Comparação empírica de diferentes métodos de imputação para modelos de regressão linear múltipla.

UFPB - Estatística Computacional

Paulo Ricardo Seganfredo Campana
Marcelo Rodrigo Portela Ferreira

22 de outubro de 2023

Resumo

Técnicas para lidar com valores ausentes em conjuntos de dados são não só importantes mas necessárias para a criação de modelos estatísticos. A imputação destes valores é uma técnica popular para solucionar este problema.

Dentre elas, iremos estudar o impacto na regressão linear múltipla de 5 métodos de imputação aplicados em 5 conjuntos de dados onde valores ausentes foram impostos de maneira aleatória, será avaliada o poder preditivo e a explicabilidade da regressão para comparar os diferentes métodos.

Palavras-chave: Imputação, Regressão.

Introdução

Para a maioria dos modelos estatísticos, incluindo os modelos regressão linear, um grande problema são as observações faltantes, também conhecidas como **NA** (*not available* / não disponível), não podemos estimar parâmetros quando uma ou mais observações estão ausentes, muito menos podemos retirar tais observações do conjunto de dados pois isso introduz um viés de seleção já que muitas vezes esta falta pode estar correlacionada com alguma variável de interesse.

Para remediar este problema existem técnicas de imputação, em que as observações ausentes de uma variável são substituídas por certos valores de modo a minimizar o viés introduzido

por essa manobra. Entre os métodos de imputação mais comuns temos imputação por média e mediana, em que todas as observações faltantes são substituídas por uma única estatística calculada por todas as outras observações. Outros métodos utilizam modelos de regressão secundários para imputação, fazendo uso da correlação entre as variáveis regressoras.

Metodologia

As seguintes análises foram realizadas na linguagem de programação R (R Core Team 2023) utilizando o *framework* de modelagem estatística *tidymodels* (Kuhn e Wickham 2020). os códigos da simulação estão disponíveis de maneira reproduzível no Github (Campana 2023).

Foram usados 5 conjuntos de dados para comparação, um conjunto de dados sintéticos e 4 incluídos em pacotes populares do R:

- Dados gerados:
  - 500 observações de 5 variáveis com distribuição conjunta Normal Multivariada.
  - $\mathcal{N}_5(0, \Sigma)$ em que Σ trás uma moderada correlação entre as variáveis.
  - y gerado como: $y_i = x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 + \varepsilon_i$.
  - ε de distribuição $\mathcal{N}(0, 5)$.
- `iris` (Fisher 1936).
- `diamonds` (Wickham 2016).
- `penguins` (Gorman, Williams, e Fraser 2014).
- `concrete` (Yeh 2006).

A seguinte tabela resume os conjuntos de dados:

Tabela 1: Conjuntos de dados utilizados para comparação

Dados	Observações	V. Quantitativas	V. Qualitativas
Dados gerados	500	5	0
`iris`	150	3	1
`diamonds`	53940	6	3
`penguins`	344	3	3
`concrete`	1030	8	0

Para cada conjunto, todas as variáveis regressoras numéricas tiveram 10% de suas observações escolhidas ao acaso substituídas por **NA**, um processo que iremos chamar de "ocultação", e aplicados 5 diferentes métodos de imputação, ajustado um modelo linear e observado as estatísticas de performance do modelo, este procedimento foi repetido em uma simulação Monte Carlo para obter estimativas pontuais e intervalares das métricas de performance.

Não foram selecionadas para ocultação as variáveis qualitativas pois não é possível aplicar imputação por média, mediana ou por modelo linear, porém há a possibilidade de imputação por moda e modelos *KNN* e *Bag* em futuros trabalhos. A ocultação foi feita de forma independente para cada variável.

Os métodos de imputação usados foram: imputação por média e mediana e por modelos lineares, *k-nearest neighbors* e *Bagged trees*, para estes três últimos, é criado submodelos para cada variável que necessita imputação de modo que as outras variáveis regressoras do conjunto de dados tentam prever o valor daquela que está sendo imputada.

