TDE02

Marcelo Wzorek Filho, Rafael Leal Machado

June 2024

1 Introdução

Neste exercício, vamos realizar uma análise de regressão linear múltipla para prever a variável *wheel-base* com base nas variáveis *normalized-losses* e *height*.

Para mais detalhes sobre o código utilizado, consulte o projeto no GitHub.

2 Variáveis presentes na base e seus significados

- **normalized-losses**: Indica as perdas normalizadas em termos de seguro de automóveis.
- **wheel-base**: Representa a distância entre os centros das rodas dianteiras e traseiras de um veículo, medida em polegadas.
- **height**: Refere-se à altura do veículo, medida em polegadas.

3 Variável a ser prevista (variável dependente)

A variável *wheel-base* será a variável dependente que pretendemos prever com base nas outras duas variáveis independentes (*normalized-losses* e *height*).

4 Cálculo do coeficiente de correlação

Para calcular o coeficiente de correlação entre *wheel-base* e as outras variáveis (*normalized-losses* e *height*), usaremos o método *corrcoef* do NumPy. O coeficiente de correlação (r) varia entre -1 e 1, onde:

- $r = 1$: Correlação positiva perfeita
- $r = -1$: Correlação negativa perfeita
- $r = 0$: Não há correlação linear

Escolheremos as duas variáveis independentes com maior correlação absoluta com a variável dependente *wheel-base*.

5 Passo-a-passo da Regressão Linear Múltipla

A regressão linear múltipla visa encontrar uma equação que melhor se ajusta aos dados disponíveis. A equação resultante terá a forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Onde:

- Y é a variável dependente (nesse caso, *wheel-base*).
- X_1 e X_2 são as variáveis independentes (*normalized-losses* e *height*).
- β_0 é o intercepto.
- β_1 e β_2 são os coeficientes de regressão para as variáveis independentes.
- ϵ é o erro aleatório.

Para calcular os coeficientes β_0, β_1, e β_2, usaremos o método de mínimos quadrados ordinários (OLS).

6 Avaliação da qualidade das previsões

Uma medida comum para avaliar a qualidade das previsões em regressão linear é o coeficiente de determinação R^2, que varia entre 0 e 1. Quanto mais próximo de 1, melhor o modelo se ajusta aos dados. Isso nos diz quanto da variabilidade da variável dependente é explicada pelas variáveis independentes no modelo.

O coeficiente de determinação R^2 é calculado como:

$$R^2 = 1 - \frac{\text{Soma dos quadrados dos resíduos}}{\text{Soma total dos quadrados}}$$

Onde:

- Soma dos quadrados dos resíduos: Soma dos quadrados das diferenças entre os valores observados e os valores previstos.
- Soma total dos quadrados: Soma dos quadrados das diferenças entre os valores observados e a média dos valores observados.