ATIVIDADE AVALIATIVA I

Guilherme Rocha Duarte[1]

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.

[1]Aluno do curso de Ciência de Dados e Inteligência Articial 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

Usando o ChatGPT como auxilio nesse estudo, foi utilizado o seguinte *prompt*:

> 1. Crie um exemplo hipotético de regressão linear múltipla com uma matriz composta por uma variável resposta e 3 variáveis explicativas
> 2. Calcule os coeficiente betas com o uso de álgebra matricial, com e sem o uso da linguagem de programação (python).
> 3. Faça a regressão linear com o uso de uma função já existente no Python. Compare os resultados dos coeficientes beta resultantes do modelo com o uso da função e com o uso dos cálculos matriciais.

Os dados gerados no número 1, foram organizados em forma de tabela. Estão dessa maneira:

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.

Após essa geração, foi pedido o seguinte:

> gere a matriz de covariancia e correlação dos dados

também foi pedido para que todos os resultados gerados sejam analisados e explicados:

> agora explique a relação entre matriz de covariancia e correlação. lembre-se de deixar os calculos necessarios explicitos

2.1 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

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

2.2 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

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

2.3 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.

2.4 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.

2.5 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.

2.6 Matriz de Covariância

A covariância entre duas variáveis X e Y é dada por:

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

onde X_i e Y_i são os valores das variáveis X e Y, e $\bar{X}$ e $\bar{Y}$ são as médias dessas variáveis.

A matriz de covariância Σ para um conjunto de variáveis X_1, X_2, X_3 é:

$$\Sigma = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \text{Cov}(X_1, X_3) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \text{Cov}(X_2, X_3) \\ \text{Cov}(X_3, X_1) & \text{Cov}(X_3, X_2) & \text{Var}(X_3) \end{pmatrix}$$

2.7 Matriz de Correlação

A correlação entre duas variáveis X e Y é dada por:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

onde σ_X e σ_Y são os desvios-padrão de X e Y, respectivamente.

2.8 Relação entre as Matrizes

Para converter a matriz de covariância Σ na matriz de correlação R, utilizamos:

$$R_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii} \cdot \Sigma_{jj}}}$$

onde R_{ij} é o elemento da matriz de correlação, Σ_{ij} é o elemento correspondente da matriz de covariância, e Σ_{ii} e Σ_{jj} são as variâncias das variáveis i e j.

2.9 Exemplo com os Dados

Dada a matriz de covariância:

$$\Sigma = \begin{pmatrix} 930.0 & 23.5 & -192.5 \\ 23.5 & 0.7 & -5.5 \\ -192.5 & -5.5 & 47.0 \end{pmatrix}$$

Calculamos os desvios-padrão:

$$\sigma_{\text{Área}} = \sqrt{930.0} = 30.49, \quad \sigma_{\text{Quartos}} = \sqrt{0.7} = 0.84, \quad \sigma_{\text{Idade}} = \sqrt{47.0} = 6.85$$

Agora, calculamos os elementos da matriz de correlação:

$$R_{12} = \frac{23.5}{30.49 \times 0.84} = 0.921, \quad R_{13} = \frac{-192.5}{30.49 \times 6.85} = -0.921, \quad R_{23} = \frac{-5.5}{0.84 \times 6.85} = -0.959$$

Finalmente, a matriz de correlação R é:

$$R = \begin{pmatrix} 1.000 & 0.921 & -0.921 \\ 0.921 & 1.000 & -0.959 \\ -0.921 & -0.959 & 1.000 \end{pmatrix}$$

A matriz de correlação é derivada da matriz de covariância ao dividir cada covariância pelos produtos dos desvios-padrão das variáveis correspondentes. Ela oferece uma comparação padronizada das relações entre as variáveis, facilitando a interpretação das forças e direções das relações lineares.