Para o modelo de regressão linear, os conjuntos de dados são divididos em duas partes: o modelo é ajustado com os dados de treinamento que compõem 80% do total e as métricas de performance, as estatísticas da raiz do erro quadrático médio ($RMSE$) e o coeficiente de determinação (R^2) são estimadas com o restante dos 20% dos dados, os dados de validação. Esta separação é importante para evitarmos o que é conhecido como *overfitting*.

Devido a cada etapa deste procedimento levar um tempo considerável e a baixa variação entre cada etapa, houve apenas 1200 iterações de simulação de Monte Carlo em 12 processos em paralelo durante 3 horas. Os resultados serão apresentados em tabelas e gráficos a seguir.

Resultados

Segundo a Tabela 2, o método de imputação por submodelos: linear, *K-nearest neighbors* e *Bagged trees* apresenta em todos os casos um $RMSE$ menor e R^2 maior, indicando que as imputações feita por estes métodos são mais eficientes para encontrar os reais estimadores dos coeficientes do modelo de regressão. Em específico, a imputação por modelo linear parece se destacar um pouco das demais, porém apresenta desvio padrão destas métricas superior aos outros métodos.

Tabela 2: Resultados da simulação de Monte Carlo para a performance dos diferentes métodos de imputação nos conjuntos de dados estudados.

	RMSE		R^2	
Método	Média	Desvio padrão	Média	Desvio padrão
Normal Multivariada				
Média	5.636	0.517	0.899	0.021
Mediana	5.636	0.517	0.899	0.021
Modelo linear	4.626	0.429	0.931	0.018
K-nearest neighbors	5.070	0.393	0.918	0.016
Bagged trees	5.006	0.373	0.920	0.016
iris				
Média	0.184	0.025	0.942	0.017
Mediana	0.185	0.025	0.942	0.017
Modelo linear	0.172	0.029	0.950	0.020
K-nearest neighbors	0.173	0.026	0.949	0.017
Bagged trees	0.173	0.026	0.949	0.017
diamonds				
Média	1556.614	30.204	0.848	0.005
Mediana	1563.247	31.547	0.846	0.006
Modelo linear	1144.434	59.352	0.918	0.009
K-nearest neighbors	1195.835	22.883	0.910	0.003
Bagged trees	1147.990	20.706	0.917	0.003
penguins				
Média	301.666	23.569	0.859	0.027
Mediana	302.028	23.560	0.859	0.027
Modelo linear	294.092	26.702	0.866	0.030
K-nearest neighbors	294.130	22.980	0.866	0.026
Bagged trees	293.793	22.920	0.866	0.025
concrete				
Média	11.484	0.517	0.530	0.045
Mediana	11.529	0.529	0.527	0.046
Modelo linear	10.574	0.836	0.601	0.068
K-nearest neighbors	10.978	0.495	0.570	0.042
Bagged trees	10.959	0.497	0.572	0.042

Pelos intervalos de confiança da Figura 1 vemos que de fato os métodos de imputação por média e mediana não obtêm resultados tão bons quanto imputação por submodelos e o método de submodelo linear se destaca apenas em alguns conjuntos de dados: Normal Multivariada e `concrete`.

