UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia

Probabilidade e Estatística

RELATÓRIO DO SEMINÁRIO:

Análise de dados usando a Regressão Linear simples

Allan M. de Almeida -

Gabrielly M. da S. Barbosa

Maurício A.

Marcus Vinicius A. Silva

Maria Luiza F. P. Cesar

Vinícius F Araújo

07/02/2024 São Carlos - SP

1. INTRODUÇÃO

A Estatística desempenha um papel fundamental na interpretação e modelagem de dados, permitindo que possamos extrair informações valiosas a partir de observações do mundo real. Dentro desse contexto, a Regressão Linear Simples é uma das técnicas estatísticas mais utilizadas para analisar a relação entre duas variáveis quantitativas. Essa técnica busca modelar a dependência entre uma variável independente (explicativa) e uma variável dependente (resposta).

A regressão linear simples é amplamente aplicada em diversas áreas, como economia, engenharia, ciências sociais e inteligência artificial, permitindo prever tendências, analisar padrões e tomar decisões com base em dados.

Neste seminário, exploraremos o uso da regressão linear simples, através de uma aplicação prática e interpretando seus resultados.

2. O PROBLEMA 11.59

Nosso exercício escolhido para explorar o uso da regressão linear simples foi o 11.59 do Capítulo 11 do livro *Probabilidade e Estatística para Engenharia e Ciências - Ronald N. Walpole*. O exercício em questão trata-se do seguinte:

"Um experimento foi desenvolvido pelo Departamento de Engenharia de Materiais do Instituto Politécnico e Universidade Estadual da Virgínia, para estudar as propriedades de fragilização do hidrogênio com base nas medidas de pressão eletrolítica do hidrogênio. A solução usada foi 0,1 *N* NaOH e o material usado era um tipo de aço inoxidável. A densidade de carregamento catódico da corrente foi controlada e variada em quatro níveis. A pressão efetiva do hidrogênio foi observada como resposta. Os dados são apresentados a seguir:

Série	**Densidade de carregamento de corrente, x (mA/cm^2)**	**Pressão efetiva do hidrogênio, y (atm)**
1	0,5	86,1
2	0,5	92,1
3	0,5	64,7
4	0,5	74,7
5	1,5	223,6

Série	Densidade de carregamento de corrente, x (mA/cm²)	Pressão efetiva do hidrogênio, y (atm)
6	1,5	202,1
7	2,5	132,9
8	2,5	413,5
9	2,5	231,5
10	2,5	466,7
11	2,5	365,3
12	3,5	493,7
13	3,5	382,3
14	3,5	447,2
15	3,5	563,8

Tabela 1”

- Trazendo esses dados para um gráfico obtêm-se:

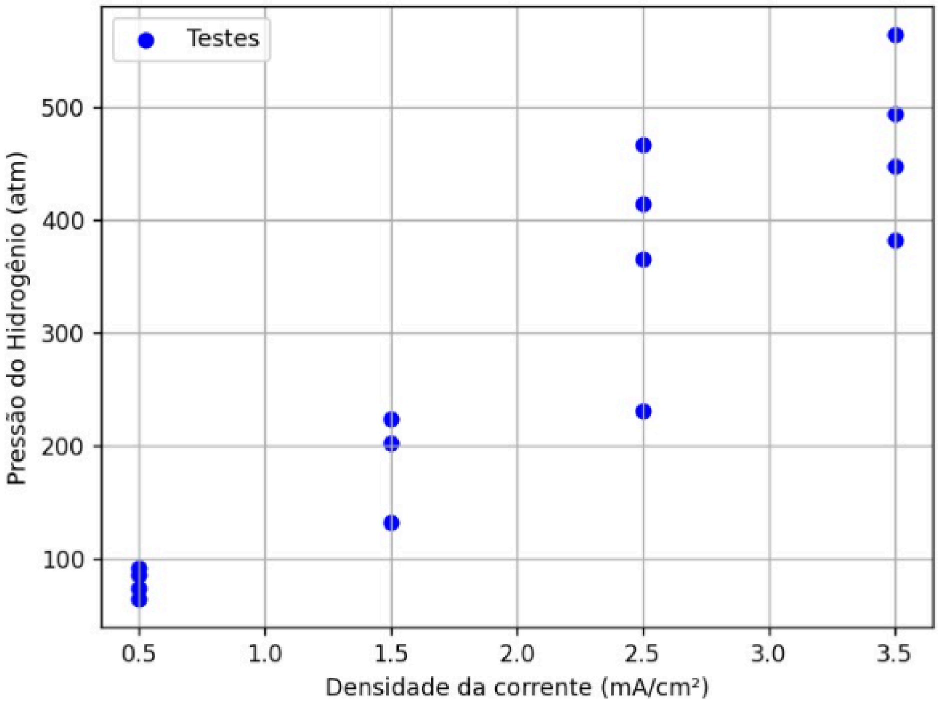

Figura 1. Gráfico de dados

3. DESENVOLVIMENTO E MÉTODOS

Com isto, a ideia inicial foi de criar um outro gráfico, mas utilizando apenas as variáveis de maneira direta, a fim de analisar a reta da regressão linear e avaliar os resultados. Essa abordagem permitirá visualizar a tendência dos pontos, sendo possível avaliar a adequação do modelo estatístico e compreender melhor o efeito da densidade de carregamento catódico na pressão efetiva do hidrogênio, contribuindo para o entendimento do fenômeno da fragilização deste material. Para isso, fizemos um algoritmo na linguagem de programação *Python* que nos ajuda a chegarmos a reta a qual desejamos:

```python
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
import matplotlib.pyplot as plt

x = np.array([0.5, 0.5, 0.5, 0.5, 1.5, 1.5, 1.5, 2.5, 2.5, 2.5, 2.5, 3.5, 3.5, 3.5, 3.5]).reshape(-1, 1)

y = np.array([86.1, 92.1, 64.7, 74.7, 223.6, 202.1, 132.9, 413.5, 231.5, 466.7, 365.3, 493.7, 382.3, 447.2, 563.8])

model = LinearRegression()
model.fit(x, y)

intercepto = model.intercept_
inclinacao = model.coef_[0]

print(f"Intercepto (β₀): {intercepto}")
print(f"Inclinação (β₁): {inclinacao}")

y_pred = model.predict(x)
print(f"Previsões: {y_pred}")

r2 = r2_score(y, y_pred)
print(f"Coeficiente de Determinação (R²): {r2}")

if inclinacao >= 0:
    print("Equação da reta: \n", f"y = {intercepto:.2f} + {inclinacao:.2f}x" )
else:
    print("Equação da reta: \n", f"y = {intercepto:.2f} {inclinacao:.2f}x")

plt.scatter(x, y, color='blue', label='Testes')

plt.plot(x, y_pred, color='red', label='Reta ajustada')

plt.grid(True)
plt.xlabel('Densidade da corrente (mA/cm²)')
plt.ylabel('Pressão do Hidrogênio (atm)')
plt.legend()
plt.show()
```

Figura 2. Algoritmo em Python

● Obtendo o gráfico com a reta a seguir:

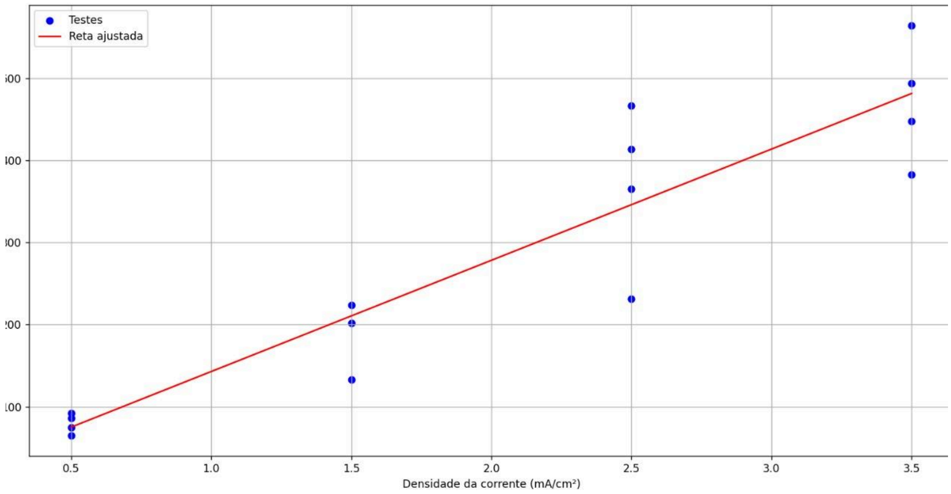

Figura 3. Reta da Regressão Linear

Através disso, conseguimos a inclinação (β1) da reta de regressão é um dos principais coeficientes na **Regressão Linear Simples**, pois descreve a direção e a intensidade da relação entre a variável independente (X) e a variável dependente (Y). Sua interpretação pode ser feita da seguinte maneira:

- **Se β1 > 0 (relação positiva):** indica que existe uma **relação diretamente proporcional** entre as variáveis. Ou seja, à medida que X aumenta, Y também tende a aumentar. Esse comportamento sugere que a variável independente exerce um impacto positivo sobre a variável resposta.
- **Se β1 < 0 (relação negativa):** indica que há uma **relação inversamente proporcional** entre as variáveis. Nesse caso, um aumento em X está associado a uma diminuição em Y. Isso sugere que a variável independente tem um efeito negativo sobre a variável dependente, reduzindo seus valores à medida que cresce.
- **Se β1 ≈ 0 (sem relação significativa):** sugere que **não há uma relação linear forte entre as variáveis**. Isso significa que mudanças em X não impactam diretamente os valores de Y, podendo indicar que outros fatores influenciam a variável resposta ou que a relação entre as variáveis não é linear.
- Em nossa análise, obtivemos uma inclinação de **(β1): 135.40337837837842**.

Dessa forma também conseguimos o coeficiente R^2. Este coeficiente é uma métrica estatística utilizada para avaliar a qualidade do ajuste de um modelo de regressão linear. Ele indica a proporção da variabilidade da variável dependente (Y) que é explicada pela variável independente (X) no modelo. Sua interpretação pode ser feita da seguinte maneira:

- Se $R^2 = 1$: O modelo **explica 100% da variação dos dados**. Isso significa que todos os pontos ajustam-se perfeitamente à reta de regressão, sem erros residuais. Esse cenário, porém, é raro em dados reais, pois sempre há algum grau de variabilidade não explicado pelo modelo.
- Se $R^2 = 0$: Indica o **pior cenário possível**, no qual o modelo não consegue explicar nenhuma variação na variável dependente. Isso significa que os valores de Y são completamente independentes dos valores de X, sugerindo que a regressão linear não é adequada para descrever a relação entre as variáveis.
- Se $0 < R^2 < 1$: O modelo explica parcialmente a variação dos dados. Quanto maior o valor de R^2, mais forte é a relação linear entre X e Y, indicando que a variável independente tem um impacto significativo sobre a variável dependente.
- Em nossa análise, obtivemos um coeficiente de determinação **(R^2) de: 0.8632097106761265**

Ao analisar o gráfico da regressão linear, percebe-se que para **valores menores da variável independente (X)**, as estimativas obtidas pelo modelo estão **bem ajustadas** aos pontos observados. Isso indica que, nessa região, a relação linear entre as variáveis é mais consistente e a reta de regressão consegue explicar com maior precisão a variação da variável dependente (Y).

No entanto, **conforme os valores de X aumentam, observa-se um aumento na dispersão entre os pontos experimentais e a reta de regressão**. Esse fenômeno sugere que, para valores mais altos da variável independente, o modelo linear pode não ser o mais adequado para capturar a complexidade dos dados, tornando o gráfico inconclusivo.

