATAMIENTO

RESUMEN

Los datos ausentes son comunes en las evaluaciones educacionales. Por ello el uso de métodos adecuados se hace fundamental para reducir el impacto de la pérdida de información. El objetivo de este estudio es comparar el desempeño de cuatro métodos de tratamientos de datos ausentes (imputación por el promedio, listwise deletion, máxima verosimilitud e imputación múltiple), en base al uso de modelos de regresión aplicados a los datos de la evaluación educacional realizada en el estado de Ceará. Se utilizaron informaciones de 7.000 estudiantes, simulando diversos escenarios de acuerdo al porcentual y al tipo de ausencia. La imputación por el promedio presentó el peor desempeño en todos los escenarios simulados y los demás métodos mostraron resultados similares entre ellos. También se verificó que el uso de variables auxiliares en la estimación por máxima verosimilitud e imputación múltiple redujo el sesgo de las estimaciones de parámetros importantes del modelo cuando la ausencia simulada no se debe a la casualidad.

PALABRAS CLAVE TRATAMIENTO DE DATOS AUSENTES • EVALUACIÓN DE LA EDUCACIÓN • DESEMPEÑO ACADÉMICO • SIMULACIÓN.

MISSING DATA IN EDUCATIONAL ASSESSMENT: A COMPARISON OF DATA TREATMENT METHODS

ABSTRACT

Missing data are common in educational assessments. Using the appropriate methods has, therefore, become essential to reduce the impact of the loss of information. The present study aims to compare the performance of four methods for dealing with missing data (mean imputation, listwise deletion, maximum likelihood and multiple imputation), all based on regression models applied to the educational assessment of data collected in the State of Ceará. Information about 7,000 students was used, simulating various scenarios according to the percentage and the type of the missing data. The mean imputation method showed the worst performance in all simulated scenarios and the other methods showed similar results among themselves. Moreover, the use of auxiliary variables in the estimation by maximum likelihood and multiple imputation proved to reduce the bias of estimates of some important parameters of the model, when the simulated missing data is not random.

KEYWORDS TREATMENT OF MISSING DATA • EDUCATION ASSESSMENT • ACADEMIC PERFORMANCE • SIMULATION.

INTRODUÇÃO

Apesar de a literatura estatística já abordar o tratamento de dados ausentes há décadas, o assunto ainda é um enigma na pesquisa social aplicada. Muitos pesquisadores não utilizam as técnicas adequadas por falta de familiaridade e, em geral, empregam métodos simples de eliminação ou substituição, que na maioria das situações não são apropriados (COX et al., 2014; MCKNIGHT et al., 2007; PEUGH; ENDERS, 2004). Trata-se, sem dúvida, de um assunto delicado, pois a ausência da informação pode ser causada por diversos fatores, apresentar diferentes padrões e distorcer os resultados da pesquisa, uma vez que a maioria das técnicas estatísticas foi desenvolvida para dados completos.

A preocupação com o tratamento de dados ausentes na pesquisa social pode ser observada em algumas revisões de literatura. Peugh e Enders (2004) revisaram estudos publicados nas áreas de educação e psicologia aplicada com o objetivo de inventariar como os pesquisadores dessas áreas reportavam a ausência de informação e os procedimentos usados nas análises. Os autores verificaram que a presença

e o tratamento de dados ausentes eram raramente reportados. Na maioria dos casos, os valores ausentes tinham que ser identificados pela comparação do tamanho da amostra e o número de graus de liberdade nas diversas análises realizadas. Além disso, em quase todos os trabalhos em que foi identificada a presença de valores ausentes, foram utilizados métodos tradicionais de eliminação ou substituição. Segundo os autores, entre 1999 e 2003, aumentou o número de estudos em que os valores ausentes foram reportados, entretanto, técnicas mais sofisticadas baseadas na estimação por máxima verossimilhança ou imputação múltipla foram pouco utilizadas.

Resultados semelhantes foram encontrados por Rousseau *et al.* (2012), que apresentam uma revisão rigorosa de artigos publicados no *British Educational Research Journal*, no período de 2003 a 2007. Dos 68 estudos selecionados, mais de um terço não indicava qualquer informação sobre dados ausentes. Na metade dos trabalhos em que a presença de valores ausentes foi reportada, o tratamento adotado não estava explícito e, entre os que mencionaram o tratamento, a maioria utilizou métodos de eliminação de observações.

Discutindo também quais seriam os motivos que levam a tamanha variação nos relatos dos estudos analisados, Rousseau *et al.* (2012) acreditam que muitos pesquisadores não mencionam os valores ausentes por ignorar a natureza desses dados, como eles influenciam os resultados, quais métodos estão disponíveis para análise e como reportá-los. Os autores afirmam ainda que não existe na área de educação um procedimento que possa ser usado pelos pesquisadores – como acontece na psicologia com os manuais da APA (*American Psychology Association*) – e que a literatura estatística relacionada ao tema é muito técnica para esse público. Eles ressaltam também que os pesquisadores não mencionam a existência de valores ausentes, ou reportam resumidamente, em função do aumento na complexidade das interpretações e dos efeitos negativos nos resultados gerados pela falta de parte da informação.

No Brasil, considerando os estudos em avaliação educacional, verifica-se também que poucos pesquisadores mencionam a presença de valores ausentes nos dados analisados.

Pode-se destacar o estudo de Oliveira, Belluzzo e Pazello (2013), que indicam a retirada das observações com dados faltantes e admitem que esse procedimento pode comprometer os resultados observados. Em Xerxenevsky (2012), as observações incompletas também são excluídas, mas nesse caso a autora afirma que isso não impactaria nos resultados, em função do tipo de dado ausente. Em alguns trabalhos foram utilizadas variáveis indicadoras no ajuste dos modelos como uma forma de reduzir o impacto das informações ausentes (MACEDO, 2004; RODRIGUES; RIOS-NETO; PINTO, 2011; SOARES; ALVES, 2003).

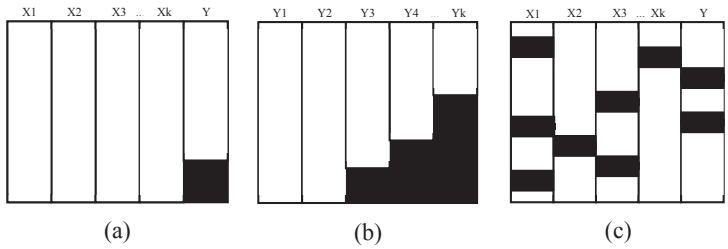
O presente estudo visa a contribuir para o melhor entendimento do impacto dos valores ausentes nos resultados de avaliações educacionais. Para isso, além de discutir conceitos relacionados e apresentar alguns métodos de tratamento e análise de dados com informações faltantes, o artigo procura comparar o desempenho de diferentes procedimentos de análise em diferentes situações, tendo como base dados reais oriundos de uma avaliação educacional de larga escala realizada no Brasil. Espera-se que este trabalho seja útil para o melhor entendimento do problema, mesmo para leitores que não estão familiarizados com análise de dados e conceitos estatísticos.

TIPOS DE DADOS AUSENTES

Algumas classificações relacionadas a dados ausentes são encontradas na literatura. Por exemplo, McKnight *et al.* (2007) utilizam uma classificação de ausência de acordo com a fonte: casos ausentes, variáveis ausentes ou ocasiões ausentes. Os casos ausentes acontecem quando o participante não fornece qualquer informação, o que ocorre com a ausência do indivíduo selecionado no momento da coleta de dados. Variáveis ausentes são observadas quando o participante não fornece parte da informação requerida, por exemplo, quando ele não responde um ou mais itens do questionário. Ocasiões ausentes são comuns em estudos longitudinais e ocorrem quando o participante não está presente em todos os momentos de coleta de dados.

Os dados ausentes podem também ser classificados de acordo com o padrão de não resposta (SCHAFER; GRAHAM, 2002). No padrão univariado apenas uma variável apresenta valores ausentes. Por exemplo, em um modelo de regressão, a ausência ocorre somente na variável dependente (Figura 1a). O padrão monótono é geralmente observado em estudos longitudinais em decorrência do abandono de participantes ao longo das avaliações (Figura 1b). Nesse caso, supondo-se que *Y* seja a característica acompanhada ao longo do tempo, os indivíduos com valor ausente em Y_t também apresentarão valores ausentes em Y_{t+1}, Y_{t+2} e assim por diante, para qualquer *t*. Por fim, como mostrado na Figura 1c, no padrão arbitrário os valores ausentes podem ocorrer em uma ou mais variáveis para qualquer observação.

