Correlação e Regressão Linear

cognitiza

Your Data & AI school

Prof. Altino Dantas

Análise multivariada

- Até agora estudamos medidas numéricas para uma única variável (média, desvio padrão, variância etc);
- Eventualmente, precisamos analisar a relação entre duas ou mais variáveis;
- Para isso, duas medidas são comumente usadas:
  - Covariância;
  - Coeficiente de correlação.

Análise multivariada

- A covariância mede a força da relação linear entre duas variáveis;
- A fórmula da covariância de uma amostra é dada por:

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

onde,
x_i = valor da ocorrência i da variável x
y_i = valor da ocorrência i da variável y
~x = média da variável x
~y = média da variável y
N é a quantidade de exemplos de elementos das amostras

Observe que para o cálculo da covariância, os valores relativos às duas variáveis devem estar emparelhados

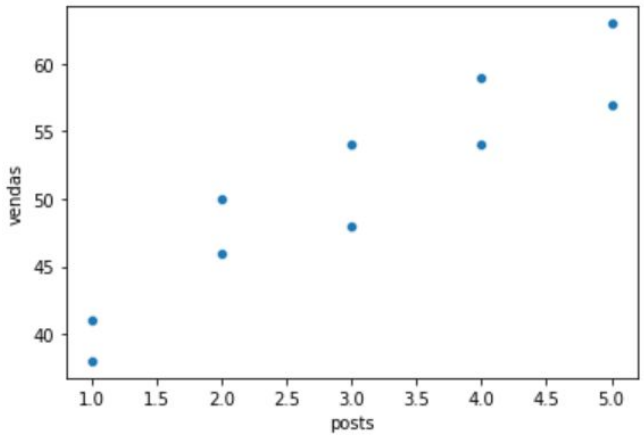
Análise multivariada

Exemplo: uma grande loja de varejo deseja investigar a relação entre o número de posts patrocinados no instagram (x) e o volume de vendas na semana seguinte à postagem (y). A tabela ao lado resume os dados:

semana	posts	vendas
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

Análise multivariada

```
sns.scatterplot(data=dados, x="posts", y="vendas")
```



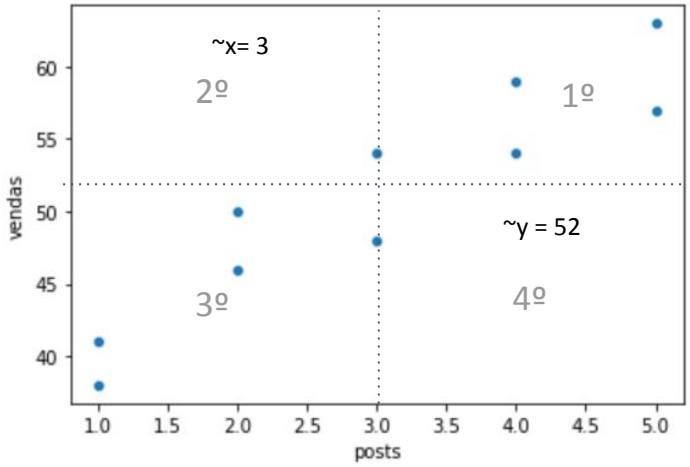
semana	posts	vendas
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

Pela análise do gráfico do volume de vendas y e número de posts patrocinados, aparentemente existe uma relação linear forte entre as duas variáveis. Fazendo as contas, nesse caso, obteremos **11 como valor de covariância**.

Análise multivariada

S_{xy} = covariância entre x e y

- Se S_{xy} é positivo, os pontos que exercem maior influência estão localizados nos quadrantes I e III, indicando uma relação linear positiva entre x e y.
- Se S_{xy} é negativo, os pontos que exercem maior influencia estão localizados nos quadrantes II e IV, indicando uma relação linear negativa entre x e y.
- Finalmente, se S_{xy} é nulo ou próximo de zero, os pontos se encontram distribuídos uniformemente ao longo dos 4 quadrantes, não indicando nenhuma relação linear entre x e y.



