Regressão Linear

Curso: Estatística e Probabilidade

Prof. Neemias Martins

PUC Campinas

neemias.silva@puc-campinas.edu.br
neemias.org

Correlação

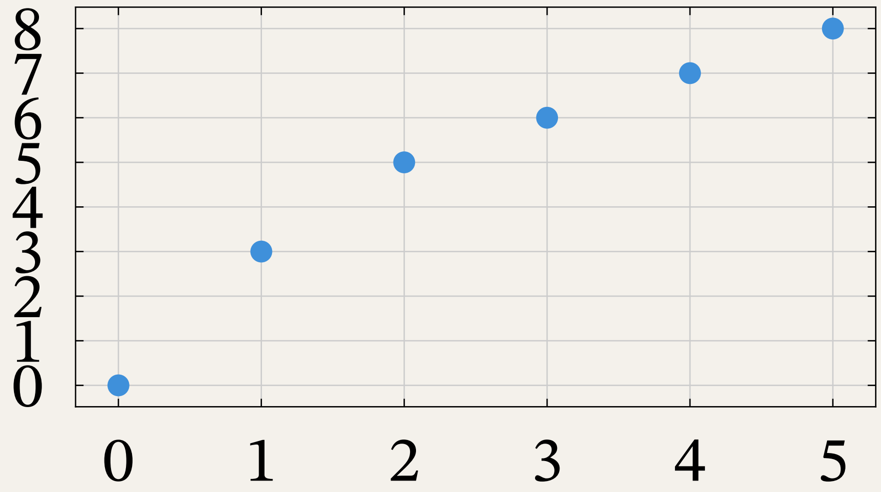
Uma *correlação* existe entre duas variáveis quando os valores de uma variável são associados aos valores da outra.

Uma correlação é *linear* se o gráfico de dispersão dos dados podem ser aproximados por uma reta.

Exemplo

Considere as variáveis. Obtenha o gráfico de dispersão.

x : Horas de estudo	0	1	2	3	4	5
y : Notas	0	3	5	6	7	8



Note que o pontos no gráfico de dispersão pode ser aproximado por uma reta. Logo a correlação é linear.

Coeficente de Correlação

O coeficiente r de correlação linear (também conhecido como coeficiente de Pearson de correlação linear) mede o quão correlacionados os valores das variáveis x e y de uma amostra estão. Ele é calculado através da expressão:

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - \left(\sum x\right)^2}\sqrt{n \sum y^2 - \left(\sum y\right)^2}}.$$

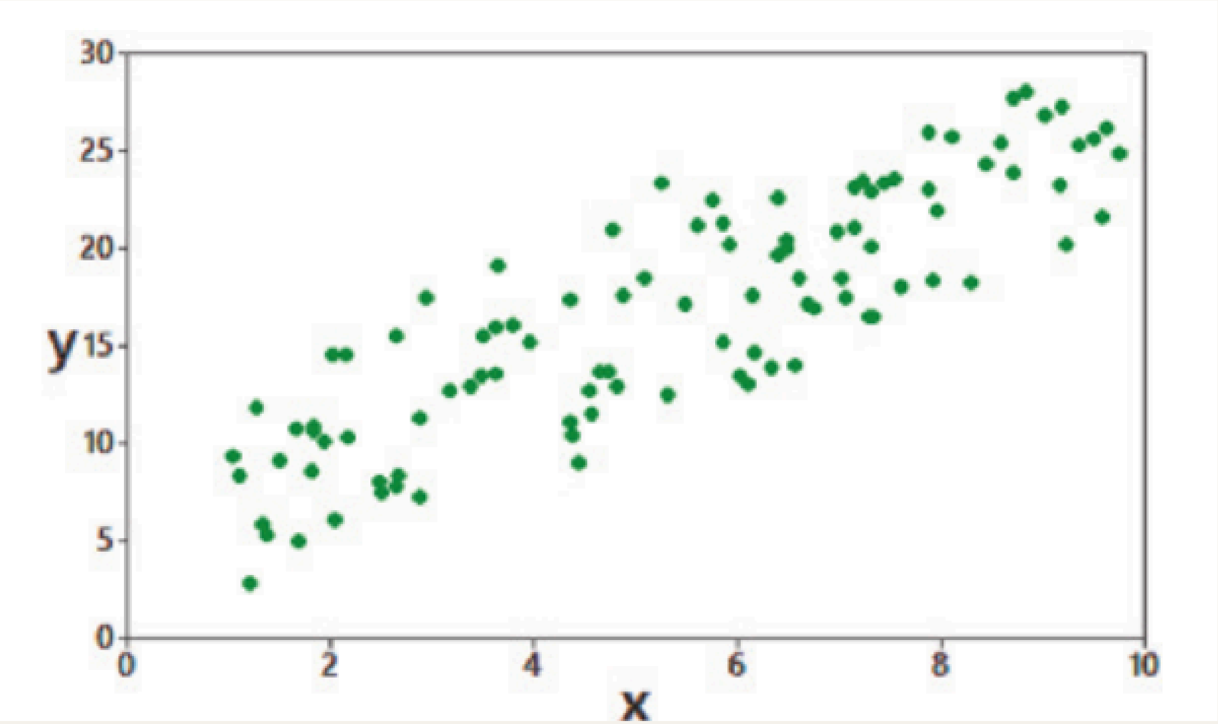
Tal coeficiente também pode ser descrito através do z-scores de x e y : z_x e z_y.

