XXIII SEMEAD
Seminários em Administração

novembro de 2020
ISSN 2177-3866

Avaliação de métodos de tratamento de dados ausentes em contextos de mediação

SAULO BARROS DE MELO
UNIVERSIDADE DE BRASÍLIA (UNB)

ELAINE RABELO NEIVA
UNIVERSIDADE DE BRASÍLIA (UNB)

Agradecimento à orgão de fomento:
FAP- DF - Fundação de Apoio à Pesquisa do Distrito Federal

Avaliação de métodos de tratamento de dados ausentes em contextos de mediação

1 INTRODUÇÃO

A problemática da existência de dados ausentes se mostra como questão onipresente na pesquisa científica. Neste sentido, cada vez mais se observa a criação e refinação de métodos de tratamentos de dados ausentes, sendo fundamental compreender as circunstâncias para executar a melhor metodologia possível (Schouten, Lugtig, Vink, 2018).

A literatura indica que não confrontar esse problema pode levar estudos a resultados inválidos (Peeters et. al, 2015). Apesar disso, todos métodos de tratamento de valores ausentes podem vir a trazer vieses às pesquisas, dessa forma, faz-se necessário encarar essa recorrente problemática não como questão marginal à pesquisa em percurso, mas sim como fator estruturante desta.

Embora a avaliação de valores ausentes não seja uma questão recente na literatura (Kuijk et al, 2016), observa-se que o avanço recente da capacidade computacional e popularização de softwares livres estão tornando métodos de imputação, antes considerados complexos, mais acessíveis.

Apesar disso, neste contexto, observa-se que muitos pesquisadores ainda utilizam de métodos tidos como menos refinados ao se depararem com esta problemática em suas pesquisas. Dessa forma, analisar comparativamente vieses, a fim de evitar utilização metodológica menos favorável a determinado contexto, se mostra essencial para o bom desempenho de estudos.

Apesar da referida necessidade de análise, observa-se contextos nos quais são raros estudos que se associam especificamente à temática de tratamento de valores ausentes. Neste sentido destaca-se a análise de mediação. A mediação consiste em um método estatístico de avaliação de relações causais e é amplamente utilizado em pesquisas em Administração e nos mais diversos campos científicos, apesar disso, estudos direcionados a discutir mediação e valores ausentes são raros (Zhang, Wang, 2013).

Dessa forma, a pesquisa avalia comparativamente três métodos de tratamento de dados ausentes (imputação por média, análise de casos completos e imputação múltipla) sob diversos parâmetros dentro do contexto de mediação simples, a fim de geração de diferentes cenários para análise. Os parâmetros avaliados pelo estudo foram a proporção de valores ausentes, o tamanho da amostra e os três tipos de geração dados ausentes propostos por Rubin (1976), gerando múltiplos cenários submetidos ao método Monte Carlo para posterior análise comparativa de vieses.

Embora seja evidente que a falta de tratamentos de valores ausentes possa levar a resultados incorretos, muitos pesquisadores ainda ignoram esta relevante questão (Peeters et al, 2015). Neste sentido, e levando em conta a situação da análise mediação no contexto da pesquisa de valores ausentes, propõe-se a seguinte questão de pesquisa: em que medida os a proporção de dados ausentes e tamanho da amostra afetam os vieses de três métodos de tratamentos de dados ausentes em contextos de mediação simples?

Diante dos pontos apresentados, entende-se como objetivo do estudo avaliar o desempenho dos métodos de tratamento de dados ausentes dentro dos cenários propostos. Dessa forma, busca-se, também, compreender a relação dos parâmetros com os vieses obtidos nos múltiplos cenários, podendo, dessa forma, averiguar o impacto destes nos resultados gerais.

2 REFERENCIAL TEÓRICO

2.1 TIPOS DE DADOS AUSENTES

Os tipos de valores faltantes, conforme a classificação predominante na literatura, podem ser gerados a partir de três mecanismos (Rubin, 1976); valores ausentes completamente ao acaso (MCAR), valores ausentes ao acaso (MAR) e valores ausentes não ao acaso (MNAR).

Uma vez que a probabilidade de falta para todos os casos de um determinado conjunto de dados é a mesma, pode-se configurar o mecanismo de ausência como MCAR. Ou seja, nestas situações a ausência não estabelece relações com outros elementos do conjunto.

O método de geração de valores ausentes MAR configura-se em situações quando as probabilidades de ausência de casos de uma determinada variável estão condicionadas a valores de outras variáveis.

Por fim, caso o método de geração de dados não se configure como MCAR ou MAR, configura-se como MNAR. Nestes casos, a probabilidade de ausência de casos está condicionada a valores da própria variável avaliada.