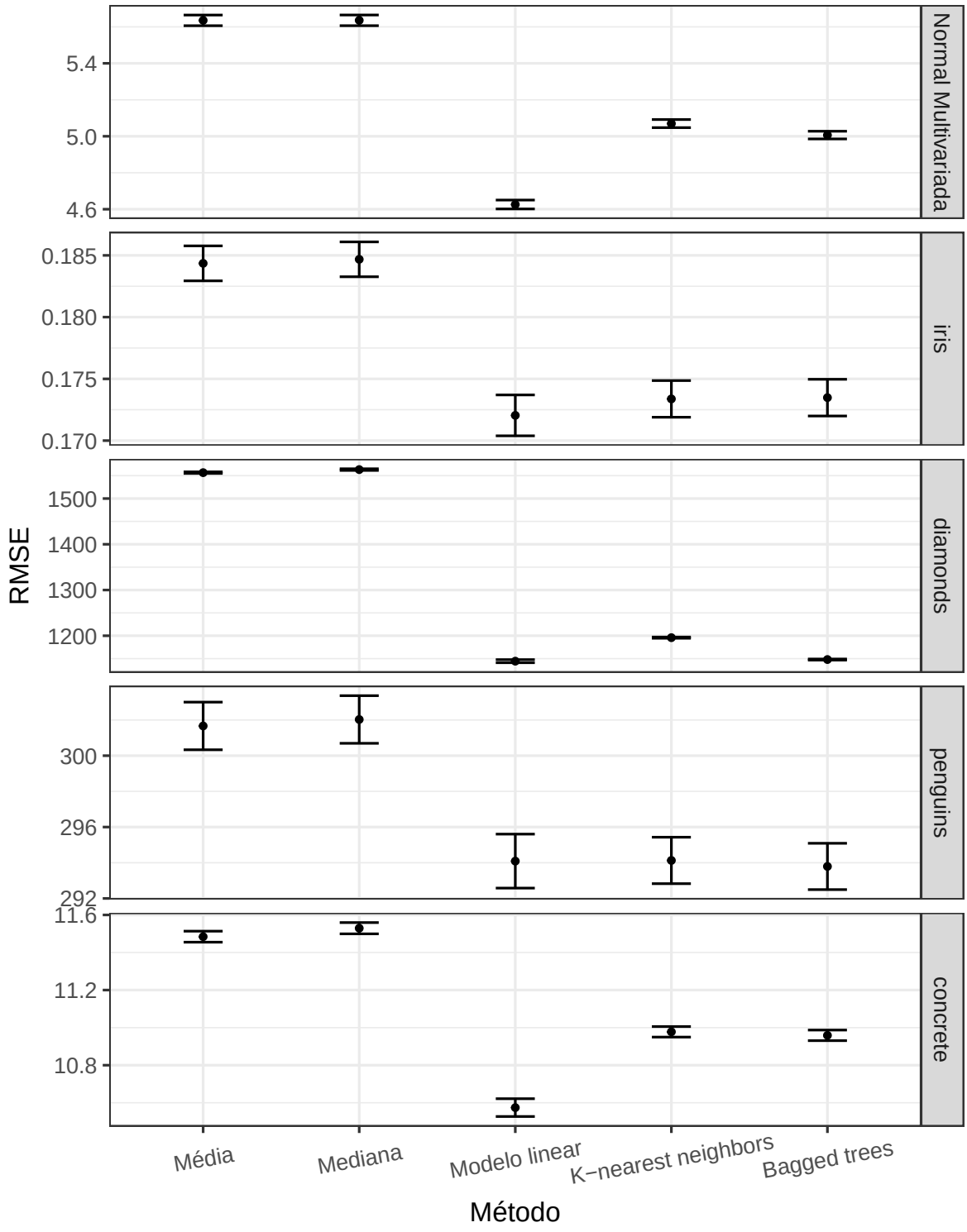

Figura 1: Intervalos de confiança para a raiz do erro quadrático médio da regressão com os diferentes métodos de imputação.

Conclusões

Desta forma, vemos que as imputações por média e mediana são ferramentas muito boas para podermos construir modelos onde há observações ausentes, porém quando estamos na situação de um modelo multivariado, os métodos de imputação por submodelos utilizam a correlação entre as variáveis regressoras para imputação, isso trás benefícios que são significantes em relação a imputação por média ou mediana.

Referências

Campana, Paulo R. S. 2023. «Código de simulação». *GitHub repository*. GitHub. https://github.com/PauloCampana/UFPB/blob/main/P5/comp/trabalho/trabalho.qmd.

Fisher, Ronald Aylmer. 1936. «The use of multiple measurements in taxonomic problems». *Annals of Eugenics* 7 (2): 179–88. https://doi.org/https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Gorman, Kristen B., Tony D. Williams, e William R. Fraser. 2014. «Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis)». *PLOS ONE* 9 (3): e90081. https://doi.org/10.1371/journal.pone.0090081.

Kuhn, Max, e Hadley Wickham. 2020. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.* https://www.tidymodels.org.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing* (versão 4.3.1). Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Yeh, I-Cheng. 2006. «Analysis of Strength of Concrete Using Design of Experiments and Neural Networks». *Journal of Materials in Civil Engineering* 18 (4): 597–604. https://doi.org/10.1061/(ASCE)0899-1561(2006)18:4(597).