FIGURA 1 – Padrões de não resposta: (a) univariado, (b) monótono e (c) arbitrário



Fonte: Adaptado de Schafer e Graham (2002).

Entre as diversas classificações encontradas na literatura, aquela proposta por Rubin (1976, 1987) é a mais importante. Segundo o autor, os valores ausentes são gerados por três mecanismos distintos que relacionam a propensão de ausência aos dados observados: valores ausentes completamente ao acaso (MCAR, do inglês *Missing Completely at Random*); ausentes ao acaso (MAR, *Missing at Random*); e ausentes não ao acaso (MNAR, *Missing not at Random*).

Os dados são classificados como ausentes completamente ao acaso (MCAR) quando a ocorrência não está relacionada a qualquer variável observada no estudo ou à própria variável que apresenta os valores faltantes. Dessa forma, o mecanismo gerador desses dados não está relacionado a qualquer

característica observada; daí a denominação de ocorrência completamente ao acaso. Esse tipo de ausência poderia ser interpretado como a retirada de uma amostra aleatória de observações do banco de dados completo. Por exemplo, valores faltantes na proficiência em leitura em uma avaliação educacional podem ser consequência de eventos diversos, como adoecimento do estudante.

Os dados faltantes são classificados como MAR quando a ocorrência está relacionada aos valores observados de outras variáveis, mas independe do valor da variável em questão. Por exemplo, pode-se supor que alunos com baixa renda familiar apresentam maiores taxas de ausência nas provas, no entanto, para uma determinada faixa de renda familiar, não é observada qualquer relação entre a ausência e o desempenho (PEUGH; ENDERS, 2004). Nesse caso, o nome dado ao mecanismo pode gerar dúvida (COLLINS; SCHAFER; KAM, 2001; GRAHAM, 2009), uma vez que a ausência de dados não é de fato ao acaso. Pode-se, então, interpretar da seguinte forma: a taxa de valores ausentes está relacionada à renda, mas, quando controlada a faixa de renda, a ausência passa a ser aleatória (a incidência é aleatória depois do controle de uma variável observada). Na prática não é possível confirmar se a ocorrência de valores ausentes está apenas relacionada com as outras variáveis e não com os valores da própria variável, pois não se conhecem os valores faltantes. No exemplo, entre os indivíduos de baixa renda, não é possível verificar se os valores ausentes são dos alunos com menor proficiência (ENDERS, 2010).

Por fim, os dados faltantes são identificados como MNAR quando a ocorrência está relacionada aos valores da própria variável analisada. Esses dados são mais difíceis de serem identificados, já que a ocorrência está relacionada a valores não observados. Na avaliação educacional, esse fato pode ser exemplificado pela maior taxa de valores faltantes em testes de proficiência entre os alunos com menor desempenho (mesmo depois de controladas outras variáveis).

A classificação proposta por Rubin está diretamente relacionada ao impacto da ausência de informação e à escolha da abordagem mais apropriada para análise dos dados. Os que menos influenciam as análises e os resultados são os

dados ausentes do tipo MCAR, uma vez que a amostra de valores completos pode ser vista como representativa da população. Quando os dados faltantes são do tipo MAR, a ausência pode ser considerada ignorável, pois se faz necessária a modelagem adicional do mecanismo de ausência de dados no processo de estimação. Por fim, os dados MNAR também são chamados de não ignoráveis, dado que o mecanismo gerador de ausência deve ser modelado para que sejam obtidas boas estimativas dos parâmetros de interesse (ALLISON, 2002).

MÉTODOS DE TRATAMENTO E ANÁLISE DE DADOS INCOMPLETOS

Nesta seção, são apresentados métodos tradicionais baseados na eliminação e substituição de valores, além de dois procedimentos mais sofisticados baseados em métodos de estimação.

MÉTODOS DE ELIMINAÇÃO E IMPUTAÇÃO SIMPLES

De acordo com as revisões de literatura mencionadas anteriormente, os procedimentos baseados na *eliminação de observações* são utilizados na maior parte dos estudos em que os dados ausentes são identificados (CHEEMA, 2014; PEUGH; ENDERS, 2004; ROUSSEAU et al., 2012). Acredita-se também que muitos pesquisadores acabam utilizando esses métodos inconscientemente, por ser o procedimento padrão dos pacotes estatísticos (ROUSSEAU et al., 2012). Os procedimentos baseados na eliminação de observações são *listwise deletion* (ou *complete-case analysis*) e *pairwise deletion* (*available case analysis*).

LISTWISE DELETION

Neste procedimento, todos os casos com um ou mais valores ausentes nas variáveis observadas são retirados; logo, consideram-se apenas as observações completas. Como consequência, o poder dos testes estatísticos se reduz com a perda de informação e pode gerar estimativas viesadas dos parâmetros quando os dados não são MCAR.

Entretanto, é importante ressaltar que esse método apresenta vantagens que o tornam atrativo em algumas situações. Trata-se de um método simples que não gera maior

complexidade nas análises e interpretação de resultados e, como mencionado anteriormente, é o procedimento padrão dos pacotes estatísticos. Em algumas situações o *listwise deletion* mostra desempenho semelhante a técnicas mais sofisticadas, como, por exemplo, quando os dados faltantes são do tipo MCAR, o tamanho da amostra é grande e o número de faltantes é relativamente pequeno (ENDERS, 2010). Ainda, no ajuste de modelos de regressão, quando os valores ausentes dependem de uma variável presente no modelo (tipo MAR), o procedimento tem bom desempenho (ALLISON, 2002; SCHAFER; GRAHAM, 2002).

PAIRWISE DELETION

A utilização do procedimento *listwise deletion* gera a perda de uma parcela considerável da informação contida no banco de dados, principalmente quando o número de variáveis envolvidas no estudo aumenta. O procedimento *pairwise deletion* surge então como alternativa para reduzir essa perda de informação.

Nesse caso, a parcela de dados utilizada nos cálculos é maior, uma vez que são consideradas observações completas para pares de variáveis e não todo o conjunto de informações de um indivíduo. Esse procedimento é muito comum quando são empregadas técnicas baseadas em correlações, dado que a estimação da correlação depende apenas de um par de valores. No entanto, tal abordagem é criticada porque as amostras utilizadas nos diferentes cálculos não são as mesmas, o que pode gerar problemas de estimação e inconsistências de resultados. Ainda, esse procedimento também pode gerar viés se os dados ausentes não são MCAR (ALLISON, 2002).

As *técnicas de imputação simples* podem ser mais atrativas do que os métodos de eliminação, uma vez que não existe descarte de informação, mas são muito criticadas. Com a utilização de técnicas de imputação simples, os valores imputados são tratados como conhecidos nas análises, o que pode distorcer os resultados. Além disso, alguns pesquisadores evitam o uso da imputação por acreditar que estariam "construindo os dados" (GRAHAM, 2009). A seguir

são apresentados os principais métodos de imputação simples encontrados na literatura.

IMPUTAÇÃO PELA MÉDIA

Essa técnica consiste na substituição dos valores ausentes pela média da variável, obtida a partir dos valores válidos observados na amostra. Segundo Enders (2010), trata-se de um procedimento antigo, cuja autoria é frequentemente atribuída a Wilks (1932). A imputação pela média é ainda muito empregada, mas produz distorções mesmo quando os dados são MCAR. Com a utilização desse procedimento, a variabilidade dos dados pode ser subestimada e, como consequência, há um aumento da chance de se rejeitarem hipóteses nulas em testes de significância, além de afetar a estimação de medidas de associação.

O método de imputação pela média é possivelmente o pior tratamento de dados ausentes (ENDERS, 2010) e fornece estimadores viesados para todos os parâmetros, com exceção da média, para qualquer que seja o tipo de dado ausente analisado.

IMPUTAÇÃO PELA REGRESSÃO

A imputação dos dados pela regressão foi proposta por Buck (1960) e, em geral, oferece melhores resultados do que a imputação pela média. Nesse caso, o valor ausente é substituído pelo valor obtido pela equação de regressão utilizando as demais variáveis que apresentam valores completos como preditoras. Esse método pode ser indicado quando poucas variáveis possuem observações faltantes. Já nas situações em que muitas variáveis têm dados ausentes, tal técnica perde a atratividade, pois o número de equações a serem estimadas é grande (ENDERS, 2010).