2.2 MÉTODOS DE IMPUTAÇÃO

2.2.1 Imputação pela média

O método de imputação pela média consiste na simples substituição dos valores ausentes pela média da variável do caso em questão. Entende-se o método como um procedimento de baixa complexidade.

Infere-se que a simplicidade desta metodologia é um dos fatores responsáveis para ainda ser observado seu emprego por pesquisadores, uma vez que o método é tido como possivelmente o pior tipo de tratamento de dados ausentes (Enders, 2010).

2.2.2 Casos completos

Dentre os vários métodos de tratamentos de valores ausentes, destaca-se o método de casos completos. A utilização desta metodologia se faz a partir da retirada das observações nas quais observa-se valores ausentes.

O referido método se mostra vastamente utilizado em estudos, uma vez que faz parte de procedimentos tomados como padrão em diversos softwares de análise, como R, SAS e SPSS. Tal metodologia reduz a amostra, o poder estatístico e se mostra mais apropriado em cenários de baixo percentual de valores ausentes, tendo como vantagem a economia de tempo por parte dos pesquisadores (Saunders et al, 2006).

2.2.3 Imputação múltipla

O método de imputação múltipla consiste na criação de m conjuntos de dados completos objetivando a mensuração das estimativas de parâmetros de cada um destes para posterior cálculo da estima geral a partir da média aritmética dessas m estimativas. Nestes casos, os valores ausentes podem ser estimados a partir de métodos como regressão logarítmica, correspondência média preditiva e outros.

Rubin (1987) apresentou as bases metodológicas e estatísticas do método de imputação desenvolvido por ele e desde então o método tem sido visto de forma geral pela literatura como a melhor para tratamento de dados ausentes em diversos cenários (Buuren, 2018).

2.3 MEDIAÇÃO

A análise de mediação consiste em um método estatístico para avaliação de relações causais. Esta avaliação busca compreender a relação como uma variável independente X afeta uma variável dependente Y. Neste caso se é inserida uma variável mediadora M na relação apresentada a fim de mensurações de parâmetros causais desta.

Uma variável mediadora é aquela que possui um papel interveniente, também chamada de mediador, este pode ser entendido como o mecanismo pela qual X influencia Y. Ou seja, a variação da variável independente gera alteração no mediador que gera variação na variável dependente (Hayes, 2017). Um modelo de mediação simples pode ser representado por meio das seguintes equações:

(1)

$$\boldsymbol{M = i_1 + aX + e_M}$$

(2)

$$\boldsymbol{Y = i_2 + c'X + bM + e_y}$$

Sendo i_1 e i_2 os interceptos da regressão, e_M e e_y os erros ao estimar M e Y, respectivamente, e a, b, e c' são os coeficientes de regressão dadas as variáveis antecedentes do modelo.

Ou seja, um modelo de mediação busca avaliar a relação entre a variável dependente e independente a partir de dois diferentes caminhos; direto e o moderado para posteriormente se mensurar o efeito direto, indireto e total da relação para o estabelecimento de relações causais.

O referido tipo análise está presente em diversos campos da ciência com múltiplas aplicações e assim como outros tipos de análise, a mediação está sujeita à onipresente problemática de valores ausentes, mesmo assim são raros estudos que direcionados a essas duas questões juntas (Zhang, Wang, 2013).

3 MÉTODOS E TÉCNICAS DE PESQUISA

3.1 TIPOLOGIA E DESCRIÇÃO GERAL DOS MÉTODOS DE PESQUISA

O presente estudo configura-se como de abordagem quantitativa. A pesquisa busca por meio de técnicas de simulação estatísticas computacionais, no software livre R, alcançar seus objetivos.

As bases de dados alvos de análise foram geradas por meio de simulações por método de Monte Carlo, conforme especificações de cada cenário e foram integralmente elaborados pelo pesquisador.

A metodologia geral, pode ser resumida, de forma simplificada, a partir da seguinte esquematização composta por 6 etapas:

1- Definição dos parâmetros do modelo estrutural a ser avaliado.
2- Geração do conjunto original de dados variando conforme especificidades propostas.
3- Amputação de dados conforme proporção e tipo de dados ausentes.
4- Tratamento de dados ausentes conforme métodos escolhidos.
5- Análise de mediação com bases tratadas.
6- Análise comparativa de resultados dos conjuntos amputados com os originais afim de estimação de vises de cada método de tratamento em cada cenário.

