

Lista IV

Regressão linear

Bruno Marques Schaefer

18/06/2025

Abstract

Use o pacote `tidyverse` para realizar os exercícios. Todas as questões requerem interpretações ou explicações por escrito. Leia as instruções com atenção antes de começar a escrever o seu código

Definições

- Leia com calma os enunciados, escreva os códigos e exporte os resultados.
- Apague o enunciado da versão final (mantenha só o número da questão)

Nesta lista final, seu objetivo será analisar dados de bolsas de estudo da Capes usando modelos de regressão linear. Para tanto, usaremos uma base de dados com informações de bolsistas que receberam auxílio da Capes para estudar fora do país em 2012. Na pasta dados que acompanha esta lista, há um arquivo nomeado como:

- `bolsas_internacionais_capes_2012.xlsx`

Este arquivo contém uma base de dados com informações de mais de 18 mil bolsistas. Na base, cada linha indica uma bolsa e informações a ela referentes como duração da bolsa em meses (`estadia_em_meses`), a instituição de origem da pessoa beneficiada (`instituicao_ensino_origem`), a região onde fica essa instituição (`regiao`), além do valor recebido da Capes (`valor_recebido_total`). Será necessário carregar essa base para realizar alguns dos exercícios a seguir.

Tarefas

1) Regressão linear simples

Usando o banco de dados de bolsas da Capes, rode uma regressão linear na qual a variável dependente é o `valor_recebido_total` e, a independente, a variável `estadia_em_meses`. Reporte:

- a) Um gráfico do tipo *scatterplot* que ilustre a relação entre as variáveis e adicione a linha do melhor modelo linear predito pelo modelo
- b) No gráfico *scatterplot* reporte o coeficiente de correlação de Pearson.
- c) Reporte os coeficientes e os erros-padrão do modelo e os interprete brevemente

Dicas:

- Use ggplot2 para fazer o gráfico, mas lembre-se de especificar argumentos para adicionar a linha MQO sem intervalos de confiança (por padrão, a função necessária adiciona intervalos de confiança)
- É possível usar modelsummary ou texreg para gerar facilmente tabelas formatadas com os resultados da regressão

2) Variáveis independentes categóricas

Rode uma regressão linear na qual a variável dependente é o valor_recebido_total e, a independente, a região do país onde a instituição de origem da pessoa beneficiada está localizada.

- a) Faça um gráfico de barras que indique a média de valor_recebido_total por região do país
- b) Compare os coeficientes do modelo de regressão com as médias do gráfico anterior

Dica:

- É importante levar em conta os valores da constante, ou *intercept*, do modelo de regressão para interpretar corretamente os coeficientes das variáveis categóricas

3) Relação multivariada

Combinando as abordagens anteriores, rode um modelo que tenha como dependente valor_recebido_total e, como independentes, regioao e estadia_em_meses. A partir disso:

- a) Interprete os coeficientes do modelo, discutindo também T-valores e P-valores
- b) Escolha e faça um gráfico que mostre simultaneamente a relação da região e do tempo de estadia sobre com os valores de bolsa recebidos