Como a substituição é feita por um valor predito baseado nas relações entre as variáveis, o valor substituído está exatamente na linha que descreve a relação, o que pode resultar na superestimação das covariâncias entre as variáveis (PEUGH; ENDERS, 2004). Por outro lado, a subestimação da variabilidade é menos acentuada do que a observada com a utilização da imputação pela média e o procedimento gera

estimativas não viesadas quando os dados são do tipo MAR.

Para atenuar os possíveis problemas decorrentes desse método, pode-se utilizar substituição pela regressão estocástica (*stochastic regression imputation*). Nesse caso o valor a ser substituído é composto pelo valor predito pela regressão mais uma parcela aleatória, proveniente da distribuição dos resíduos do modelo. Esse método também pode ser encontrado na literatura como *single random imputation* (CHEEMA, 2014).

HOT DECK IMPUTATION

Desenvolvido pelo *US Census Bureau* (ENDERS, 2010), o procedimento *hot deck imputation* propõe a substituição do valor ausente por um valor observado em unidades similares (ANDRIDGE; LITTLE, 2010). Para uma observação com ausência de informação são identificados os possíveis doadores, observações que apresentam valores válidos para a variável a ser tratada e com respostas semelhantes para as demais variáveis coletadas. Por exemplo, o valor observado de apenas um doador selecionado aleatoriamente pode ser usado nessa substituição, ou pode-se utilizar o dado do doador mais próximo, de acordo com alguma métrica preestabelecida. Em outras versões do método, os valores referentes a um conjunto de doadores são combinados, por exemplo, pela média aritmética, sendo que o resultado é então usado na substituição.

Uma característica importante desse método refere-se ao fato de a substituição ser feita sem a suposição de modelo. Porém, como Andridge e Little (2010) ressaltam, o procedimento depende da medida usada e das variáveis envolvidas na escolha de doadores. O método apresenta bons resultados quando os dados são do tipo MCAR e o número de observações é elevado.

VARIÁVEIS INDICADORAS

O uso de variáveis indicadoras de ausência no ajuste dos modelos foi proposto por Cohen e Cohen (1985). Esse método tem sido empregado pelos pesquisadores da área da educação (COX et al., 2014) e também pode ser encontrado na

literatura brasileira em avaliação educacional (MACEDO, 2004; RODRIGUES; RIOS-NETO; PINTO, 2011; SOARES; ALVES, 2003). O procedimento visa a reduzir a perda de informação e consiste na imputação dos valores faltantes pela média e inclusão de variáveis que indicam a ausência de informação. Com essa reparametrização um novo intercepto para categorias de dados ausentes é criado e, sob a hipótese de que os dados são MCAR, as estimativas dos parâmetros de interesse não são afetadas (RODRIGUES; RIOS-NETO; PINTO, 2011). Contudo, quando os valores ausentes não são MCAR, o procedimento pode gerar viés nas estimativas dos parâmetros de interesse (COX et al., 2014).

MÉTODOS DE ESTIMAÇÃO E IMPUTAÇÃO MÚLTIPLA

As abordagens que utilizam métodos de estimação para o tratamento de dados incompletos são apresentadas a seguir. Segundo Peugh e Enders (2004), diversos estudos têm apontado a superioridade desses métodos em relação aos tradicionais apresentados anteriormente. Ainda, Graham (2009) afirma que os pesquisadores deveriam empregar os procedimentos baseados na máxima verossimilhança e na imputação múltipla, pois são os melhores métodos disponíveis e baseiam-se em conceitos fortes e tradicionais da estatística.

MÁXIMA VEROSSIMILHANÇA

O tratamento dos valores ausentes pelo método da máxima verossimilhança é semelhante ao utilizado para dados completos (ENDERS, 2010). O método consiste na escolha de valores para os parâmetros do modelo que maximizam a função de verossimilhança, a qual expressa a probabilidade de se observarem os valores obtidos na amostra para um determinado modelo escolhido. Esse procedimento requer a suposição de que a ausência é do tipo MAR, além de uma suposição relacionada à distribuição. Em geral assume-se que os dados têm distribuição normal multivariada.

A estimação dos parâmetros é realizada através da maximização da função de verossimilhança, o que, na maior parte das situações, não pode ser realizado analiticamente.

Faz-se necessário o uso de métodos numéricos, como, por exemplo, o algoritmo EM. Esse método de maximização é muito popular quando os dados considerados na estimação não estão completos (ALLISON, 2002).

O algoritmo EM é um procedimento iterativo que consiste na repetição de dois passos: estimação (E) e maximização (M). O processo se inicia com uma estimação do vetor de médias e da matriz de covariâncias, utilizando apenas os dados completos. No passo E os elementos do vetor de médias e da matriz de covariâncias são utilizados para construir um conjunto de equações de regressão usadas para estimar os valores ausentes com base nas variáveis observadas. Em seguida, no passo M, o vetor de médias e a matriz de covariâncias são reestimados por meio dos estimadores de máxima verossimilhança com base em dados completos (valores observados e os substituídos no passo anterior). Em uma nova etapa E, as equações de regressão são reestimadas e os valores ausentes substituídos por novos valores. Na etapa M seguinte, esse novo banco de dados é usado para reestimar o vetor de médias e a matriz de covariâncias, sendo que o processo é repetido até que as estimativas de médias e variâncias não mudem mais (usando um critério de convergência preestabelecido).

O algoritmo EM apresenta algumas vantagens importantes que o tornam muito atrativo (ALLISON, 2002). Trata-se de um procedimento de fácil implementação, não requer a definição de muitos parâmetros e está disponível em diversos pacotes estatísticos. Entretanto, considerando o ajuste de modelos de regressão, os erros padrão associados aos coeficientes não são obtidos diretamente (ENDERS, 2010). Como alternativa, pode-se utilizar o procedimento FIML (*full information maximum likelihood*) que, assim como o algoritmo EM, gera estimativas não viesadas para os coeficientes da regressão e ainda estima diretamente os erros associados (ENDERS, 2001a).

A estimação pelo método da máxima verossimilhança possibilita a utilização de variáveis auxiliares, que não fazem parte do modelo principal de análise, mas estão associadas ao mecanismo gerador de valores ausentes ou à variável a ser

tratada. A inclusão de variáveis auxiliares aumenta a chance de satisfazer a suposição de ausência do tipo MAR, e assim melhorar a estimação (BARALDI; ENDERS, 2010; COLLINS; SCHAFER; KAM, 2001). É importante ressaltar que a inclusão de variáveis auxiliares de forma adequada não altera a interpretação dos parâmetros do modelo principal.

De forma geral, a estimação por máxima verossimilhança apresenta diversas propriedades interessantes no tratamento de dados ausentes, entretanto, não se trata de um método infalível. Entre os problemas relacionados, pode-se destacar que a suposição de distribuição normal multivariada geralmente não está satisfeita, o que pode gerar viés nas estimativas e erros associados. Nas situações em que a ausência não é ao acaso (MNAR), esse procedimento pode distorcer os resultados, apesar de apresentar desempenho superior aos métodos tradicionais, principalmente quando são utilizadas variáveis auxiliares (GRAHAM, 2009).

IMPUTAÇÃO MÚLTIPLA

Na imputação múltipla, proposta por Rubin (1987), os valores ausentes são repetidamente substituídos por valores obtidos a partir da simulação da distribuição condicional de probabilidade, tendo como resultado múltiplas versões do banco de dados. Cada versão do banco de dados é analisada de acordo com as técnicas usuais e os resultados são combinados gerando estimativas pontuais dos parâmetros de interesse (ROSE; FRASER, 2008).

Esse procedimento é composto por três etapas: imputação, análise e combinação. Na etapa de imputação, composta por dois passos, são gerados *m* novos bancos de dados. No primeiro passo (passo I), o vetor de médias e a matriz de covariâncias são estimados e é construído um sistema de equações de regressão para imputação dos valores ausentes, como no método de substituição pela regressão estocástica. No segundo passo (passo II), o vetor de média e a matriz de covariâncias são estimados novamente, e a partir dessas estimativas são geradas novas estimativas com a adição de um termo aleatório (esses novos valores correspondem à retirada de uma amostra da distribuição *a posteriori* da matriz

de covariâncias e do vetor de médias). Essas novas estimativas de médias e covariâncias são usadas no passo I seguinte, sendo que o processo se repete até que os *m* conjuntos de dados completos sejam criados.