Análise multivariada

- Nesse exemplo, as médias das variáveis foram, 3 e 52, respectivamente para x e y, já a covariância foi 11;
- Pela análise do gráfico parece haver uma correlação entre as variáveis, mas o quando grande é o valor 11 para expressão tão relação?
- Problemas da covariância:
  - Dependente da unidade de grandeza utilizada nas variáveis envolvidas - se as vendas fossem expressas em milhares em vez de dezenas, ter-se-ia 11.000 e não 11;
  - É difícil dizer se o valor da variância é grande ou pequeno, uma vez que é uma medida de base absoluta.

Análise multivariada

- Chamamos de **correlação** uma medida normalizada da covariância no intervalo [-1,1]. Por exemplo, correlação de Pearson:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = Coeficiente de correlação
x_i = Valores da variável x na amostra
$\bar{x}$ = Média da variável x na amostra
y_i = Valores da variável y na amostra
$\bar{y}$ = Média da variável y na amostra

- Ou seja, o coef. de correlação é apenas a covariância “padronizada” pelos desvios-padrões das duas variáveis (x e y).

Análise multivariada

- Quanto mais próximo de +1, mais forte é a relação linear positiva entre x e y
- Quanto mais próximo de –1, mais forte é a relação linear negativa entre x e y
- Valores próximos de zero indicam ausência de relação linear entre x e y

Análise multivariada

- Em planilhas do google existe a função:

```
=CORREL([intervalo x],[intervalo y])
// retorna o valor da correlação entre os dois intervalos x e y
```

- Para dataframes do pandas, temos:

```
dataframe.corr()

// retorna a matriz de correlação entre as variáveis do dataframe

dataframe.cov()

// retorna a matriz de covariância entre as variáveis do dataframe
```

Análise multivariada

- Checando a fórmula do coeficiente de Pearson, para os dados anteriores:

```
1 # obten a convariância entre venda e
2 covariancia = dados.cov()['vendas']['posts']
3 # calcula os desvios-padrões das duas variávies
4 std_x = dados['posts'].std()
5 std_y = dados['vendas'].std()
6 # divide a covariância pelo produto dos desvios-padrões
7 print(covariancia / (std_x * std_y))
```

```
0.9304905807411791
```

```
1 dados.corr()
```

	semana	posts	vendas
semana	1.000000	0.073855	0.055532
posts	0.073855	1.000000	0.930491
vendas	0.055532	0.930491	1.000000

Exercício

1. Reproduzir o cenário apresentado calculando o coeficiente de correlação entre posts patrocinados e vendas utilizando:
   a. Planilha do google (Google Sheets).
   b. Dataframes Pandas no Python.
2. Apresentar gráficos de dispersão.

Exercício | python

- Carregue os dados do dataset “dados-delivery.csv”;
- Verifique a covariância entre as variáveis tempo de entrega real e avaliação;
- Verifique a correlação entre as variáveis;
- Apresente o gráfico de dispersão entre as duas variáveis.

Exercício | python

- Respite o experimento anterior, mas, dessa vez, analise apenas as avaliações com valores entre 8 e 10;

Exercício | python

- Crie uma nova coluna de dados representando a diferença entre o tempo real de entrega e o tempo previsto;
- Filtre apenas os registros em que houve atraso;
- Verifique a variância e a correlação entre a nova variável e a avaliação do pedido.

Regressão Linear simples

Regressão linear simples

- Correlação entre duas variáveis, cujo gráfico aproxima-se de uma linha;
- O gráfico cartesiano que representa essa linha é denominado diagrama de dispersão;
- Essa correlação pode ser sintetizada por uma reta chamada de reta de regressão e a equação que a representa é a equação de regressão.
- Assim, a equação de regressão pode ser usada como um estimador de valores desconhecidos;
- Por exemplo, para novos valores de x consegue-se estimar quais seriam os valores de y.

Regressão linear

Figura 16.1: Gráfico de dispersão de idade e reação ao estímulo, com reta ajustada.

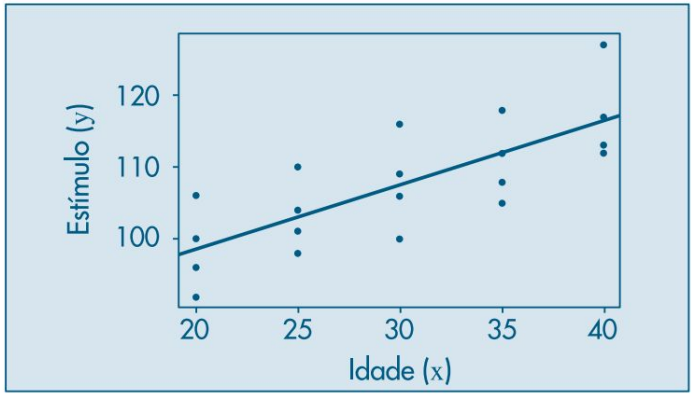


Tabela 15.1: Tempos de reação a um estímulo (*Y*) e acuidade visual (*Z*) de 20 indivíduos, segundo o sexo (*W*) e a idade (*X*).