Dito isso, para a solução deste problema da dispersão crescente dos pontos em relação à reta de regressão para valores maiores de X, podemos recorrer a uma **transformação matemática** dos dados. Uma abordagem eficaz é a **escala logarítmica**, que pode tornar o comportamento do gráfico **mais constante e confiável**. Trazendo um gráfico que permite **uma melhor visualização e interpretação dos resultados**, tornando o modelo mais confiável para previsões futuras.

```python
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
import matplotlib.pyplot as plt

x = np.array([0.5, 0.5, 0.5, 0.5, 1.5, 1.5, 1.5, 2.5, 2.5, 2.5, 2.5, 3.5, 3.5, 3.5, 3.5]).reshape(-1, 1)
y = np.array([86.1, 92.1, 64.7, 74.7, 223.6, 202.1, 132.9, 413.5, 231.5, 466.7, 365.3, 493.7, 382.3, 447.2, 563.8])

x_log = np.log10(x)
y_log = np.log10(y)

model = LinearRegression()
model.fit(x_log, y_log)

intercepto = model.intercept_
inclinacao = model.coef_[0]

print(f"Intercepto (β₀): {intercepto}")
print(f"Inclinação (β₁): {inclinacao}")

y_log_pred = model.predict(x_log)

print(y_log_pred)

r2 = r2_score(y_log, y_log_pred)
print(f"Coeficiente de Determinação (R²): {r2}")

if inclinacao >= 0:
    print("Equação da reta no espaço transformado: \n", f"log(y) = {intercepto:.2f} + {inclinacao:.2f}log(x)")
else:
    print("Equação da reta no espaço transformado: \n", f"log(y) = {intercepto:.2f} {inclinacao:.2f}log(x)")

plt.scatter(x_log, y_log, color='blue', label='Testes (log-transformados)')
plt.plot(x_log, y_log_pred, color='red', label='Reta ajustada (log-transformada)')
plt.grid(True)
plt.xlabel('Log(Densidade da corrente) (log(mA/cm²))')
plt.ylabel('Log(Pressão do Hidrogênio) (log(atm))')
plt.legend()
plt.show()
```

Figura 4. Algoritmo em Python atualizado

- E obtemos a reta atualizada:

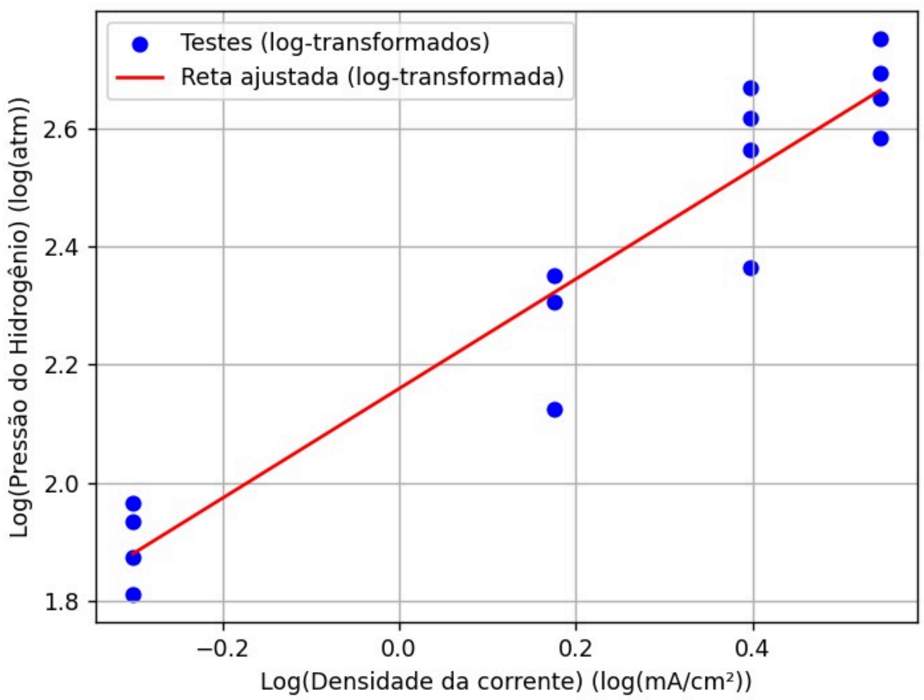

Figura 5. Reta da Regressão Linear em logarítmico

4. CONCLUSÃO

Com a introdução da escala logarítmica, conseguimos reduzir a diferença entre os valores maiores e menores, tornando a distribuição dos dados mais equilibrada e facilitando sua interpretação. Essa transformação é especialmente útil quando os dados apresentam uma variação muito ampla, pois ajuda a minimizar o impacto de valores extremos e torna a relação entre as variáveis mais clara.

Assim, a introdução da escala logarítmica não apenas melhora a representação gráfica e estatística dos dados, mas também permite que conclusões mais embasadas sejam tiradas, tornando o processo de análise mais robusto e eficiente, e neste caso, sugere que um aumento na densidade da corrente leva a um aumento na pressão do hidrogênio de forma não linear, seguindo um comportamento exponencial.

5. REFERÊNCIAS

WALPOLE, R. ; MYERS, R.; MYERS, S & YE, K. Probabilidade e Estatística para Engenharia e Ciências. Ed. Pearson, 2009.