Algumas considerações importantes são necessárias nessa etapa. Primeiro, a decisão de quais variáveis serão incluídas na etapa de imputação é um aspecto relevante para o bom desempenho da imputação múltipla (ENDERS, 2010). As variáveis presentes no modelo principal devem sempre ser incluídas. Variáveis auxiliares podem ser consideradas nessa etapa, sem correr o risco de introduzir viés nos resultados, cuidando apenas para não ser introduzido um número muito elevado de variáveis, o que poderia causar problemas de convergência. Segundo, deve-se considerar um número de iterações antes da retirada do primeiro banco de dados e entre as retiradas dos outros bancos de dados. As iterações iniciais são necessárias para estabilização da distribuição dos parâmetros e as iterações entre as retiradas asseguram a independência entre os dados gerados. Por fim, a eficiência do procedimento está relacionada com o número de bancos de dados gerados (*m*) e, de forma geral, quanto maior o percentual de valores ausentes, maior deve ser *m* (para mais detalhes ver GRAHAM; OLCHOWSKI; GILREATH, 2007).

A etapa de análise consiste na aplicação das técnicas estatísticas usuais aos *m* conjuntos de dados gerados na etapa anterior, de acordo com os objetivos da pesquisa. Como resultado dessa etapa, *m* conjuntos de estimativas para os parâmetros de interesse são gerados. Por fim, na última etapa são calculadas as estimativas dos parâmetros de interesse a partir da combinação das *m* estimativas. A média aritmética das *m* estimativas pode ser usada para gerar a estimativa combinada de um parâmetro, porém essa combinação será válida quando a distribuição do estimador se aproxima da normal. Alguns estimadores têm distribuições assimétricas, como os estimadores de variância e covariâncias, e, nesses casos, transformações podem ser utilizadas para melhorar as estimativas (ENDERS, 2010).

Os procedimentos de imputação múltipla e máxima verossimilhança são baseados nas mesmas suposições de distribuição normal multivariada e mecanismo gerador de dados

ausentes do tipo MAR. Em geral, os resultados obtidos por meio desses métodos são semelhantes, especialmente quando o tamanho da amostra é grande. Segundo Collins, Schafer e Kam (2001), os procedimentos apresentam resultados muito próximos quando é usado o mesmo banco de dados, com as mesmas suposições de relações entre as variáveis e suas distribuições.

ESTUDO COMPARATIVO

O estudo apresentado neste manuscrito visa a comparar métodos de tratamento de dados ausentes, tendo como base um conjunto de dados reais de uma avaliação educacional. Foram selecionados para essa comparação dois procedimentos tradicionais e que ainda são muito utilizados pelos pesquisadores (PEUGH; ENDERS, 2004) – a imputação simples pela média (Me) e o *listwise deletion* (LD) – e os métodos baseados na estimação por máxima verossimilhança (MV) e na imputação múltipla (IM). Também foi avaliada a utilização de variáveis auxiliares nos procedimentos MV e IM. Este estudo baseia-se na análise dos dados por meio de modelos de regressão linear múltipla, tendo como variável resposta o desempenho escolar.

DADOS

Os dados trabalhados são provenientes do Sistema Permanente de Avaliação da Educação Básica (Spaece), disponibilizados pela Secretaria de Educação do Estado do Ceará. Entre outros levantamentos, esse sistema monitora os alunos do ensino médio da rede pública cearense, tendo como medidas de desempenho escolar as proficiências em Língua Portuguesa e Matemática. As proficiências são estimadas por meio da Teoria da Resposta ao Item (TRI) (CEARÁ, 2011). Além dos testes, os alunos respondem a questionários contextuais com itens relacionados a dados socioeconômicos, hábitos de estudo e clima em sala de aula.[1]

O banco de dados original disponibilizado contém informações relativas a centenas de milhares de estudantes, com grande incidência de valores ausentes. Entretanto, para

1 O questionário contextual pode ser acessado em <http://www.spaece.caedufjf.net/downloads/2009-2/>.

a simulação realizada neste estudo, foi utilizada uma amostra composta apenas por estudantes com dados completos para as variáveis utilizadas (Quadro 1). Essa amostra contém informação referente a 7.000 estudantes matriculados no 1º ano do ensino médio em 2009 e que estavam presentes na avaliação em 2010. Esses estudantes selecionados foram avaliados por meio das provas de proficiência nos dois anos e responderam ao questionário no primeiro ano. As variáveis utilizadas no estudo são apresentadas no Quadro 1.

QUADRO 1 – Variáveis utilizadas no estudo

VARIÁVEIS	DESCRIÇÃO / CODIFICAÇÃO
MAT10	Desempenho em Matemática em 2010 – 2º ano do ensino médio / Variável contínua.
MAT09	Desempenho em Matemática em 2009 – 1º ano do ensino médio / Variável contínua.
LP09	Desempenho em Língua Portuguesa em 2009 – 1º ano do ensino médio/Variável contínua.
MASC	Sexo do aluno / Variável binária: 0, se feminino; 1, se masculino.
SUP	Pretensão de ingresso no ensino superior. Obtida a partir da questão relativa aos planos dos alunos após a conclusão do ensino médio / Variável binária: 1, se pretende; e 0, se tem outros planos.
REPETÊNCIA	Números de vezes que o aluno repetiu um ano escolar, avaliado em 2009 / Variável ordinal: 0, se nunca repetiu; 1, uma repetência; 2, duas repetências; 3, três ou mais repetências.
IDADE	Idade reportada em 2009 (em anos).
MANHÃ	Turno em que o aluno frequenta as aulas. Variável binária que assume o valor 1 se o aluno estudava de manhã em 2009 e 0 se estudava à tarde ou à noite.

Fonte: Secretaria de Educação do Estado do Ceará/Spaece (elaboração dos autores).

Entre os 7.000 estudantes selecionados, 45,3% são do sexo masculino, 43,8% pretendiam ingressar no ensino superior, 32,6% reportaram uma ou mais repetências e 42,0% estudavam no período matutino em 2009. A idade média observada foi de 15,7 anos com desvio-padrão de 1,1 anos e as notas médias em Matemática foram 246,5 e 257,1 pontos em 2009 e 2010, respectivamente, com desvios-padrão de 46,1 e 46,9 pontos. Já a nota média em Língua Portuguesa em 2009 foi de 247,5 pontos com desvio-padrão de 39,9 pontos.

PROCEDIMENTOS

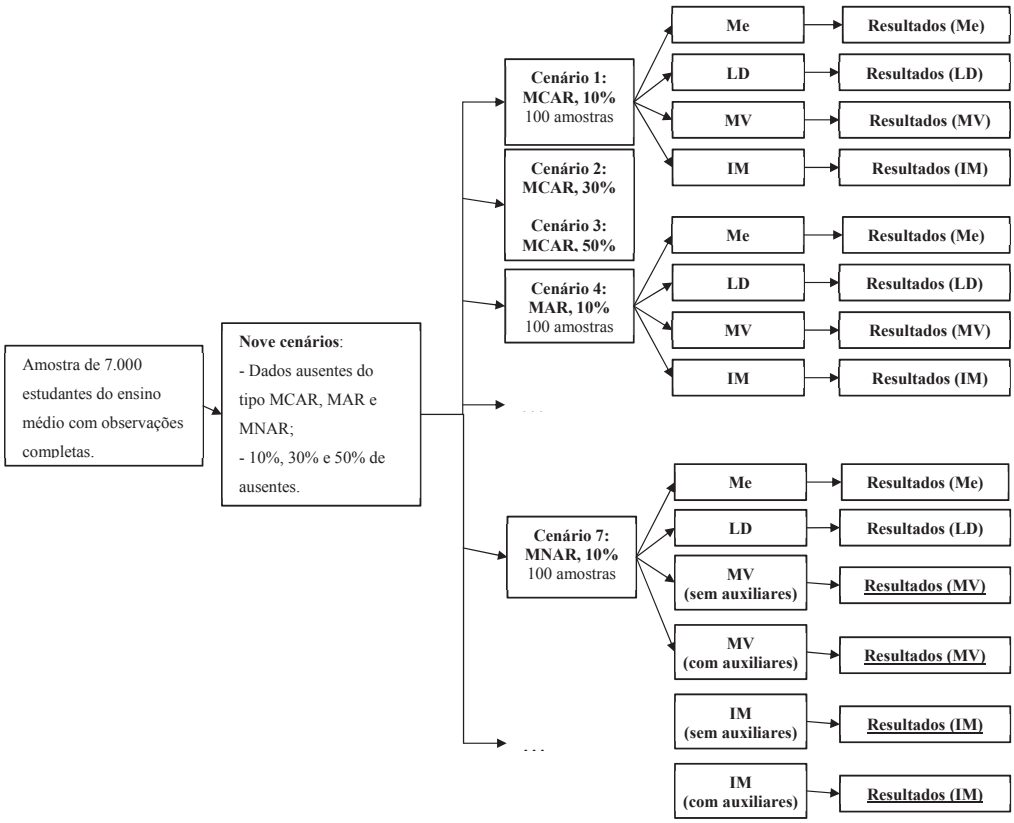
A comparação dos métodos foi realizada supondo-se que essa avaliação seja usada para comparar o desempenho em Matemática no segundo ano do ensino médio, por sexo e intenção de ingresso no ensino superior. Dessa forma, o modelo principal de análise é dado por:

$$\text{MAT10} = \beta_0 + \beta_1\text{MAT09} + \beta_2\text{MASC} + \beta_3\text{SUP} + \varepsilon \quad (1)$$

