Universidade de Brasília
Departamento de Estatística

Técnicas de Imputação para variáveis quantitativas

Sabrina Lopes França

Orientador: Eduardo Yoshio Nakano

Resumo

Este estudo pesquisa e compara dois métodos de Imputação Única que utilizam dois conceitos muito importantes e disseminados na estatística, a média e a regressão linear. Através de simulações em R, foi obtido resultados em que a imputação por regressão linear simples teve melhor performance para diferentes cenários de amostras.

1. Introdução

Atualmente, com os avanços tecnológicos, é possível a coleta e o armazenamento de grandes volumes de dados utilizados para a realização de pesquisas, aplicação em negócios, desenvolvimento de inteligências artificiais entre outros. Porém, é comum ter que lidar com dados que tenham valores ausentes em sua composição. Isso pode decorrer de vários fatores como falhas humanas no processo de coleta de dados, defeitos operacionais em equipamentos ou ferramentas (RIBEIRO, 2015), alto custo na coleta de dados (MYRTVEIT et. al., 2001) e até mesmo em decorrência de participantes de pesquisa que optam por não responder algum item de um questionário (ALLISON, 2001).

Além de ser uma limitação na análise descritiva dos dados, a problemática dos dados faltantes se revela também na incerteza dos resultados (CARVALHO, 2017). Isso porque a análise de dados incompletos pode gerar vieses de seleção, principalmente se os indivíduos que respondem e estão na análise são sistematicamnete diferentes daqueles que optaram pela não resposta (ERCOLE et al., 2010). Ademais, lidar com dados incompletos produz

menor eficiência nas estimativas, já que a amostra é reduzida (NUNES, 2007). Sendo assim, uma análise incompleta pode causar dependênciass que impactam os estudo de características amostrais e na tomada de decisões (CARVALHO, 2017).

Visando solucionar essa questão, desde os anos da década de 70 têm sido desenvolvidas técnicas estatísticas cujo objetivo é substituir os valores ausentes por dados que façam sentido (RUBIN, 1976) . Na estatística, a técnica de substituição de dados ausentes denomina-se imputação de dados. Atualmente, existem várias formas de imputação: imputação única, imputação múltipla, imputação por último caso observado, entre outras (CARVALHO,2017).

Na imputação única, tem se que os dados ausentes são preenchidos uma única vez para assim se obter o banco de dados completo. Dentro desse escopo, alguns dos métodos existentes são a imputação pela média e a imputação por regressão linear. Dessa forma, esse estudo se propõe a analisar e comparar esses dois tipos de técnicas na imputação única, analisando dados simulados com os referentes métodos aplicados.

2. Fundamentos Teóricos

2.1. Média aritmética

O conceito da média aritmética ($\bar{X}$) é bastante conhecido. Ela é uma medida de posição que se define pela soma das observações divididas pelo número total de observações (n) (MORETTIN, 2013). A média é expressa da seguinte forma:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

onde X_i é o valor da i-ésima observação, i = 1,2,...,n.

2.2. Modelo de regressão linear

Um modelo de regressão consiste em um modelo matemático que retrata a relação entre duas ou mais variáveis quantitativas. Quando essa relação se trata de apenas duas variáveis (X e Y, por exemplo) e o modelo que representa a relação for uma reta, então ele é denominado de regressão linear simples (MARTINS, 2019).

Por meio do modelo de regressão linear simples, é possível observar a relação linear entre uma variável explicativa (independente) X e uma variável resposta (dependente) Y. O modelo é representado através da seguinte equação:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

onde:

- y_i é o da valor variável resposta a ser predita, Y, na observação $i = 1, ..., n$;
- x_i representa o valor da variável explicativa, X, na observação $i = 1, ..., n$;
- ϵ_i, com $i = 1, ..., n$, representam os erros. São variáveis aleatórias que explicam a parte da variabilidade em Y que X não explica;
- β_0 e β_1 simbolizam os coeficientes (parâmetros) de regressão.