3.2 PROCEDIMENTOS DE GERAÇÃO E ANÁLISE DE DADOS

Os procedimentos para geração de amostras do estudo consistem, em síntese, na elaboração do que se pode chamar de conjunto de dados originais para posterior amputação e, por fim, tratamento de valores ausentes desses. Ou seja, tais conjuntos representam bases de dados de características especificadas a priori pelo pesquisador que após terem passado pelos processos de amputação e imputação formam-se cenários nos quais serão avaliados comparativamente e após análise de mediação se é possível estimar vieses de cada situação nos três tipos diferentes de tratamento de valores ausentes.

Neste sentido, a figura abaixo representa os critérios variantes utilizados na pesquisa para formulação dos múltiplos cenários:

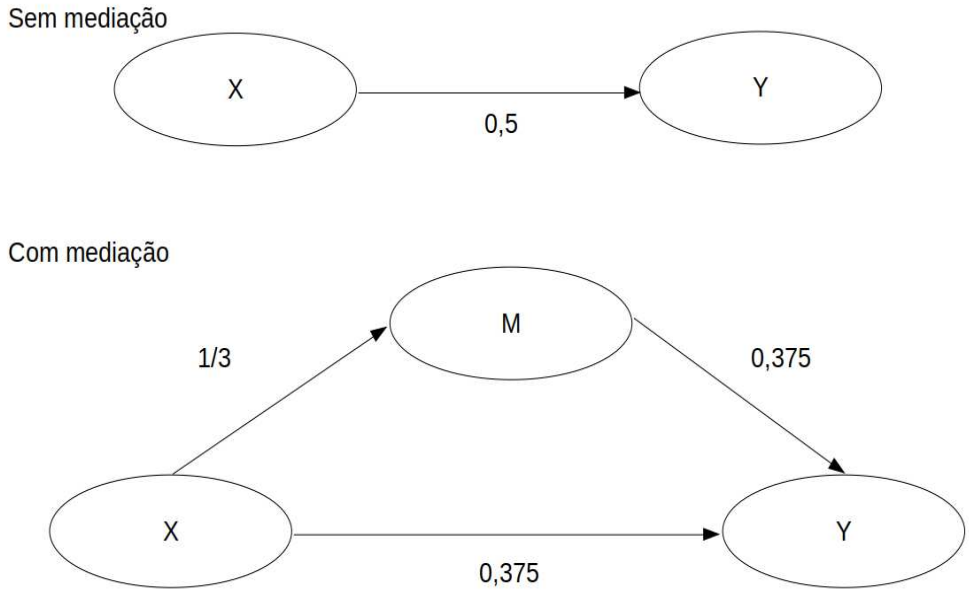
Figura 1 - Critérios para formulações de cenários

Tipo de valor ausente	MCAR	MNAR	MAR		
Tamanho da amostra	100	300	1000		
Proporção de ausentes	0,05	0,1	0,2	0,3	0,5

Fonte: Os autores.

Todos os cenários elaborados foram executados dentro de um contexto de mediação. A relação mediadora, comumente utilizada em modelos de equações estruturais, neste presente estudo foi elaborada a partir de três variáveis. As variáveis formuladas na pesquisa podem ser representadas por X, Y e M correspondendo, respectivamente pela variável independente, dependente e mediadora. O modelo e os coeficientes, definidos pelos pesquisadores, nos conjuntos de dados originais podem ser observados a partir da ilustração abaixo:

Figura 2 - Modelo proposto



Fonte: Os autores.

Tendo definido tais características do modelo, buscou-se operacionalizá-lo por meio do software R. O estudo utilizou-se dos pacotes MASS (Venables, Ripley, 2002) , Hmisc (Harrell Jr, 2020), mice (Buuren, Groothuis-Oudshoorn, 2011) e lavaan (Rosseel, 2012).

De forma inicial, tendo definido os parâmetros primários, observa-se a necessidade de elaboração do banco de dados original. O procedimento foi executado a partir da função mvnorm() do pacote MASS (Venables, Ripley, 2002), tal comando permite elaboração de simulações amostrais multivariadas de distribuição normal.

A função referida foi associada a *loops*, gerando 1000 amostras de distribuição normal, formada por 100, 300 e 1000 observações tendo três variáveis de média 0 relacionadas por meio da seguinte matriz de covariância:

Figura 3 - Matriz de variância e covariância

	X	M	Y
X	1	1/3	1/2
M	1/3	1	1/2
Y	1/2	1/2	1

Fonte: Os autores.

Tendo gerado as amostras deve-se amputar as bases conforme os parâmetros preestabelecidos. O comando ampute() de mice (Buuren, Groothuis-Oudshoorn, 2011) gerou a amputação nas 1000 amostras de cada conjunto originais criados anteriormente para geração de cada cenário. Ou seja, é por meio dessa função que se configura a proporção e tipos dos dados ausente.