A variável MAT09 foi incorporada ao modelo como uma variável de controle para que sejam consideradas as diferenças entre os alunos por meio de uma medida inicial de desempenho. Dessa forma, os coeficientes β_2 e β_3 representam as diferenças entre as proficiências médias dos grupos, controladas pela proficiência no ano anterior.

A Figura 2 apresenta o plano utilizado para a simulação. Foram gerados nove cenários diferentes de acordo com o percentual (10%, 30% e 50%) e o tipo de ausência de dados (MCAR, MAR e MNAR). Os dados ausentes foram simulados apenas na variável resposta (MAT10), ou seja, padrão univariado. Para cada cenário foram geradas 100 amostras e cada amostra foi analisada por meio dos quatro procedimentos (Me, LD, MV e IM). Os métodos MV e IM foram implementados com e sem variáveis auxiliares para os cenários com dados ausentes do tipo MNAR. Com isso, foram obtidas as estimativas médias dos coeficientes da regressão referentes aos quatro procedimentos em cada cenário.

FIGURA 2 – Plano do estudo de simulação



Fonte: Dados da pesquisa (elaboração dos autores).

O primeiro mecanismo utilizado gera valores ausentes completamente ao acaso (MCAR). Foram sorteadas amostras aleatórias de indivíduos com valores completos. Para esses indivíduos o valor da variável resposta passa a ser ausente. No segundo mecanismo, os dados ausentes simulados são do tipo MAR e, nesse caso, a ausência estava relacionada à variável MAT09, de tal forma que a ocorrência de dados faltantes é maior entre os alunos com menor proficiência em Matemática no primeiro ano do ensino médio. Por fim, no terceiro mecanismo foram simulados valores ausentes (MNAR) cuja taxa depende de outras variáveis não presentes no modelo apresentado na equação (1) e, nesse caso, a ausência simulada está relacionada com as variáveis LP09, IDADE, REPETÊNCIA e MANHÃ (ver Quadro 1). Nesse último

cenário o percentual de valores ausentes é maior entre os alunos com menor proficiência em Língua Portuguesa, mais velhos, com maior número de repetência e que estudam no período vespertino ou noturno.

O terceiro mecanismo descrito acima pode ser considerado MNAR, dado que existe uma dependência residual entre a ausência e a variável que apresenta os valores ausentes (MAT10). Essa dependência deve-se à relação existente entre LP09, IDADE, REPETÊNCIA e MANHÃ e a variável MAT10, considerando as demais variáveis do modelo. Tais variáveis foram escolhidas por estarem relacionadas com a probabilidade de ausência dos alunos no segundo ano do ensino médio (VINHA, 2016).

ANÁLISE DE DADOS

As análises foram realizadas utilizando o *software* estatístico SAS (*Statistical Analysis System*), versão 9.4. Algumas funções específicas foram usadas:

- para a análise dos dados completos e para o procedimento LD foi empregada apenas a PROC REG. Essa função tem como padrão o procedimento *listwise deletion* no ajuste do modelo para dados incompletos. Os modelos foram estimados pelo método dos mínimos quadrados ordinários;
- a estimação por MV foi realizada através do algoritmo EM. A inclusão das variáveis auxiliares (LP09, IDADE, REPETÊNCIA e MANHÃ) foi feita pelo método de estimação em dois estágios,[2] sendo que, para isso, utilizaram-se a opção EM da PROC MI e a PROC REG;
- para o procedimento IM foi empregada a PROC MI com os seguintes parâmetros: $m = 50$ para os cenários com 10% e 30% de valores ausentes; e $m = 100$ para os cenários com 50% de ausentes, 500 iterações iniciais antes da retirada da primeira amostra e 200 iterações entre as amostras. A PROC MIANALYZE foi usada para gerar as estimativas finais a partir das m amostras.

2 Esse método foi proposto por Savalei e Bentler (2009). No primeiro estágio são incorporadas todas as variáveis (de interesse no estudo e auxiliares) na estimação do vetor de médias e matriz de covariâncias pela máxima verossimilhança e, no segundo estágio, o modelo principal é ajustado considerando somente as variáveis de interesse no estudo.

RESULTADOS

As Tabelas 1, 2 e 3 apresentam os coeficientes estimados a partir dos dados completos e as estimativas médias relativas aos diferentes cenários de ausência de dados. Essas tabelas mostram também os desvios observados entre as estimativas médias e as estimativas obtidas com os dados completos, de acordo com a equação:

$$Desvio\,(\%) = \left(\frac{bi - bc}{bc}\right) \times 100\% \quad (2)$$

Onde *bi* e *bc* correspondem às estimativas obtidas para os dados incompletos e completos, respectivamente.

Considerando os resultados da amostra completa, pode-se observar que quanto maior é a proficiência em Matemática em 2009, maior tende a ser a proficiência em 2010 (o coeficiente estimado é de 0,677 pontos). Verifica-se que, controlado o desempenho anterior, os meninos têm nota média 3,017 pontos acima da nota média das meninas, e os que pretendem ingressar no ensino superior registram 6,995 pontos a mais, em média. Os resultados apresentados a seguir mostram as variações em relação a esses valores.

Primeiramente, observa-se que a imputação simples pela média mostra o pior desempenho, independentemente do mecanismo gerador e do percentual de ausência (Tabelas 1, 2 e 3). Mesmo quando a ausência é completamente ao acaso, constata-se, pela Tabela 1, que as estimativas obtidas por esse método são distantes dos valores encontrados com os dados completos.

Na Tabela 1 são apresentados os demais resultados para os dados ausentes do tipo MCAR. Com exceção da imputação pela média, os resultados são semelhantes, ou seja, as estimativas dos parâmetros do modelo de regressão mostram pequena variação em relação ao valor observado no ajuste para os dados completos (desvios de 4,1% ou menos). Verifica-se também que, mesmo com percentual elevado de valores ausentes, a variação observada nos coeficientes ainda é pequena. Em geral, o coeficiente relativo à variável MASC registra maiores desvios.

TABELA 1 – Estimativas dos coeficientes do modelo de regressão aplicado a dados completos e a dados com ausência de informação do tipo MCAR (simulada), acompanhados do desvio percentual, segundo o cenário de ausência e de tratamento

DADOS COMPLETOS		DADOS AUSENTES POR SIMULAÇÃO: TIPO MCAR			
		TRATAMENTO			
		MÉDIA	LD	MV	IM
10% AUSENTES					
Intercepto	85,85	**102,79 (19,7)**	85,90 (0,1)	85,88 (0,0)	85,86 (0,0)
MAT09	0,677	0,61 (-9,9)	0,68 (0,4)	0,68 (0,4)	0,68 (0,4)
MASC	3,017	**2,71 (-10,2)**	3,01 (-0,2)	3,02 (0,1)	3,00 (-0,6)
SUP	6,995	6,31 (-9,8)	7,02 (0,4)	7,02 (0,4)	7,01 (0,2)
30% AUSENTES					
Intercepto	85,85	**136,51 (59,0)**	85,92 (0,1)	85,85 (0,0)	85,72 (-0,2)
MAT09	0,677	**0,48 (-29,1)**	0,68 (0,4)	0,68 (0,4)	0,68 (0,4)
MASC	3,017	**2,08 (-31,1)**	2,99 (-0,9)	2,97 (-1,6)	2,94 (-2,6)
SUP	6,995	**4,95 (-29,2)**	7,00 (0,1)	7,01 (0,2)	7,03 (0,5)
50% AUSENTES					
Intercepto	85,85	**170,53 (98,6)**	85,89 (0,0)	85,88 (0,0)	85,62 (-0,3)
MAT09	0,677	**0,34 (-49,8)**	0,68 (0,4)	0,68 (0,4)	0,69 (1,9)
MASC	3,017	**1,51 (-50,0)**	3,14 (4,1)	3,14 (3,4)	3,12 (3,4)
SUP	6,995	**3,53 (-49,5)**	7,14 (2,1)	7,13 (1,9)	7,14 (2,1)

Fonte: Dados da pesquisa (elaboração dos autores).

