Lista 5: Regressão Simples, Múltipla, e Checagem dos Pressupostos
FLP 0468/FLS 6183

Prof. Manoel Galdino
Gabriel Mardegan
Pedro Reis

Para entregar até: 29 de novembro de 2024

Exercícios

Esta é a quinta e última lista de exercícios da disciplina de métodos quantitivos de pesquisa e introdução à regressão linear. Ela compreende questões sobre inferência e interpretação de regressão linear simples, verificação dos pressupostos do modelo de regressão linear, e também sobre regressão linear múltipla.

Trata-se de uma lista mais extensa que as anteriores, devido aos tópicos que abarca. No entanto, ela terá um prazo maior para resolução e envio: *29 de novembro, última semana de aula.* Muitos dos exercícios podem ajudar na estruturação do trabalho final, inclusive com o aproveitamento de parte dos códigos utilizados.

1 Importando e limpando os dados do ENEM 2023

Nesta lista, utilizaremos os microdados do ENEM relativos a 2023.

1. Faça o download dos dados do ENEM 2023 (em formato zip).

Como indica o site do INEP, os microdados são o menor nível de desagregação dos dados relativos ao exame, contendo os questionários respondidos pelos inscritos. Portanto, o download pode demorar algum tempo. Além dos questionários, a pasta contém outros materiais de referência para compreensão dos dados, como o dicionário de dados, e um documento "leia-me".

Link: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem

2. Aponte o local do arquivo dentro de seu computador, e depois leia o documento através da função "read.csv2".

3. Selecione as variáveis: "NU_NOTA_CH", "NU_NOTA_MT", "NU_NOTA_CN", "NU_NOTA_LC", "TP_SEXO", "TP_COR_RACA", "TP_ESCOLA". Do que se trata cada variável? Examine o dicionário dos dados.

4. Limpe a base de dados: elimine os dados ausentes; nas variáveis de notas das provas, retire todas as observações iguais a zero; na variável de raça, elimine as observações com raça não declarada; na variável de tipo de escola, elimine as observações que não responderam qual tipo de escola frequentaram.

2 Recortando os dados

1. Repare na quantidade de observações/linhas em seu banco de dados. Trata-se da quantidade de inscritos no exame. Para facilitar o trabalho com os dados, faremos uma amostragem utilizando a função "slice_sample" do pacote `dplyr`. Escolha uma amostra de 100000 ou 10000 observações. Não se esqueça de definir uma semente ("set.seed").

Com a base de dados organizada e limpa, podemos começar a analisar os dados com os ferramentais de regressão linear.

3 Analisando os dados: regressão linear simples

Vamos utilizar as notas do ENEM em ciências humanas para prever as notas em matemática por meio de uma regressão linear simples.

1. Antes, entretanto, construa um gráfico de dispersão ("geom_point"), e faça uma observação preliminar sobre a relação entre as variáveis.

2. Rode a regressão utilizando a função "lm", apresente os resultados através da função "summary" e interprete os coeficientes.

4 Inferência estatística

1. Calcule o intervalo de confiança dos coeficientes ao nível de confiança de 95%. Tente fazer o cálculo utilizando os valores do erro padrão apresentados no sumário/resumo da regressão, e depois confirme os valores utilizando a função "confint".

2. Um dos pilares do modelo de regressão linear é a premissa da homoscedasticidade. Quando essa premissa é violada, estamos lidando com uma situação de heteroscedasticidade. Do que se trata esses dois conceitos?

3. Em breve, veremos um teste simples para checar homoscedasticidade. Por ora, a partir dos pacotes `lmtest` e `sandwich`, recalcule o nosso modelo de regressão simples usando uma fórmula de erro-padrão consistente com heteroscedasticidade ("HC1"). Veja os resultados do modelo recalculado e discuta as diferenças em relação ao modelo original.