Vale ressaltar que a grande quantidade de simulações amostrais citada acima (n=1000), característica do método de Monte Carlo, tem por objetivo a obtenção de coeficientes estáveis no modelo.

Além disso, é importante destacar questões específicas de cada método de amputação elaborado. Em todos os cenários os valores ausentes dos conjuntos se encontravam ou na variável independente ou na moderadora, preservando, assim, a integridade da variável dependente.

No caso dos conjuntos gerados por MCAR, conforme o conceito do método sugere, os dados foram gerados de forma totalmente aleatória. Já pelo mecanismo de MNAR, foram gerados dados condicionados pelo próprio valor, os quais quanto maior forem, maior a possibilidade de serem amputados. Por fim, neste sentido, as ausências geradas nos conjuntos de dados gerados por MAR, estavam condicionadas ao valor da variável dependente, ou seja, nestes cenários, quando maior o valor da variável Y, maior as chances dos valores das variáveis X e M serem amputadas.

Tendo feito as amputações buscou-se tratar as bases com dados ausentes para posterior análise de mediação. O modelo de mediação simples utilizado foi operacionalizado no software R por meio do pacote lavaan (Rosseel, 2012) de forma padrão.

O presente estudo apresenta três métodos de tratamentos de valores faltantes. De forma inicial, a imputação por média foi a gerada a partir de funções básicas do software R por meio da substituição dos valores ausentes pela média dos valores existente da variável específica.

Nas situações de análise de casos completos não foi necessário nenhum procedimento específico para retirada casos com ausências. Isso se deve pois o pacote que executa a análise de mediação já analisa por padrão apenas as observações com todos valores completos.

Já o método de imputação múltipla se fez a partir da função mice() do pacote de mesmo nome. Tal comando foi utilizado conforme seus parâmetros padrões, ou seja, o método de imputação foi o de *Predictive Mean Matching* (PMM) com configuração de cinco imputações múltiplas.

Por fim, objetivando a compreensão do fenômeno faz-se necessário a avaliação dos vieses nos cenários criados sob uma ótica comparativa que leve em consideração a multiplicidade de proporções de dados ausentes a fim de observar os diferentes comportamentos em cada uma das análises.

A análise de vieses foi esquematizada a partir de representações gráficas elaboradas por meio da avaliação comparativa das estimativas dos coeficientes das bases tratadas com as originais. O cálculo do viés pode ser explicitado a partir da seguinte expressão:

(3)

$$Viés = 100 \times \left(\frac{1}{1000} \sum_{i=1}^{1000} \left(\frac{bi - bc}{bc} \right) \right)$$

Na equação acima bi corresponde ao coeficiente de um caminho do conjunto de dados tratados após amputação e bc a estimativa deste mesmo caminho, mas retirada a partir dados originais completos. Além disso, o viés foi multiplicado por 100 para adequação de escala. Neste caso, o conjunto de dados originais possui coeficiente de 0,375 e 0,125 para o efeito direto e indireto respectivamente, dessa forma quanto mais os coeficientes se distanciar destes valores, maior é o viés obtido pelo conjunto tratado que possuía valores ausentes.

4 RESULTADOS

As estimativas dos coeficientes de regressão obtidos nas simulações do presente estudo expõem a existência de vieses em diferentes cenários. Neste sentido, observou-se também a existência desses desvios em relação à estimativa dos conjuntos completos em todos os métodos de tratamento de dados ausentes.

A figura 4 apresenta um compilado de estimativas de regressões das simulações executadas dentro do modelo proposto pelo estudo sob uma amostra de tamanho n= 300 e variância de valor 1. As representações gráficas da esquerda apresentam os coeficientes referentes ao efeito direto, já as da direita, representam as estimativas de coeficiente do caminho que corresponde ao efeito indireto do modelo de mediação. Além disso a compilação de dados apresenta diferenciação por método de amputação. Os gráficos da primeira, segunda e terceira linha corresponde, respectivamente, aos métodos MCAR, MNAR, MAR.

As ilustrações gráficas da figura 4 indicam as estimativas de coeficientes sob diferentes proporções de dados ausentes nos três diferentes métodos de tratamentos avaliados pela pesquisa. Nos gráficos abaixo, as linhas vermelhas, verdes e azuis representam, respectivamente, por estimativas de coeficientes retirados a partir de conjunto de dados amputados tratados pelo método de caso completo (CC), imputação múltipla (MI) imputação por média (ME). Além disso as linhas tracejadas em preto nos gráficos indicam os coeficientes dos conjuntos originais.