Nota: Os valores em negrito do desvio percentual correspondem às estimativas com desvio maior que 10% em relação ao coeficiente da aplicação da regressão aos dados completos.

Pela Tabela 2 observa-se que os procedimentos LD, MV e IM têm desempenhos semelhantes quando os dados ausentes estão relacionados com a variável MAT09 presente no modelo. Em comparação com os cenários apresentados na Tabela 1, em geral, os dados ausentes do tipo MAR têm maior impacto na estimação dos coeficientes da regressão. Verifica-se que quanto maior é o percentual de ausência, maiores são os desvios observados, e os coeficientes mais afetados estão relacionados à variável MAT09 e ao intercepto do modelo. Considerando os coeficientes de MASC e SUP, variáveis não relacionadas diretamente com a ausência de dados, os desvios são pequenos (até 2,6%) quando o percentual de ausência é de 10% e 30%, e não ultrapassa 12,8% quando a ausência é simulada em metade das amostras.

TABELA 2 – Estimativas dos coeficientes do modelo de regressão aplicado a dados completos e a dados com ausência de informação do tipo MAR (simulada), acompanhados do desvio percentual, segundo o cenário de ausência e de tratamento

DADOS COMPLETOS		DADOS AUSENTES POR SIMULAÇÃO: TIPO MAR			
		TRATAMENTO			
		MÉDIA	LD	MV	IM
10% AUSENTES					
Intercepto	85,85	**99,0 (15,3)**	82,79 (-3,6)	82,75 (-3,6)	82,92 (-3,4)
MAT09	0,677	0,63 (-6,9)	0,689 (1,8)	0,689 (1,8)	0,688 (1,6)
MASC	3,017	3,14 (4,1)	2,974 (-1,4)	2,987 (-1,0)	2,999 (-0,6)
SUP	6,995	**5,62 (-19,7)**	7,017 (0,3)	7,019 (0,3)	7,034 (0,6)
30% AUSENTES					
Intercepto	85,85	**127,7 (48,7)**	**75,3 (-12,3)**	**75,17 (-12,4)**	**75,77 (-11,7)**
MAT09	0,677	**0,54 (-20,2)**	0,717 (5,9)	0,717 (5,9)	0,715 (5,6)
MASC	3,017	3,05 (1,1)	2,940 (-2,6)	3,008 (-0,3)	3,039 (0,7)
SUP	6,995	**3,62 (-48,2)**	7,049 (0,8)	7,074 (1,1)	7,081 (1,2)
50% AUSENTES					
Intercepto	85,85	**168,00 (95,7)**	**61,67 (-28,2)**	**61,35 (-28,5)**	**62,65 (-27,0)**
MAT09	0,677	**0,39 (-42,4)**	**0,767 (13,3)**	**0,768 (13,4)**	**0,763 (12,7)**
MASC	3,017	**2,48 (-17,8)**	**2,631 (-12,8)**	**2,711 (-10,1)**	2,796 (-7,3)
SUP	6,995	**2,03 (-71,0)**	6,937 (-0,8)	6,934 (-0,9)	6,984 (-0,2)

Fonte: Dados da pesquisa (elaboração dos autores).

Nota: Os valores em negrito do desvio percentual correspondem às estimativas com desvio maior que 10% em relação ao coeficiente da aplicação da regressão aos dados completos.

Os dados ausentes do tipo MNAR simulados neste estudo têm maior impacto nos resultados (Tabela 3). Para os procedimentos LD, MV e IM, o coeficiente da variável MASC é o mais afetado quando a ausência não é ao acaso. Nota-se que, quando não são consideradas variáveis auxiliares, os métodos MV e IM apresentam resultados praticamente iguais aos observados para o método LD.

No presente estudo, a utilização de variáveis auxiliares é importante para reduzir o impacto dos valores ausentes na estimação dos efeitos das variáveis MASC e SUP (Tabela 3). Considerando 30% e 50% de valores ausentes, verifica-se que os desvios observados para o coeficiente da variável MASC são reduzidos aproximadamente pela metade. Por exemplo, para o método MV, o desvio passa de 31,9% para 16,3%,

quando o percentual de ausentes é de 30%, e de 55,8% para 27,9%, quando 50% estão ausentes. O impacto da ausência de informação na variável SUP é menor, com desvios de 12,9% e 15,0% para os maiores percentuais de ausência, e com a utilização das variáveis auxiliares os desvios passam para 2,5% ou menos. Por outro lado, pode-se observar que a utilização de variáveis auxiliares gera maiores desvios na estimação dos coeficientes da variável MAT09 e do intercepto do modelo.

TABELA 3 – Estimativas dos coeficientes do modelo de regressão aplicado a dados completos e a dados com ausência de informação do tipo MNAR (simulada), acompanhados do desvio percentual, segundo o cenário de ausência e de tratamento

DADOS COMPLETOS		DADOS AUSENTES POR SIMULAÇÃO: TIPO MNAR					
		TRATAMENTO					
		MÉDIA	LD	MV		IM	
				SEM AUXILIARES	COM AUXILIARES	SEM AUXILIARES	COM AUXILIARES
10% AUSENTES							
Intercepto	85,85	**100,1 (16,6)**	84,3 (-1,8)	84,3 (-1,8)	83,9 (-2,3)	84,4 (-1,7)	83,9 (-2,3)
MAT09	0,677	0,63 (-6,9)	0,68 (0,4)	0,68 (0,4)	0,69 (1,9)	0,68 (0,4)	0,69 (1,9)
MASC	3,017	3,14 (4,1)	3,15 (4,4)	3,15 (4,4)	3,12 (3,4)	3,15 (4,4)	3,12 (3,4)
SUP	6,995	**5,55 (-20,7)**	6,96 (-0,5)	6,96 (-0,5)	6,97 (-0,4)	6,96 (-0,5)	6,97 (-0,4)
30% AUSENTES							
Intercepto	85,85	**131,4 (53,1)**	79,5 (-7,4)	79,5 (-7,4)	**74,3 (-13,5)**	79,4 (-7,5)	**74,7 (-13,0)**
MAT09	0,677	**0,53 (-21,7)**	0,71 (4,9)	0,71 (4,9)	0,72 (6,4)	0,71 (4,9)	0,72 (6,4)
MASC	3,017	3,00 (-0,6)	**3,98 (31,9)**	**3,98 (31,9)**	**3,51 (16,3)**	**3,99 (32,3)**	**3,57 (18,3)**
SUP	6,995	**3,40 (-51,4)**	**6,09 (-12,9)**	**6,09 (-12,9)**	6,82 (-2,5)	**6,09 (-12,9)**	6,83 (-2,4)
50% AUSENTES							
Intercepto	85,85	**164,7 (91,8)**	**75,3 (-12,3)**	**75,3 (-12,3)**	**67,2 (-21,7)**	**75,3 (-12,3)**	**67,3 (-21,6)**
MAT09	0,677	**0,40 (-40,9)**	0,73 (7,8)	0,73 (7,8)	**0,75 (10,8)**	0,73 (7,8)	**0,75 (10,8)**
MASC	3,017	**2,28 (-24,4)**	**4,70 (55,9)**	**4,70 (55,8)**	**3,86 (27,9)**	**4,70 (55,8)**	**3,86 (27,9)**
SUP	6,995	**2,62 (-62,5)**	**5,95 (-14,9)**	**5,95 (-14,9)**	7,06 (0,9)	**5,94 (-15,1)**	7,05 (0,8)

Fonte: Dados da pesquisa (elaboração dos autores).

Nota: Os valores em negrito do desvio percentual correspondem às estimativas com desvio maior que 10% em relação ao coeficiente da aplicação da regressão aos dados completos.