O parâmetro β_0 é chamado de intercepto ou coeficiente linear, pois caracteriza o ponto em que a reta regressora corta o eixo y quando X=0. Já o parâmetro β_1 representa a inclinação da reta.

O modelo de regressão linear tem alguns pressupostos, são eles:

- X e Y possuem relação linear;
- Os erros são independentes e possuem média nula, $E(\epsilon_i)= 0$, $i = 1, ..., n)$;
- A variância do erro é constante, $var(\epsilon_i) = \sigma^2$, $i = 1, ..., n$;
- Os erros (ϵ_i, $i = 1, ..., n$) possuem distribuição normal (RODRIGUES, 2012; SHIMAKURA, 2006).
- Na presença de duas ou mais variáveis explicativas, o modelo é denominado Modelo de Regressão Múltipla.

Para analisar o ajuste de um modelo, é possível calcular algumas métricas que serão espelho da qualidade do ajuste. Alguma delas são o erro quadrático médio e o erro absoluto médio, essas serão definidas a seguir.

2.2.1. Erro quadrático médio

No contexto desse estudo, o erro quadrático médio (EQM) é definido como a média do quadrado da diferença da variável resposta predita ($\hat{Y}$) pelo seu valor verdadeiro (Y):

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

2.2.2 Erro absoluto médio

O erro absoluto médio (EAM) é construído com base na média dos erros absolutos. Os erros advém da diferença absoluta entre a variável resposta predita ($\hat{Y}$) e seu valor observado (Y):

$$EAM = \frac{1}{n}\sum_{i=1}^{n} |\hat{Y}_i - Y_i|$$

2.4. Imputação única

Os métodos chamados de imputação única, ou simples, têm por objetivo imputar os dados faltantes através de uma substituição que ocorre uma única vez (ENGELS, 2003). Dentro da imputação única, existem alguns métodos, como a substituição por um valor de tendência central, "Hot deck", média predita por regressão, estimativa de máxima verossimilhança, entre outros. Nesse estudo o foco será analisar a substituição pela média e a substituição por regressão.

A imputação pela média ocorre substituindo os dados faltantes pela média dos dados observados na variável de interesse. A média utilizada pode ser a média geral dos valores observados ou a média do grupo mais similar onde existem dados faltantes, identificado pelas variáveis categóricas do banco de dados (NUNES, 2007; CASTRO, 2014).

Por outro lado, a imputação por regressão se utiliza da regressão simples ou múltipla para prever os valores a serem imputados. A regressão opera com uma ou mais variáveis existentes para predizer os dados faltantes de outra variável altamente correlacionada com as anteriores (NUNES, 2007).

Primeiro estima-se as equações de regressão que vão predizer as variáveis com os dados faltantes. E depois, são gerados os valores preditos a serem imputados, a partir das regressões estimadas (CASTRO, 2014).

3. Metodologia

Com o objetivo de comparar os diferentes tipos de imputação, por média e por regressão, foram realizadas simulações que permitissem a aplicação desses métodos. Para todo esse processo, foi utilizado o software R, versão 4.0.4.

A simulação consistiu em gerar diferentes amostras variando seu tamanho (20, 50, 100 e 200) e sua porcentagem de dados faltantes (5%, 20%, 30%, 50%) a fim de se comparar a imputação para diferentes cenários. As amostras constituíram-se de quatro variáveis, sendo três variáveis explicativas e a variável resposta :

- x_3: dados gerados da distribuição exponencial com parâmetro $\lambda = 0.2$;
- x_2: dados gerados na distribuição bernoulli com probabilidade de sucesso p = 0.4;
- x_1: dados gerados da função:

$$2x_2 - 0.1x_3 + \epsilon_1$$

  Onde ϵ_1 representa dados gerados de uma distribuição normal com parâmetros $\mu = 0$ e $\sigma = 0.5$. Note que x_1 é uma covariável que depende de x_2 e x_3, isso é importante para a aplicação do método de imputação da regressão linear que será visto a seguir.

  =

- y : variável resposta do modelo, definida por:

$$3x_1 - 0.5x_2 + x_3 + \epsilon$$

  Onde ϵ representa dados gerados de de uma distribuição normal com parâmetros $\mu = 0$ e $\sigma = 5$.

