

ATIVIDADE AVALIATIVA I

Guilherme Rocha Duarte¹

RESUMO

Este documento explora a aplicação de regressão linear múltipla utilizando um conjunto de dados hipotético. Foram gerados coeficientes beta usando álgebra matricial e uma função de regressão existente no Python.

Palavras-chave: matriz, álgebra matricial, regressão linear múltipla.

ABSTRACT

This document explores the application of multiple linear regression using a hypothetical data set. Beta coefficients were generated using matrix algebra and an existing regression function in Python.

Key Words: matrix, matrix algebra, multiple linear regression.

¹Aluno do curso de Ciência de Dados e Inteligência Artificial pelo IESB.

1 INTRODUÇÃO

Este documento apresenta um exemplo prático de regressão linear múltipla, onde se visa prever o preço de casas com base em variáveis explicativas como área, número de quartos e idade da propriedade. São realizados cálculos dos coeficientes beta utilizando álgebra matricial e a função `LinearRegression` da biblioteca `scikit-learn`.

2 MÉTODOS

Foram utilizados dados hipotéticos de preços de casas com base em três variáveis: Tamanho (m²), Número de Quartos e Idade (anos). Os dados gerados foram os seguintes:

Preço	Área (m ²)	Quartos	Idade (anos)
350	120	3	10
450	150	4	5
300	100	2	20
500	180	4	2
400	130	3	8

Tabela 1 – Dados hipotéticos utilizados no estudo.

Matriz Original

A matriz original A representa os dados coletados no estudo, onde as linhas correspondem a diferentes imóveis e as colunas representam as características desses imóveis: preço, área em metros quadrados, número de quartos e idade em anos.

Matriz Original A

$$A = \begin{bmatrix} 350 & 120 & 3 & 10 \\ 450 & 150 & 4 & 5 \\ 300 & 100 & 2 & 20 \\ 500 & 180 & 4 & 2 \\ 400 & 130 & 3 & 8 \end{bmatrix}$$

Matriz Transposta

A matriz transposta A^T é obtida trocando-se as linhas por colunas da matriz original. Nesse caso, cada coluna da matriz transposta representa uma característica específica dos imóveis, agrupando os valores correspondentes de cada imóvel.

Matriz Transposta A^T

$$A^T = \begin{bmatrix} 350 & 450 & 300 & 500 & 400 \\ 120 & 150 & 100 & 180 & 130 \\ 3 & 4 & 2 & 4 & 3 \\ 10 & 5 & 20 & 2 & 8 \end{bmatrix}$$

Regressão Linear em Python

A matriz X representa as variáveis explicativas utilizadas no modelo de regressão linear para prever o preço dos imóveis. Cada linha corresponde a um imóvel e cada coluna a uma característica desse imóvel.

Matriz X antes de adicionar o intercepto

$$X = \begin{bmatrix} 120 & 3 & 10 \\ 150 & 4 & 5 \\ 100 & 2 & 20 \\ 180 & 4 & 2 \\ 130 & 3 & 8 \end{bmatrix}$$

- **Colunas:**

- **Área (m²):** Primeira coluna, representando o tamanho do imóvel em metros quadrados.
- **Quartos:** Segunda coluna, representando o número de quartos no imóvel.
- **Idade (anos):** Terceira coluna, representando a idade do imóvel em anos.

Adicionando o Intercepto

No código Python, uma coluna de 1s é adicionada à matriz X , permitindo que o modelo calcule um termo de intercepto durante o ajuste da regressão linear.

Matriz X após adicionar o intercepto

$$X = \begin{bmatrix} 1 & 120 & 3 & 10 \\ 1 & 150 & 4 & 5 \\ 1 & 100 & 2 & 20 \\ 1 & 180 & 4 & 2 \\ 1 & 130 & 3 & 8 \end{bmatrix}$$

- **Intercepto:** A nova primeira coluna contém apenas 1s, o que permite ao modelo calcular o coeficiente beta associado ao intercepto.

Ajuste do Modelo de Regressão Linear

O modelo de regressão foi ajustado utilizando álgebra matricial, onde os coeficientes β foram calculados através da seguinte equação:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Em seguida, o modelo foi ajustado utilizando a função `LinearRegression` da biblioteca `scikit-learn` no Python. Os coeficientes obtidos foram comparados para verificar a consistência dos resultados.

3 RESULTADOS

Os coeficientes β obtidos para o modelo de regressão linear são apresentados a seguir:

- **Intercepto (β_0): 122.40**

O coeficiente de intercepto indica que, na ausência de qualquer efeito das variáveis explicativas (Área, Quartos e Idade), o preço médio inicial estimado de um imóvel seria de 122.40 unidades monetárias.

- **Coeficiente para Área (β_1): 1.86**

Este coeficiente sugere que, para cada metro quadrado adicional na área do imóvel, o preço aumenta em aproximadamente 1.86 unidades monetárias, mantendo as demais variáveis constantes.

- **Coeficiente para Quartos (β_2): 12.66**

O coeficiente para o número de quartos indica que, para cada quarto adicional, o preço do imóvel aumenta em 12.66 unidades monetárias, assumindo que a área e a idade permanecem inalteradas.

- **Coeficiente para Idade (β_3): -1.79**

Este coeficiente negativo mostra que, para cada ano a mais de idade do imóvel, o preço diminui em 1.79 unidades monetárias, dado que as demais variáveis permanecem constantes.

Os coeficientes obtidos pela álgebra matricial e pela função de regressão linear do `scikit-learn` foram idênticos:

- **Coeficientes Beta (álgebra matricial):** [122.40075614, 1.86200378, 12.66540643, -1.79584121]
- **Coeficientes Beta (scikit-learn):** [122.40075614, 1.86200378, 12.66540643, -1.79584121]

A igualdade entre os coeficientes mostra que ambos os métodos, tanto a álgebra matricial quanto o uso da biblioteca `scikit-learn`, são consistentes e precisos. Isso reforça a confiabilidade dos cálculos matriciais e do algoritmo de regressão linear do Python, validando a implementação e os resultados obtidos.

4 CONCLUSÃO

A identidade dos coeficientes beta obtidos pelos dois métodos confirma a precisão do cálculo matricial e sua equivalência com as implementações computacionais. Este resultado destaca a relevância da álgebra matricial como uma ferramenta eficaz para a análise de regressão linear múltipla, especialmente em contextos que demandam robustez e confiabilidade nos cálculos.