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

UFPB - Estatística Computacional

Paulo Ricardo Seganfredo Campana
Marcelo Rodrigo Portela Ferreira

15 de outubro de 2023

Resumo

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).

Foram usados 5 conjuntos de dados para comparação: o primeiro foi gerado números aleatórios de distribuição Normal Multivariada com 4 variáveis, vetor de médias $\vec{\mathbf{0}}$ e matriz de covariância com moderada correlação entre as variáveis e adicionado uma variável y como combinação linear das outras 4 mais um erro normal. Os outros conjuntos de dados são compostos por dados disponíveis em pacotes populares do R: `iris` (Fisher 1936), `diamonds` (Wickham 2016), `penguins` (Gorman, Williams, e Fraser 2014) e `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
Normal Multivariada	500	4	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 foram escolhidas para terem 10% de suas observações escolhidas ao acaso substituídas por **NA**, 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 escolhidas 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 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.

Devido a cada etapa deste procedimento levar um tempo considerável e a baixa variação entre cada etapa, houve apenas 100 iterações de simulação de Monte Carlo, que em 12 processos em paralelo demorou ***TODO*** horas. Os resultados serão apresentados em tabelas e gráficos a seguir.

Resultados

Conclusões

Referências

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).