Tendo caracterizado as variáveis, foram geradas amostras com os dados completos. Em sequência os dados foram reduzidos, isto é, foram simulados dados faltantes para as amostras na variável x_1 (dependente de x_2 e x_3), variando, como já citado, a porcentagem de dados ausentes nas amostras. Com os dados reduzidos realizou-se primeiro a imputação pela média, substituindo os dados faltantes de x_1 pela média da variável. E para essa dada amostra, foi também realizada a imputação por regressão linear, onde se ajustou o modelo linear para x_1 como sendo variável resposta e x_2 e x_3 variáveis explicativas, o que possibilitou a obtenção de médias preditas que foram usadas na imputação de x_1.

Nesse processo foram utilizadas 1000 réplicas de Monte Carlo para gerar cada amostra, e para aferir as imputações foram calculadas as medidas de EQM e EAM, com base nos valores originais dos dados completos. Essas medidas foram baseadas na média das 1000 réplicas de Monte Carlo.

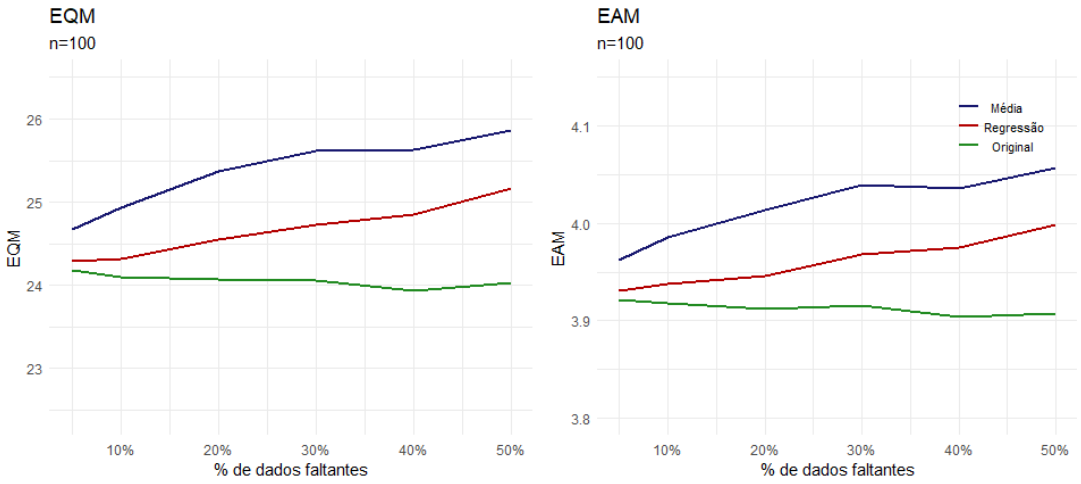
Além disso, os valores de EQM e EAM do modelo considerando os dados completos foram calculados com o intuito de comparar o desempenho dos dois métodos de imputação adotados nesse trabalho.

As simulações foram ilustradas através de gráficos de linhas e tabela, como será mostrado na próxima seção.

4. Resultados

Para produzir diferentes cenários de imputação, as amostras geradas foram variadas segundo seu tamanho (n) e sua porcentagem de dados faltantes (df's), os resultados dessas variações são apresentados na Tabela 1.