Indivíduo	Y	W	X	Z
1	96	H	20	90
2	92	M	20	100
3	106	H	20	80
4	100	M	20	90
5	98	M	25	100
6	104	H	25	90
7	110	H	25	80
8	101	M	25	90
9	116	M	30	70
10	106	H	30	90
11	109	H	30	90
12	100	M	30	80
13	112	M	35	90
14	105	M	35	80
15	118	H	35	70
16	108	H	35	90
17	113	M	40	90
18	112	M	40	90
19	127	H	40	60
20	117	H	40	80

Regressão linear simples

- A reta de regressão é obtida a partir da observação de **n** amostras e pode ser representado pelo seguinte modelo:

$$E(Y|x) = \mu(x) = \alpha + \beta x,$$

- Ou seja, pretendemos estimar um *Y*, dado que *x* ocorra e para tanto, precisamos descobrir os coeficientes α *e* β;
- Uma forma de encontrar os valores do coeficientes é a através da minimização do erro do modelo sobre as amostras selecionadas.

Regressão linear simples

- Utilizando a derivada de uma função dos erros quadráticos em relação aos coeficientes da equação, definem-se as fórmulas para α *e* β como*:*

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}.$$

Regressão linear simples

- Voltemos ao exemplo das idades e tempos de reações a estímulos. Com os dados da tabela, podemos ajustar o modelo dispondo de:

y_i: tempo de reação do i-ésimo indivíduo,
x_i: idade do i-ésimo indivíduo,
e_i: desvio, $i = 1, 2, \ldots, 20$.

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}.$$

Pelos dados podemos calcular:

$$n = 20, \quad \sum y_i = 2.150, \quad \sum x_i = 600, \quad \sum x_i y_i = 65.400,$$
$$\bar{y} = 107{,}50, \quad \bar{x} = 30, \quad \sum x_i^2 = 19.000.$$

E substituir nas fórmulas de α e β

$$\hat{\beta} = \frac{65.400 - (20)(30)(107{,}50)}{19.000 - (20)(30)^2} = 0{,}90,$$
$$\hat{\alpha} = 107{,}50 - (0{,}90)(30) = 80{,}50,$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},$$

O que nos dá o modelo ajustado:

$$\hat{y}_i = 80{,}50 + 0{,}90x_i, \quad i = 1, 2, \ldots, 20.$$

Regressão linear simples

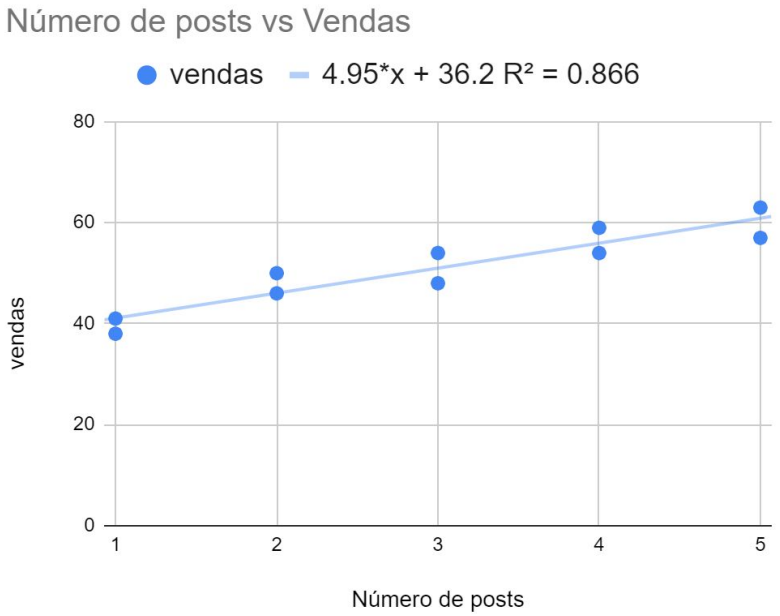
- Com esse modelo podemos prever, por exemplo, o tempo médio de reação para pessoas de 20 anos, que será indicado por ^y (20) e determinado por:

$$\hat{y}(20) = 80{,}50 + (0{,}90)(20) = 98{,}50.$$

- Veja que o modelo consegue estimar corretamente o tempo médio de resposta a estímulos de acordo com os dados, e melhor do que isso, o mesmo modelo pode ser usado para estimar o tempo de resposta para idades de fora do conjunto de amostras.