2.10 Relação entre o Coeficiente de Correlação e o Coeficiente de Determinação

O coeficiente de correlação linear (r) está relacionado ao coeficiente de determinação (R^2) pelo quadrado do valor de r. Em um modelo de regressão, o coeficiente de determinação é expresso como:

$$R^2 = r^2$$

Onde:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Aplicando ao exemplo dado (usando o ChatGPT), a saída do método OLS em Python mostra os coeficientes, a estatística R^2, entre outras métricas:

> baseado nesses dados data = 'Preço': [350, 450, 300, 500, 400], 'Área': [120, 150, 100, 180, 130], 'Quartos': [3, 4, 2, 4, 3], 'Idade': [10, 5, 20, 2, 8] faça (lembre-se que tem o intercepto):
> - Calcule a matriz de covariância e correlação com o uso de álgebra matricial, com e sem o uso da linguagem de programação (python ou R).
> - Explique a relação existente entre a matriz de correlação e a matriz de covariância.
> - Calcule a soma dos quadrados dos resíduos (SQR) com e sem o uso da linguagem de programação (python ou R).
> - A partir do modelo de regressão linear com o uso de uma função já existente no Python ou R, compare dos itens 2, 3 e 4 do modelo gerados pelo uso da função de regressão linear no R ou Python e pelo uso dos cálculos matriciais.

Após os resultados gerados pela inteligência artificial, também foi calculado o R^2:

```
R-squared: 0.981
```

Nesse caso, o R^2 ajustado não é definido, pois não há graus de liberdade suficientes para calcular uma estimativa válida.

2.11 Soma dos Quadrados dos Resíduos (SQR)

A Soma dos Quadrados dos Resíduos (SQR) é uma medida da variação nos dados que não é explicada pelo modelo de regressão. Ela é usada para avaliar o quão bem o modelo se ajusta aos dados observados.

Dado um modelo de regressão linear:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

onde:

- y_i são os valores observados da variável dependente,
- $\hat{y}_i$ são os valores preditos pelo modelo,
- $\epsilon_i = y_i - \hat{y}_i$ são os resíduos (erros),
- n é o número de observações.

A SQR é definida como:

$$\text{SQR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Interpretação

- **Valor baixo de SQR:** Indica que os valores preditos estão próximos dos valores observados, sugerindo um bom ajuste do modelo.
- **Valor alto de SQR:** Indica que os valores preditos estão distantes dos valores observados, sugerindo um ajuste ruim do modelo.

A SQR é crucial na análise de regressão, pois ajuda a entender a quantidade de variabilidade que o modelo não consegue capturar. Quanto menor a SQR, melhor o ajuste do modelo aos dados.

Para calcular a soma dos quadrados dos resíduos foi utilizado o *prompt*:

> Calcule a soma dos quadrados dos resíduos (SQR) com e sem o uso da linguagem de programação (python)

O resultado obtido foi:

```
SQR: 0.0829
```

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.

A matriz de covariância fornece uma medida de como duas variáveis variam em relação uma à outra. No caso das variáveis 'Área', 'Quartos' e 'Idade', a matriz de covariância mostra a relação entre cada par dessas variáveis. A matriz de correlação, por sua vez, normaliza essas covariâncias pelo produto dos desvios padrão das variáveis, resultando em valores entre -1 e 1 que indicam a força e a direção da relação linear entre as variáveis.

Essas matrizes são importantes para entender as relações lineares entre as variáveis e como elas se inter-relacionam no contexto do modelo de regressão.

A Soma dos Quadrados dos Resíduos (SQR) quantifica a quantidade de variabilidade nos dados que não é explicada pelo modelo de regressão. Um valor de SQR baixo, como o obtido (0.0829), indica que o modelo de regressão se ajusta bem aos dados, com os valores preditos pelo modelo estando muito próximos dos valores observados. Isso sugere que o modelo é eficaz em capturar a variabilidade dos dados.

O coeficiente de determinação (R^2) mede a proporção da variância total dos dados explicada pelo modelo de regressão. No caso, o valor de R^2 obtido foi 0.981, indicando que aproximadamente 98.1% da variância nos dados é explicada pelo modelo. Isso representa um ajuste ótimo do modelo, mostrando que ele é capaz de explicar quase toda a variabilidade observada nos dados.

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.

Os resultados indicam que o modelo de regressão linear se ajusta muito bem aos dados, como evidenciado pelo baixo valor de SQR e pelo alto valor de R^2. A matriz de

covariância e a matriz de correlação confirmam as relações lineares entre as variáveis incluídas no modelo. Em conjunto, esses resultados demonstram que o modelo é robusto e eficaz na explicação das variáveis em questão.