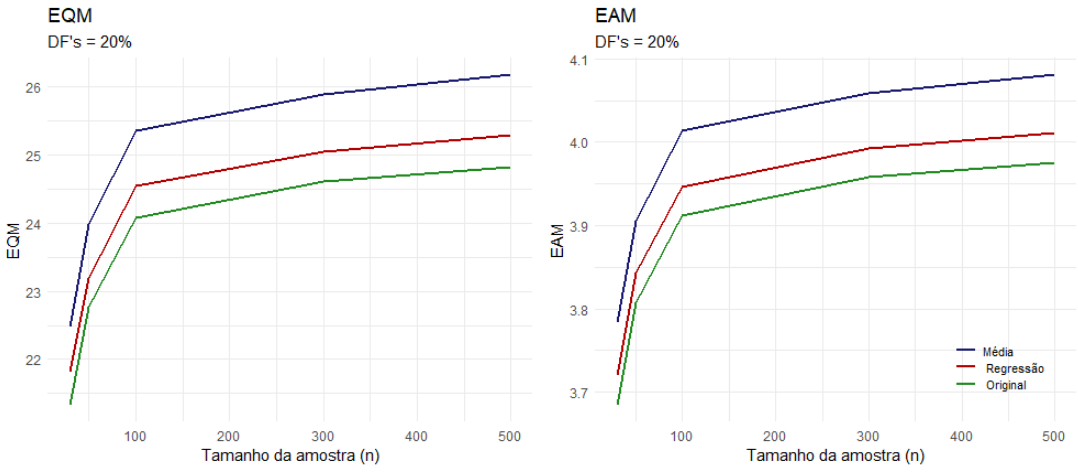
Figura 1: Gráfico de linhas para o EQM e o EAM das imputações com n = 100 fixado e variando a porcentagem de df's



No gráfico de linhas da Figura 1, fixou-se o tamanho da amostra igual a 100 e observou-se o comportamento do EQM e EAM para as diferentes porcentagens de dados faltantes na amostra. E como observado, o comportamento para as duas medidas de erro se mostra bastante similar. É possível verificar que a linha azul, que representa a imputação pela média, apresentou valor de erro superior para todas as porcentagens de dados faltantes. Já a linha vermelha, que nos mostra a imputação por regressão, com porcentagens de dados faltantes menores, se aproxima bastante da medida de erro para o conjunto de dados completos (linha verde) e se mantém abaixo da linha da imputação por média em todo gráfico. Isso indica a melhor performance da imputação por regressão para uma amostra de tamanho 100, com diferentes percentis de dados ausentes.

No gráfico da Figura 2 pode-se ver o comportamento do EQM e do EAM quando há

Figura 2: Gráfico de linhas para o EQM e o EAM das imputações com df's = 20% fixado e variando o tamanho da amostra



20% dos dados ausentes para diferentes tamanhos de amostras. Como na Figura1 vê-se que ambos gráficos para os diferentes erros possuem comportamentos muito semelhantes. Primeiramente o que se observa é a crescente dos valores de erro a medida que o tamanho da amostra cresce. Ademais a imputação pela média apresentou maiores valores de EQM e EAM quando comparados com aqueles erros obtidos pela imputação por regressão.

A Tabela 1 apresenta os resultados para diferentes tamanhos de amostras e proporções de dados faltantes. Percebe-se, primeiramente, que valores dos erros são maiores para as amostras maiores. Identifica-se que as menores medidas de erro são para a amostra de tamanho 20, com 5% de dados faltantes (EQM - média = 20.8 e EQM - RL = 19.78). Observa-se também que para todos os tamanhos de amostra (20, 50, 100 e 500) os erros apresentam um padrão: para todos os casos, em todas as proporções de dados faltantes, o erro quadrático médio e o erro absoluto médio apresentam maiores valores para a imputação por média, enquanto que os EQM's e os EAM's para a imputação por regressão são valores intermediários entre os erros pela imputação por média e os erros da amostra com dados completos. Isso indica a melhor performance da imputação por regressão independente do tamanho da amostra ou da proporção de dados faltantes.

Tabela 1: Valores dos EQM's e EAM's para as imputações com diferentes tamanhos de amostras e proporções de dados faltantes