O primeiro gráfico expõe um viés expressivo na imputação pelo método de imputação pela média, neste caso a partir de 30% de ausência se observou uma diferença significativa

($p < 0,05$) com os demais métodos. Neste sentido, os métodos de imputação múltipla e por análise de casos completos performaram de forma mais satisfatória em todos os cenários, não havendo diferença significativa ($p < 0,05$) entre estes, nos coeficientes de efeito direto sob todas proporções de amputação.

Já no gráfico B, observa-se também um maior desvio do método de imputação por média, porém, neste caso com um viés que subestimou o coeficiente, nesta situação, o método se diferenciou significativamente ($p < 0,05$) dos demais a partir da proporção de 20% de valores ausentes.

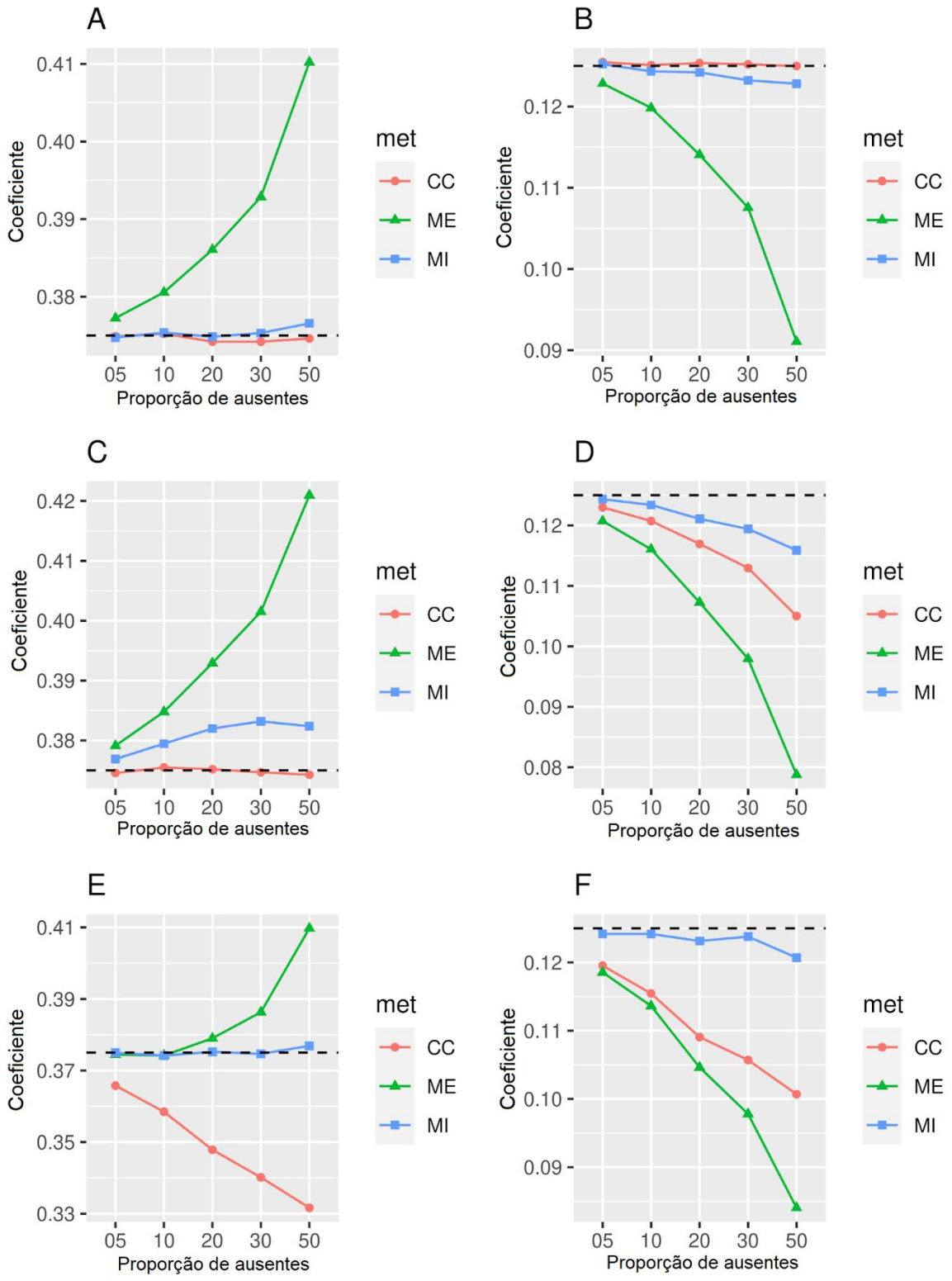
As estimativas dos coeficientes do efeito direto em cenários amputados pelo método MNAR podem ser vistas por meio do gráfico C. A ilustração indica um maior desvio do método de imputação pela média e possui diferença significativa ($p < 0,05$) com CC a partir dos 20% de amputação e 30% com MI. Nesta avaliação o método de análise de casos completos se performou com menores desvios.

