Roteiro

Camille Menezes

Maio de 2023

1 Introdução

- Nessa atividade, realizamos uma análise de Regressão Linear Múltipla com o objetivo de investigar a relação entre algumas variáveis e a prevalência de obesidade em dados de 403 afro-americanos residentes no Estado da Virginia, nos Estados Unidos;
- A primeira etapa dessa análise consistiu na manipulação dos dados e uma análise descritiva e exploratória, onde as características das variáveis quantitativas foram destacadas;
- Em seguida, ajustamos um modelo de regressão linear múltiplo os dados, em que a variável resposta foi o Índice de Massa Corporal. Identificamos quais variáveis têm um efeito significativo e a proporção da variação dos dados explicada pela regressão.
- Propusemos outro modelo de regressão, com menos variáveis, comparamos o ajuste dos dois modelos e interpretamos os coeficientes estimados;
- A avaliação da bondade de ajuste do modelo foi realizada por meio da análise de variância.
- E por último, uma série de gráficos de diagnóstico foram analisados. Esses gráficos nos ajudaram a identificar pontos atípicos, verificar a normalidade, homocedasticidade dos resíduos e avaliar a influência de observações sobre os resultados do modelo.

2 Manipulação dos dados

- Os dados inicialmente apresentavam essas variáveis: colesterol total, glicose estabilizada, colesterol bom, razão entre colesterol total e colesterol bom, hemoglobina glicada, município de residência, idade em anos, sexo, altura, peso, pressão sanguínea sistólica e diastólica (1o e 2o medidas), cintura e quadril;
- Nós convertemos altura, a cintura e o quadril para metros e o peso para quilos;
- Acrescentamos duas variáveis calculadas a partir das variáveis existentes: a variável IMC que é a divisão entre o peso e a altura ao quadrado e a RCQ (Relação Cintura Quadril) que foi calculada dividindo a cintura pela medida do quadril. Então reduzimos a quantidade de variáveis no modelo retirando o peso, a altura, a cintura e o quadril.
- Por haver muitos valores ausentes, as variáveis que denotam as pressões sanguíneas sistólica e diastólica na 2o medida também foram retiradas dos dados.

- Além disso, as linhas que contêm valores ausentes foram removidas. Após a remoção, restaram 377 observações no conjunto de dados.
- Então, as variáveis que foram utilizadas no modelo de regressão são: glicose estabilizada, razão entre colesterol total e colesterol bom, hemoglobina glicada, idade em anos, pressão sistólica, pressão diastólica e as novas variáveis IMC e RCQ.

3 Análise descritiva

- Ao observar os *boxplots* das variáveis, é possível notar que há muitos valores discrepantes, principalmente nas variáveis glicose estabilizada e hemoglobina glicada;
- Nos acreditamos que esses valores não devem ser retirados da análise, pois eles fazem parte da natureza das variáveis e parecem indicar alguma informação sobre a presença de diabetes nesses indivíduos, já que as duas avaliam a glicose no sangue;
- As variáveis pressão sanguínea diastólica, IMC e RCQ aparentam uma leve simetria;
- Quando se observa os gráficos de dispersão, as variáveis não aparentam ter uma relação linear visível, com exceção das relações entre as variáveis:
  - glicose estabilizada e hemoglobina glicada;
  - pressão sanguínea diastólica e sistólica.

  além disso, os coeficientes de correlação são baixos. Ele é maior que 0,6 apenas para essas duas relações.

4 Modelo de regressão linear múltiplo

- Essa tabela apresenta uma sumarização da regressão linear em que as características dos afro-americanas são utilizadas para predizer o seu IMC.
- É possível notar que a estimativa do intercepto, apesar de ser um valor possível do IMC, não possui interpretabilidade. Pois, o indivíduo não pode ter valores de RCQ, glicose estabilizada, pressão sistólica e diastólica iguais a zero, por exemplo.
- Apenas as variáveis razão colesterol total e colesterol bom, idade e pressão diastólica foram significativas ao nível de 10%.
- Além disso, o coeficiente de determinação foi baixíssimo, esse modelo de regressão explica apenas 9% da variabilidade do IMC.
- Então, parece que esse modelo de regressão linear múltiplo não descreve bem a natureza da variável IMC.