N	% df's	EQM	EQM - média	EQM - RL	EAM	EAM - média	EAM - RL
20	5	19.68	20.08	19.78	3.53	3.57	3.54
20	20	20.12	21.18	20.53	3.57	3.66	3.60
20	30	20.22	21.25	20.75	3.56	3.66	3.61
20	50	19.78	21.41	21.13	3.54	3.68	3.64
50	5	23.06	23.47	23.17	3.84	3.87	3.84
50	20	22.90	24.09	23.33	3.82	3.92	3.86
50	30	23.13	24.58	23.77	3.83	3.96	3.89
50	50	22.92	24.77	24.06	3.81	3.96	3.90
100	5	23.89	24.39	23.99	3.90	3.94	3.91
100	20	23.94	25.19	24.40	3.91	4.01	3.94
100	30	24.06	25.64	24.73	3.91	4.04	3.96
100	50	23.79	25.59	24.90	3.90	4.04	3.99
500	5	24.78	25.30	24.90	3.97	4.01	3.98
500	20	24.78	26.10	25.23	3.97	4.07	4.01
500	30	24.89	26.50	25.57	3.98	4.11	4.04
500	50	24.77	26.66	25.89	3.97	4.12	4.06

5. Conclusão

Esta pesquisa teve como propósito comparar dois tipos de imputação única: a imputação por média e a imputação por regressão linear simples.

Através de simulações, foi possível verificar que para vários cenários, onde se foi estabelecido diferentes tamanhos de amostras e diferentes proporções de dados faltantes, chegando até um caso extremo de 50% de df's, a imputação por média se mostrou menos eficiente do que a imputação por regressão, qualquer que seja o tamanho da amostra ou porcentagem de dados faltantes.

Em observação, entende-se que a imputação por regressão é um método de melhor performance do que o método realizado pela média, por sua natureza matemática. Note que, a imputação por regressão em um cenário em que não há correlação entre as variáveis irá resultar em uma imputação pela média, pois o intecepto de um modelo linear sem

covariáveis é, na realidade, a média. Assim, pode-se considerar que a imputação pela média é um caso particular da imputação por regressão.

Ademais, para futuros estudos recomenda-se estudar outros métodos de imputação única e com mais variedades de cenários para realizar comparações.

Bibliografia

[1] ALLISON, Paul D. Missing data. Sage publications, 2001.

[2] CARVALHO, Melissa Mello. Dados faltantes em análises: uma revisão sobre métodos estatísticos flexíveis a incompletude. In: II Simpósio de Métodos Numéricos em Engenharia. 2017.

[3] CASTRO, Isabela Queirós. Uma aplicação de métodos de imputação no estudo de fatores associados ao baixo peso ao nascer. Monografia. Universidade Federal de Juiz de Fora, Juiz de Fora, 2014.

[4] ENGELS, Jean Mundahl; DIEHR, Paula. Imputation of missing longitudinal data: a comparison of methods. Journal of clinical epidemiology, 2003.

[5] ERCOLE, Flávia Falci; CARNEIRO, Mariângela; CHIANCA Tânia Couto M.; DUARTE, Denise. Efeito da imputação de dados faltantes em banco de dados de infecção de sítio cirúrgico em pacientes ortopédicos em Belo Horizonte. 2010.

[6] MARTINS, Maria Eugénia Graça. Regressão linear simples. Revista de Ciência Elementar, 2019, 7.3.

[7] MORETTIN, Pedro A.; BUSSAB, Wilton O. Estatística básica. Saraiva Educação SA, 2017.

[8] MYRTVEIT, Ingunn; STENSRUD, Erik; OLSSON, Ulf H. . Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods. IEEE Transactions on Software Engineering, 2001.

[9] NUNES, Luciana Neves. Métodos de imputação de dados aplicados na área da saúde. 2007.

[10] RIBEIRO, Elisalvo Alves, et al. Imputação de dados faltantes via algoritmo EM e rede neural MLP com o método de estimativa de máxima verossimilhança para aumentar a acurácia das estimativas. 2015.

[11] RODRIGUES, Sandra Cristina Antunes. Modelo de regressão linear e suas aplicações. 2012. PhD Thesis. Universidade da Beira Interior (Portugal).

[12] RUBIN, Donald B. Inference and missing data. Biometrika, 1976, 63.3: 581-592.

[13] SHIMAKURA, Silvia. Modelo de regressão linear simples. Notes: CE003 - Estatística II, 2006. Disponível em: http://leg.ufpr.br/ silvia/CE003/node81.html