Os coeficientes do efeito indireto dos cenários amputados a partir do método MNAR podem ser vistos no gráfico D. Observa-se, neste caso, uma performance com menor viés da imputação múltipla em comparação aos outros métodos em todas proporções de dados ausentes. Nesta avaliação, se é observado uma diferença significativa ($p < 0,05$) entre os todos os métodos desde 20% de valores amputados.

A avaliação dos desvios dos coeficientes do efeito direto dos casos amputados pelo método de MAR pode ser vista a partir da ilustração gráfica E. Neste caso, os coeficientes retirados a partir da imputação múltipla apresentaram baixos desvios. A avaliação CC obteve expressivos desvios sendo significantemente diferentes das estimativas dos outros métodos a partir de 10% de amputação ($p < 0,05$).

A análise dos coeficientes dos efeitos indiretos nos cenários amputados por MAR, visto no gráfico F, expõe uma evidente superioridade da imputação múltipla nestes casos. Os coeficientes de bases imputadas por MI são significativamente diferentes ($p < 0,05$) sob todas proporções de ausência a partir de 10%, o método performou melhor de forma significativa. Além disso, a análise de casos completos apresentou resultados similares à imputação pela média, com exceção dos cenários com 30% e 50% de amputação, nestes a diferença entre métodos fora dada como significativamente diferente ($p < 0,05$).

Figura 4 - Coeficientes por proporção de ausentes



Fonte: Os autores.

As estimativas obtidas a partir das bases amputadas por MCAR demonstram que a análise de casos completos e por múltipla imputação trouxeram resultados adequados, ou seja, sem vieses significativos em todos os cenários propostos. Já a imputação por média potencializou o efeito direto em até 9,39% e reduziu o efeito expressivamente as estimativas

de coeficientes indiretos, neste caso no cenário de 50% de amputação, obteve-se 27,16% de desvio.

Nos casos amputados por MNAR apenas a análise por casos completos obteve resultados sem vieses significativos ($p < 0,05$) no caso do efeito direto, porém ao se avaliar o coeficiente do efeito indireto observa-se um comportamento diferente. Neste caso a imputação múltipla volta se apresentar como método de tratamento com menor desvio. A título de ilustração, observa-se uma redução em 36,99% e 15,98% das estimativas dos coeficientes de conjuntos com 50% de amputação tratados, respectivamente por ME e CC enquanto os conjuntos imputados por MI sofreram uma redução de 7,45% quando se comparado com o conjunto de dados original.