DISCUSSÃO

O estudo comparativo apresentado confirma resultados observados por outros autores. Por exemplo, verificou-se que a imputação pela média tem desempenho inferior na maioria das circunstâncias, o que corrobora a afirmação de McKnight *et al.* (2007) de que esse método não deveria ser utilizado pelos pesquisadores.

O desempenho dos procedimentos LD, MV e IM para análise de dados com ausência do tipo MCAR encontrado neste estudo é semelhante ao observado nos trabalhos de Enders (2001a) e de Schafer e Graham (2002). O primeiro autor relata que os coeficientes estimados para os quatro métodos avaliados (LD, MV, IM e *pairwise deletion*) apresentam pequenas variações em relação aos verdadeiros parâmetros. Já Schafer e Graham (2002) avaliaram apenas os métodos MV e IM e também verificaram pequena variação nas estimativas em relação aos parâmetros quando a ausência é completamente ao acaso.

No estudo de Enders (2001a) também foram simulados cenários com dados ausentes do tipo MAR. De forma geral, as estimativas dos coeficientes do modelo geradas pelos métodos avaliados mostram desvios relativamente pequenos. O autor afirma que o método MV apresentou melhor desempenho, porém a diferença entre os métodos é significativa apenas na estimação de um dos coeficientes do modelo. No estudo de Schafer e Graham (2002), não foram observados desvios expressivos nas estimativas dos coeficientes da regressão (para os métodos MV e MI) quando a ausência é ao acaso. Vale ressaltar que esses autores utilizaram dados simulados nas comparações.

O maior impacto dos valores ausentes do tipo MNAR também é reportado por outros autores. No estudo de Schafer e Graham (2002), os procedimentos MV e MI registraram estimativas para os coeficientes da regressão distantes dos valores verdadeiros. Em Enders (2001a), desvios expressivos foram observados em apenas uma variável e, nesse caso, o procedimento LD alcançou os melhores resultados.

Em Langkamp, Lehman e Lemeshow (2010) foram simulados dados ausentes do tipo MNAR, com percentual de

ausência variando de 10% a 40%. Nesse estudo foram utilizados dados completos extraídos de uma pesquisa da área de saúde. Os métodos LD e IM mostram desempenhos similares quando o percentual de ausentes é igual a 10%, mas, quando o percentual é maior, o IM gera estimativas mais próximas dos resultados observados para dados completos.

Uma amostra de dados completos obtidos a partir de dados reais também foi empregada no estudo de Young e Johnson (2013). Os métodos IM, MV e LD apresentam viés semelhante na estimação dos coeficientes da regressão quando os dados ausentes são do tipo MNAR, porém os autores ressaltam que os métodos IM e MV têm melhor desempenho na estimação dos erros padrão.

Croninger e Douglas (2005) usaram dados simulados baseados em variáveis presentes em uma pesquisa realizada entre ex-alunos de uma universidade. Os dados considerados nas comparações apresentam padrão arbitrário de ausência, contendo ausentes do tipo MCAR, MAR e MNAR. Os métodos MV e IM registraram resultados superiores em relação a métodos tradicionais, entre eles o LD, com menores desvios em relação aos coeficientes observados para a amostra completa.

Os benefícios da utilização de variáveis auxiliares são relatados no trabalho de Collins, Schafer e Kam (2001). Esse estudo tinha como objetivo principal avaliar o uso de variáveis auxiliares nos procedimentos MV e IM e, para isso, foram simulados diversos cenários variando os tipos de dados ausentes e a relação entre a taxa de ausência e as variáveis auxiliares. Em geral, os autores identificaram que o uso dessas variáveis é importante quando o percentual de ausentes é superior a 25% e a relação entre a variável auxiliar (usada no mecanismo gerador de valores ausentes) e a variável que apresenta os valores ausentes é forte. No presente estudo, tendo em vista que o modelo de regressão foi empregado para a comparação do desempenho entre grupos de estudantes (segundo o sexo e a intenção de ingresso no ensino superior), verificou-se também o benefício do uso de variáveis auxiliares. Apesar do aumento do viés nas estimativas do intercepto do modelo e do coeficiente da variável utilizada como controle (MAT09), observou-se a redução dos desvios

nas estimativas dos coeficientes das variáveis usadas na comparação dos grupos (MASC e SUP).

CONCLUSÕES E RECOMENDAÇÕES

A ausência de informação faz parte de praticamente toda pesquisa quantitativa e, portanto, deve-se fazer o máximo para reduzir seu impacto e assim evitar conclusões equivocadas a partir dos resultados observados. Segundo Allison (2002), os pesquisadores que pretendem mitigar os riscos associados aos dados ausentes devem escolher a estratégia de análise com cuidado, devendo considerar as características da ausência de informação. Nesse contexto, o presente estudo tem como objetivo principal contribuir para o melhor entendimento do tratamento e análise de dados incompletos.

Neste trabalho foram apresentadas algumas classificações de dados ausentes encontradas na literatura, em especial a proposta por Rubin (1976, 1987). Essa classificação é amplamente utilizada e extremamente útil na escolha do tratamento a ser empregado, no entanto outras informações devem ser analisadas. O pesquisador deve também considerar o objetivo da pesquisa e a técnica principal de análise de dados, a fonte e o padrão de ausência de informação, o tamanho da amostra, o percentual de ausência e a disponibilidade de variáveis que podem servir como auxiliares.

Dada a diversidade de situações e tipos de dados ausentes, não existe um procedimento único e infalível para a análise de dados. Inúmeras abordagens têm sido desenvolvidas ao longo dos anos e neste artigo foram apresentados resumidamente alguns dos procedimentos mais encontrados na literatura. Buscou-se aqui uma apresentação simples, logo muitos detalhes foram omitidos. Para os leitores interessados, Enders (2010) e McKnight *et al.* (2007) são boas referências, apresentando conceitos e métodos de forma acessível. Para os que buscam uma leitura mais teórica, ver Rubin (1987) e Allison (2002).

Outros procedimentos para tratamento de dados ausentes podem ser encontrados na literatura. Por exemplo, é comum na área social o uso de escalas para mensuração de

traços latentes (ansiedade, qualidade de vida, motivação) e, nesses casos, os dados ausentes de um item da escala podem ser substituídos pela média dos demais, sob a suposição de que todos representam medidas válidas do mesmo traço latente (SCHAFER; GRAHAM, 2002). Nos estudos longitudinais, a substituição pode ser feita pelo último valor observado para o mesmo indivíduo (MCKNIGHT et al., 2007). Nas ocasiões em que existem casos ausentes (um ou mais indivíduos sem resposta para todo o questionário), podem-se utilizar procedimentos de reponderação (*reweighting*), em que são calculados novos pesos amostrais para os casos presentes na amostra (SCHAFER; GRAHAM, 2002). Ainda, quando os dados ausentes são do tipo MNAR, os modelos de seleção e de mistura de padrões são muito úteis (FITZMAURICE et al., 2009).

O estudo de comparação realizado teve como foco o ajuste de modelos de regressão aplicados aos dados de uma avaliação educacional. Em geral, os desvios observados nas estimativas dos coeficientes da regressão quando 10% dos dados da variável resposta estavam ausentes foram pequenos. O impacto da ausência de dados depende do tipo de ausência (menor para MCAR e maior para MNAR) e aumenta quando o percentual de ausência é maior. A substituição simples pela média mostrou resultados insatisfatórios em todos os cenários. O procedimento *listwise deletion* apresenta resultados semelhantes aos procedimentos baseados na máxima verossimilhança e imputação múltipla nos cenários simulados, e as variáveis auxiliares foram importantes para redução dos desvios.

Os resultados do estudo comparativo aqui exposto devem ser interpretados com cautela. Optou-se pela utilização de apenas uma técnica de análise de dados na comparação dos métodos, o que limita a extrapolação dos resultados. Foram empregados modelos de regressão por se tratar de uma técnica amplamente usada na pesquisa aplicada. Em geral, os sistemas de avaliação educacional coletam diversas informações dos estudantes, escolas, professores e diretores, no entanto poucas variáveis foram utilizadas para simplificar a apresentação dos resultados.