5 Checagem do modelo

Conseguimos executar o modelo de regressão sobre as variáveis de interesse, interpretamos seus resultados e calculamos seus intervalos de confiança para fazermos inferências sobre o nosso fenô-

meno. Porém, como vimos, para a inferência a partir do modelo ser válida, seus pressupostos devem se manter em pé.

Em vista disso, precisamos fazer testes de checagem do modelo de regressão linear.

1. Gere os resíduos do modelo executado anteriormente.
2. Crie uma nova tabela com as seguintes variáveis: os resíduos gerados no item anterior e a variável preditora do nosso modelo.
3. Analisando apenas os resíduos: calcule sua média e gere um gráfico de densidade (ou um histograma) desses resíduos. O que sua distribuição nos diz? Como isso se relaciona com o que se espera teoricamente dos resíduos da regressão linear?
4. Falemos do gráfico quantil-quantil (Q-Q): nas suas palavras, qual é a utilidade do gráfico Q-Q? O que esse gráfico nos mostraria quando o comportamento dos resíduos é "normal"? E qual seria o comportamento desse gráfico em sua situação "anormal"?
5. Usemos as funções "qqnorm()" e "qqline()" sobre os nossos resíduos. O que podemos concluir a partir da nossa visualização?
6. Em um gráfico de dispersão, visualize a relação entre os resíduos (no eixo y) e os valores do preditor (no eixo x) do nosso modelo. Adicione ao gráfico uma reta de regressão linear usando "geom_smooth(method = "lm")". O que a visualização nos diz sobre a relação entre os resíduos e o preditor? Como isso se relaciona com a relação teoricamente esperada pelos pressupostos do modelo?
7. Tratemos o quadrado dos resíduos: examinemos a relação entre o preditor e o quadrado dos resíduos e visualizemos num gráfico de dispersão com uma reta ajustada. Dado o que esperamos teoricamente da variância condicional dos resíduos, o que o gráfico gerado nos diz?
8. Repita o processo do item anterior, mas, desta vez, use os resíduos absolutos: as conclusões mudam?
9. Usando a função "row_number()", ao lado das colunas dos resíduos e do preditor, gere uma nova coluna chamada "id" que identifica numericamente cada observação.
10. Nossa amostra foi selecionada aleatoriamente de uma base maior, logo, é improvável que haja uma relação entre a ordem da observação (se ela é a primeira, a décima, a última observação, etc.) e o valor dos resíduos. Mas, para boas condutas, geremos um gráfico de dispersão (com a reta ajustada) entre o id da observação (no eixo x) e os resíduos. Há algum padrão claro na relação? Podemos sugerir que os erros estão correlacionados?
11. Em um parágrafo curto, o que podemos concluir sobre a checagem geral do nosso modelo?

6 Regressão múltipla

Vamos adicionar mais variáveis preditoras para prever a nota de matemática.

1. Na nossa tabela de dados, converta as variáveis de sexo, raça, e escola em categóricas (use a função “factor()”). Além disso, usando a função “relevel()”, torne o sexo masculino, a raça branca, e a escola pública como os níveis de referência de suas respectivas variáveis – isso será muito importante para interpretar os coeficientes de regressão de variáveis categóricas.

2. Com a variável da nota de matemática como a variável dependente, execute a regressão linear com as demais variáveis como preditoras.

3. Apresente os resultados através da função “summary” e interprete os coeficientes – atente-se com a interpretação das variáveis numéricas e das variáveis categóricas.

4. Discuta as diferenças entre o modelo de regressão simples executado anteriormente e o modelo de regressão múltipla executado agora.

5. Recalcule o modelo de regressão múltipla usando uma fórmula de erro-padrão consistente com heteroscedasticidade (“HC1”) e discuta as diferenças em relação ao modelo de regressão múltipla original.

6. Refaça o processo de checagem dos modelos no modelo de regressão múltipla. O que você conclui sobre cada checagem e no geral? Suas conclusões sobre a checagem do modelo de regressão múltipla diferem das conclusões feitas sobre o modelo de regressão simples? Se sim ou se não, o que isso indica?