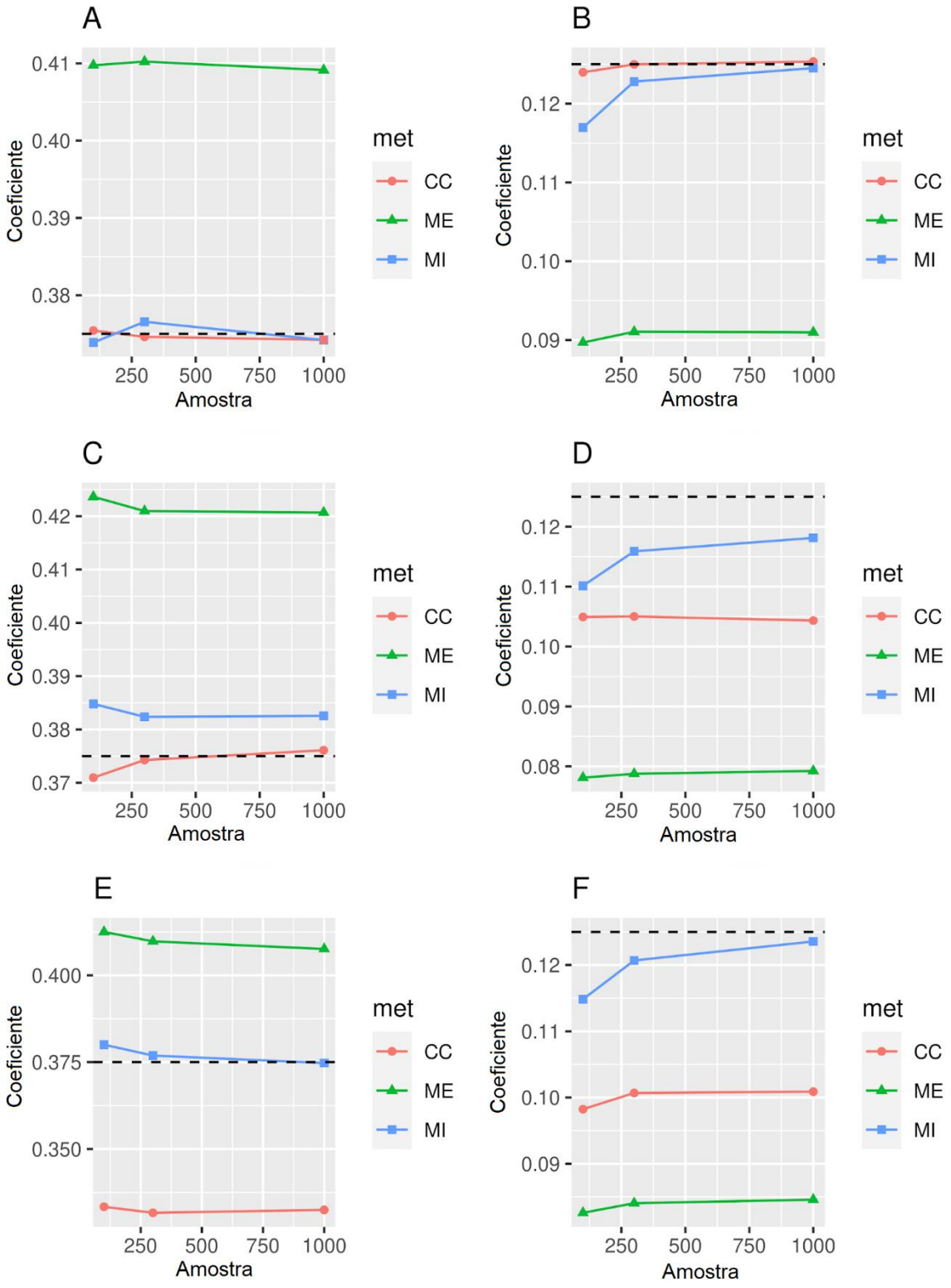
Em todos os cenários amputados por MAR a imputação múltipla se destacou como melhor método para tratamento. Na avaliação do efeito indireto, observou-se as estimativas retiradas de MI sem vieses significativos ($p < 0,05$) em todos casos. A imputação pela média e análise de casos completos obtiveram tendência de vieses distintas, ME potencializou as estimativas e CC subestimou os coeficientes. A análise do efeito indireto expões estimativas de coeficientes menores que as omitidas pelos dados originais, neste caso, a imputação múltipla obteve resultados mais satisfatórios. Enquanto ME e CC apresentaram, respectivamente, coeficientes 32,76% e 19,45% menores, IM subestimou, em média, apenas 3,45% do coeficiente do efeito indireto no cenário com 50% de amputação.

Vale ressaltar novamente que a avaliação acima se fez apenas a partir de cenários obtidos a partir de uma geração amostra n=300. Dessa forma, faz se necessário ainda a análise comparativa levando em consideração os três tamanhos amostrais por este estudo avaliado.

As representações gráficas da figura 5 possuem a mesma estrutura da figura 4 porém avalia comparativamente o tamanho da amostra e não a proporção de valores ausentes. Os gráficos apresentam as médias das estimativas de coeficientes do efeito direto, indireto sob os três métodos de amputação em um contexto de geração de 50% de valores ausentes.

A avaliação dos efeitos do tamanho da amostra no efeito direto, se apresentam com baixa variação, observando, de forma geral, uma tendência de aumento de performance diretamente proporcional ao tamanho da amostra. Já nos casos relativos aos efeitos indiretos, observou-se uma maior variação conforme eventual mudança amostral. Neste sentido destaca-se a performance do método de imputação múltipla, a partir dos gráficos é possível ver uma tendência mais forte de aproximação do coeficiente ao valor obtido pelas bases originais. A título de ilustração, o viés obtido em bases geradas por MCAR com 50% de amputação, por MI com a amostra de n=100 indicou um desvio médio negativo de 6,44% enquanto o de amostra dez vezes maior, sob o mesmo cenário, gerou desvio negativo de apenas 0,39%.

Figura 5 - Coeficientes por tamanho da amostra



Fonte: Os autores.

5 CONCLUSÃO E RECOMENDAÇÃO

A presença de valores ausentes, como visto, é um fenômeno que não distingue o

campo de estudo e está presente de forma onipresente no fazer científico. Esta questão pode ser vista como uma problemática inevitável a qual pesquisadores terão que lidar, sendo essencial seu entendimento a fim de evitar eventual desajustes em pesquisas que podem levar até mesmo ao seu comprometimento total.