O impacto das informações ausentes e a comparação dos métodos foram feitos apenas com as estimativas médias

dos coeficientes, porém outros aspectos poderiam ser considerados. Uma análise rigorosa deveria avaliar não somente o viés dos estimadores, mas também o erro quadrático médio[3] (EQM), estimativas dos erros padrão, cobertura de intervalos de confiança, conclusões de testes de significância e medidas de qualidade de ajuste dos modelos (COLLINS; SCHAFER; KAM, 2001; ENDERS, 2001b). Além disso, nas comparações apresentadas, os coeficientes estimados a partir dos dados completos foram tratados como parâmetros do modelo proposto. Deve-se considerar que existe incerteza associada a esses coeficientes e que eles podem ser afetados por problemas de especificação do modelo, uma vez que não foram estimados a partir de população e sim de uma amostra.

3 O EQM representa a média das diferenças entre as estimativas individuais e o parâmetro de interesse, elevadas ao quadrado. Dessa forma, tal medida considera, além do viés, a variância das estimativas (COLLINS; SCHAFER; KAM, 2001).

Uma vez que o presente estudo teve como base os dados reais de uma avaliação, com variáveis contínuas e categóricas, a suposição de normalidade multivariada não está satisfeita. No entanto, acredita-se que esse fato não comprometa os resultados obtidos, pois as estimativas dos coeficientes da regressão são pouco afetadas pela não normalidade (ALLISON, 2002).

Recomenda-se a utilização dos procedimentos baseados na máxima verossimilhança e imputação múltipla. Além de apresentar bons resultados na estimação dos coeficientes, esses procedimentos são mais apropriados para a estimação de erros padrão, o que resulta em testes de hipóteses mais confiáveis. Os procedimentos MV e IM geram resultados similares, no entanto, com o uso de pacotes estatísticos adequados, o MV pode ser mais vantajoso pela facilidade de implementação. O procedimento *listwise deletion* alcança resultados semelhantes aos procedimentos mais sofisticados quando as variáveis auxiliares não são incorporadas nas análises; logo pode ser atrativo dada a enorme simplicidade. Por fim, a imputação simples pela média deve ser evitada.

REFERÊNCIAS

ALLISON, Paul D. *Missing data*. Thousand Oaks, CA: Sage, 2002.

ANDRIDGE, Rebecca R.; LITTLE, Roderick J. A. A review of hot deck imputation for survey non-response. *International Statistical Review*, New Jersey, v. 78, n. 1, p. 40-64, 2010.

BARALDI, Amanda N.; ENDERS, Craig K. An introduction to modern missing data analyses. *Journal of School Psychology*, Amsterdam, v. 48, p. 5-37, 2010.

BUCK, S. F. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, Series B*, London, v. 22, n. 2, p. 302-306, 1960.

CEARÁ. Secretaria da Educação. *SPAECE – 2011*. Matemática, 3º ano: ensino médio. Fortaleza: SEE, UFJF, 2011. p. 1-22. (Boletim Pedagógico, v. 3).

CHEEMA, Jehanzeb R. A review of missing data handling methods in education research. *Review of Educational Research*, Thousand Oaks, CA, v. 20, n. 10, p. 1-20, 2014.

COHEN, Jacob; COHEN, Patricia. *Applied multiple regression and correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum, 1985.

COLLINS, Linda M.; SCHAFER, Joseph L.; KAM, Chi-Ming. A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods*, Washington, v. 6, n. 4, p. 330-351, 2001.

COX, Bradley E. et al. Working with missing data in higher education research: a primer and real world. *The Review of Higher Education*, Baltimore, v. 37, n. 3, p. 377-402, Spring 2014.

CRONINGER, Robert G.; DOUGLAS, Karen M. Missing data and institutional research. In: UMBACH, P. D. (Ed.). *Survey research*: emerging issues of technology, policy, and analysis. San Francisco: Wiley Interscience Periodicals, 2005. p. 33-49.

ENDERS, Craig K. The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement*, Thousand Oaks, CA, v. 61, n. 5, p. 713-740, 2001a.

ENDERS, Craig K. The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, Washington, v. 6, n. 4, p. 352-370, 2001b.

ENDERS, Craig K. *Applied missing data analysis*. New York: Guilford, 2010.

FITZMAURICE, Garret et al. *Longitudinal data analysis*. Boca Raton: Chapman & Hall, 2009.

GRAHAM, John W. Missing data analysis: making it work in the real world. *Annual Review of Psychology*, Palo Alto, CA, v. 60, p. 549-576, 2009.

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

LANGKAMP, Diane L.; LEHMAN, Amy; LEMESHOW, Stanley. Techniques for handling missing data in secondary analyses of large surveys. *Academic Pediatrics,* Amsterdam, v. 10, n. 3, p. 205-210, maio/jun. 2010.

MACEDO, Glaucia Alves. *Fatores associados ao rendimento escolar de alunos da 5ª série (2000):* uma abordagem longitudinal do valor adicionado e da heterogeneidade. 2004. 212f. Dissertação (Mestrado em Demografia) – Faculdade de Ciências Econômicas, Universidade Federal de Minas Gerais, Belo Horizonte, 2004.

MCKNIGHT, Patrick E. et al. *Missing data*: a gentle introduction. New York: Guilford, 2007.

OLIVEIRA, Pedro Rodrigues; BELLUZZO, Walter; PAZELLO, Elaine Toldo. The public–private test score gap in Brazil. *Economics of Education Review*, Amsterdam, v. 35, p. 120-133, 2013.

PEUGH, James L.; ENDERS, Craig K. Missing data in educational research: a review of reporting practices and suggestions for improvement. *Review of Educational Research*, Thousand Oaks, CA, v. 74, n. 4, p. 525-556, Winter 2004.

RODRIGUES, Clarissa Guimarães; RIOS-NETO, Eduardo Luiz Gonçalves; PINTO, Cristine Campos de Xavier. Diferenças intertemporais na média e distribuição do desempenho escolar no Brasil: o papel do nível socioeconômico, 1997-2005. *Revista Brasileira de Estudos de População*, Belo Horizonte, v. 28, n. 1, p. 5-36, jan./jun. 2011.

ROSE, Roderick A.; FRASER, Mark W. A simplified framework for using multiple imputation in social work research. *Social Work Research*, Oxford, v. 32, n. 3, p. 171-178, 2008.

ROUSSEAU, Michel et al. Reporting missing data: a study of selected articles published from 2003-2007. *Quality & Quantity*, Berlin, v. 46, n. 5, p. 1393-1406, 2012.

RUBIN, Donald B. Inference and missing data. *Biometrika*, Oxford, v. 63, n. 3, p. 581-592, 1976.

RUBIN, Donald B. *Multiple imputation for nonresponse in surveys.* New York: Wiley, 1987.

SAVALEI, Victoria; BENTLER, Peter M. A two-stage approach to missing data: theory and application to auxiliary variables. *Structural Equation Modeling*, London, v. 16, n. 3, p. 477-497, 2009.

SCHAFER, Joseph L.; GRAHAM, John W. Missing data: our view of the state of the art. *Psychological Methods*, Washington, v. 7, n. 2, p. 147-177, 2002.

SOARES, José Francisco; ALVES, Maria Teresa Gonzaga. Desigualdades raciais no sistema brasileiro de educação básica. *Educação e Pesquisa*, São Paulo, v. 29, n. 1, p. 147-165, jan./jun. 2003.

VINHA, Luís Gustavo do Amaral. *Estudos longitudinais e tratamento de dados ausentes em avaliações educacionais.* 2016. 124f. Tese (Doutorado em Psicologia Social, do Trabalho e das Organizações) – Instituto de Psicologia, Universidade de Brasília, Brasília, DF, 2016.

WILKS, S. S. Moments and distributions of estimates of population parameters from fragmentary samples. *The Annals of Mathematical Statistics*, New York, v. 3, p. 163-195, 1932.

XERXENEVSKY, Lauren Lewis. *Programa Mais Educação*: avaliação do impacto da educação integral no desempenho de alunos no Rio Grande do Sul. 2012. 143f. Dissertação (Mestrado em Economia do Desenvolvimento) – Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2012.

YOUNG, Rebekah; JOHNSON, David. Methods for handling missing secondary respondent data. *Journal of Marriage and Family*, New Jersey, v. 75, n. 1, p. 221-234, 2013.

LUIS GUSTAVO DO AMARAL VINHA
Professor adjunto do Departamento de Estatística da Universidade de Brasília (UnB), Brasília, Distrito Federal, Brasil
lgvinha@gmail.com

JACOB ARIE LAROS
Professor associado do Instituto de Psicologia da Universidade de Brasília (UnB), Brasília, Distrito Federal, Brasil
jalaros@gmail.com

Recebido em: 09 MAIO 2017
Aprovado para publicação em: 11 OUTUBRO 2017