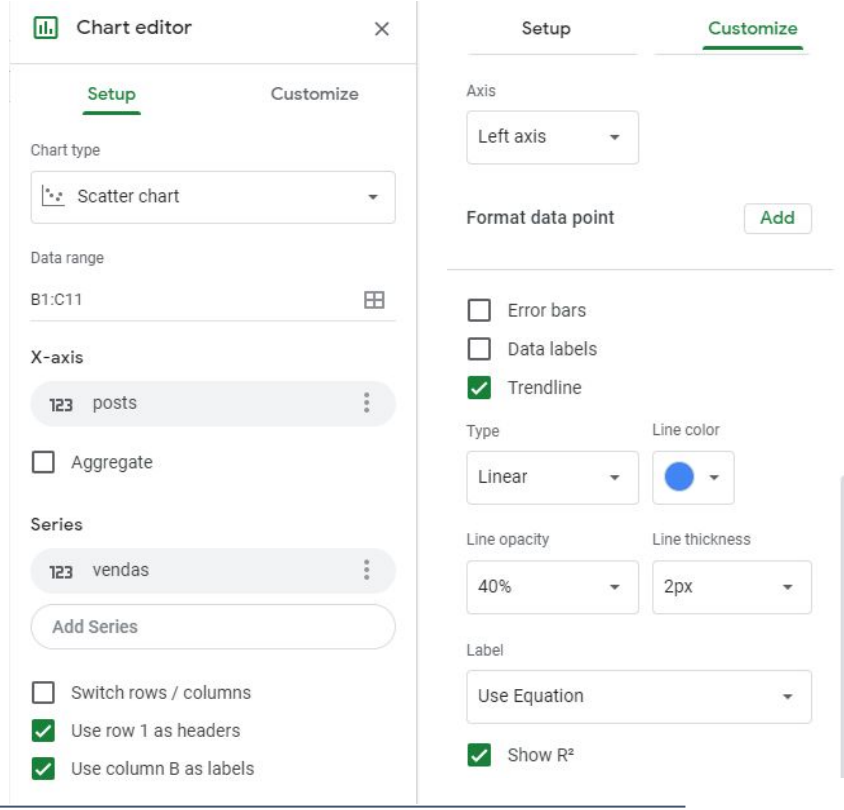
Encontrando os coeficientes no Google Sheets

- Como você percebeu, é trivial criar um procedimento computacional para calcular os coeficientes de uma regressão linear simples;
- Ainda assim, caso não queira fazê-lo, é possível obter rapidamente a reta de regressão com a ferramentas de gráficos do Google Sheets;
- Veja ao lado que $\alpha = 4,95$ e $\beta = 36,2$. Com isso, é possível estimar o valor das vendas para 8 posts patrocinados, 75,8



Encontrando os coeficientes no Google Sheets

- Selecione as duas colunas de dados que representam X e Y e insira um gráfico do tipo **Scatter** (gráfico de dispersão);
- Depois, em Customize-> series, marque **Trendline,** em **type** escolha **Linear** e no label **Use Equation;**
- **R^2** é uma métrica de qualidade do modelo;
- Os parâmetros podem ser obtidos diretamente com a fórmula: `=LINEST(C2:C11,B2:B11)`, onde `C2:C11,B2:B11` seriam os intervalos dos valores das variáveis Y e X



Exercício

- O diretor de uma cadeia de restaurantes de fast-food acredita que o volume de vendas de suas lojas seja positivamente relacionado à população situada em um raio de 3 km. Determine essa relação.
- Qual seria a estimativa de vendas para duas novas lojas cujas respectivas populações do entorno fossem de 5 mil e 30 mil.

Loja	População (mil)	Vendas (R$ mil)
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Obrigado.