A pesquisa não teve por objetivo servir como imperativo, impondo qual método de tratamento de dados ausentes deve ser usado, apesar disso, o estudo desencoraja a utilização da imputação por média, uma vez se foram obtidos expressivos vieses sob a maioria dos cenários. Neste sentido, resta-se a análise por casos completos e imputação múltipla como opções potencialmente úteis para obtenção adequada de estimadores.

Os resultados obtidos nos cenários amputados por MCAR indicam a geração de estimativas adequadas sob imputação múltipla e análise de dados completos em todas as proporções, dessa forma recomenda-se ambos métodos para situações análogas.

A avaliação dos resultados fornecidos pelos cenários amputados por MNAR, expõe um caso onde a análise de casos completos apresenta estimativas mais adequadas para mensuração do efeito direto, em contrapartida o efeito indireto apresenta grandes desvios, sugerindo-se, assim, que a imputação múltipla seja o método mais adequado para situações de mediação como as dos cenários propostos.

Já a análise dos resultados dos casos de MAR sugere uma expressiva vantagem na utilização da imputação múltipla, dado que a análise de casos completos e imputação por média forneceram estimativas do efeito indireto subestimadas de forma expressiva.

O presente estudo se mostra relevante na medida em que amplia o entendimento a respeito dos efeitos de dados ausentes em contextos de mediação. Isso se deve, pois, a pesquisa além de apresentar uma multiplicidade de cenários, mensura a o tratamento de imputação por média, tal mensuração não é observada em avaliações comparativa sob contextos de mediação como os feitos por Zhang e Wang (2013). Já em termos práticos, a pesquisa auxilia na tomada de decisões de pesquisadores, a fim de que esses tomem decisões mais acertadas em suas pesquisas ao se depararem com a problemática apresentada.

Por fim, a infinita possibilidade de cenários no fazer científico mostra-se fator limitante para uma compreensão integralmente precisa de todos eventuais casos. Ou seja, o presente estudo não me mostra capaz de estipular com precisão todos contextos, mas sim trazer uma perspectiva geral a partir dos cenários preestabelecidos. Neste sentido, situações onde se observa mais de um tipo de geração de valores ausentes não foram estipulados nesta pesquisa, recomendando-se, assim, futuros estudos que se levem em consideração esta referida questão.

REFERÊNCIAS

BUUREN, S. V., **Flexible Imputation of Missing Data**, 2a Ed. New York: Chapman and Hall/CRC, 2018.

BUUREN, S. V.; GROOTHUIS-OUDSHOORN, K. mice: Multivariate Imputation by Chained Equations in R. **Journal of Statistical Software**, v.45 n.3, p. 1-67, 2011.

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

HARRELL JR, F. E.; DUPONT, C**. Hmisc: Harrell Miscellaneous. R package version 4.4-0**, 2020, Disponível em: <https://CRAN.R-project.org/package=Hmisc> Acesso em: 11 jun. 2020.

HAYES, A. F. **Introduction to mediation, moderation, and conditional process analysis:**

A regression based approach. 2a ed. New York, The Guilford Press, 2017.

LITTLE, R.J.A.; RUBIN, D.B. **Statistical analysis with missing data**. 2a ed. New York, Wiley-Interscience, 2002.

PEETERS, M.; ZONDERVAN-ZWIJNENBURG, M.; VINK, G.; SCHOOT R. V., How to handle missing data: A comparison of diferent approaches, **European Journal of Developmental Psychology**, v.12, n.4, p. 377-394, 2015.

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

SANDER MJ VAN KUIJK, S. M. J. V.; VIECHTBAUER, W.; PEETERS, L. L.; SMITS, L. Bias in regression coefficient estimates when assumptions for handling missing data are violated: a simulation study. **Epidemiology Biostatistics and Public Health**. v.13, n.1, 2016.

SAUNDERS, J. A.; MORROW-HOWELL, N.; SPITZNAGEL, E.; DORÉ, P.; PROCTOR, E. K.; PESCARINO, R. Imputing Missing Data: A Comparison of Methods for Social Work Researchers, ***Social Work Research***, v.30, n.1, p.19–31, 2006

SCHOUTEN, R. M.; LUGTIG P.; VINK, G. Generating missing values for simulation purposes: a multivariate amputation procedure, **Journal of Statistical Computation and Simulation**, v.88, n.15, p.2909-2930, 2018.

VENABLES, W. N.; Ripley, B. D. **Modern Applied Statistics with S**. 4a ed. Springer, New York, 2002.

YVES ROSSEEL . lavaan: An R Package for Structural Equation Modeling. **Journal of Statistical Software**, v.48, n.2, p. 1-36, 2012.

ZHANG Z; WANG L. Methods for mediation analysis with missing data**. Psychometrika**, v.78, n.1, p. 154-184, 2013.