$$r = \sum \frac{z_x z_y}{n-1}.$$

Coeficente de Correlação

A correlação r é sempre um número entre -1 e 1.

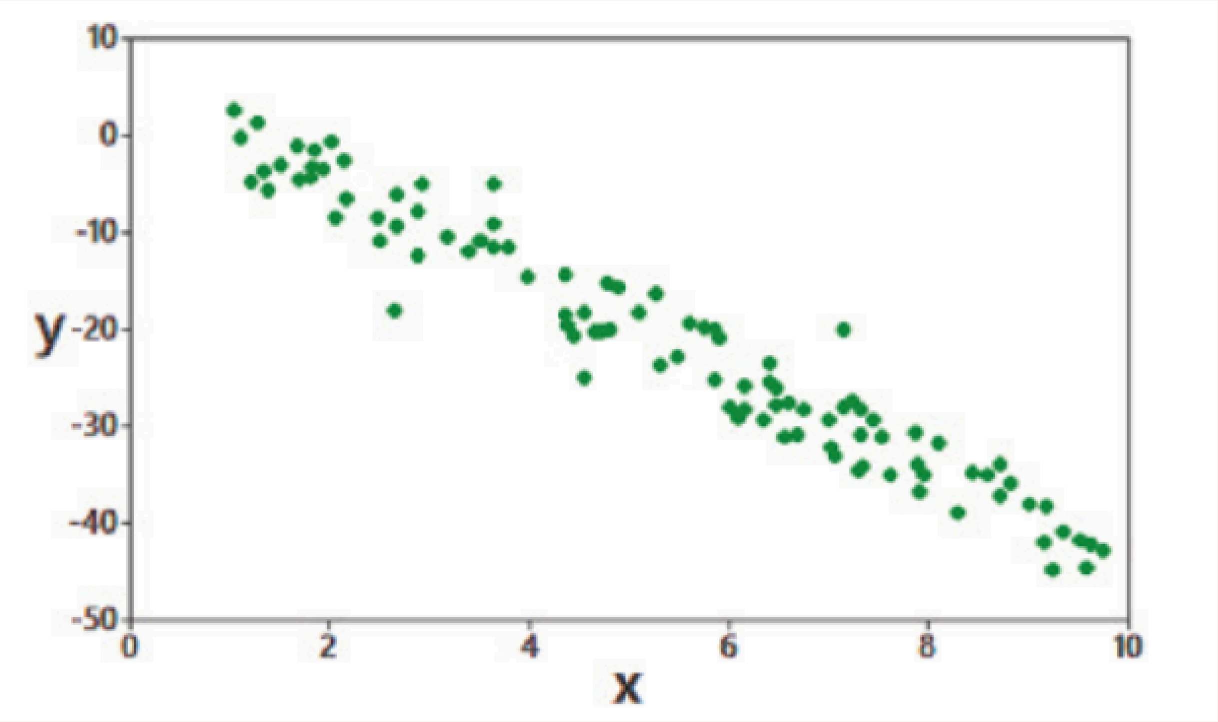
- Se $r = 1$, então há uma correlação positiva.



Coeficente de Correlação

A correlação r é sempre um número entre -1 e 1.

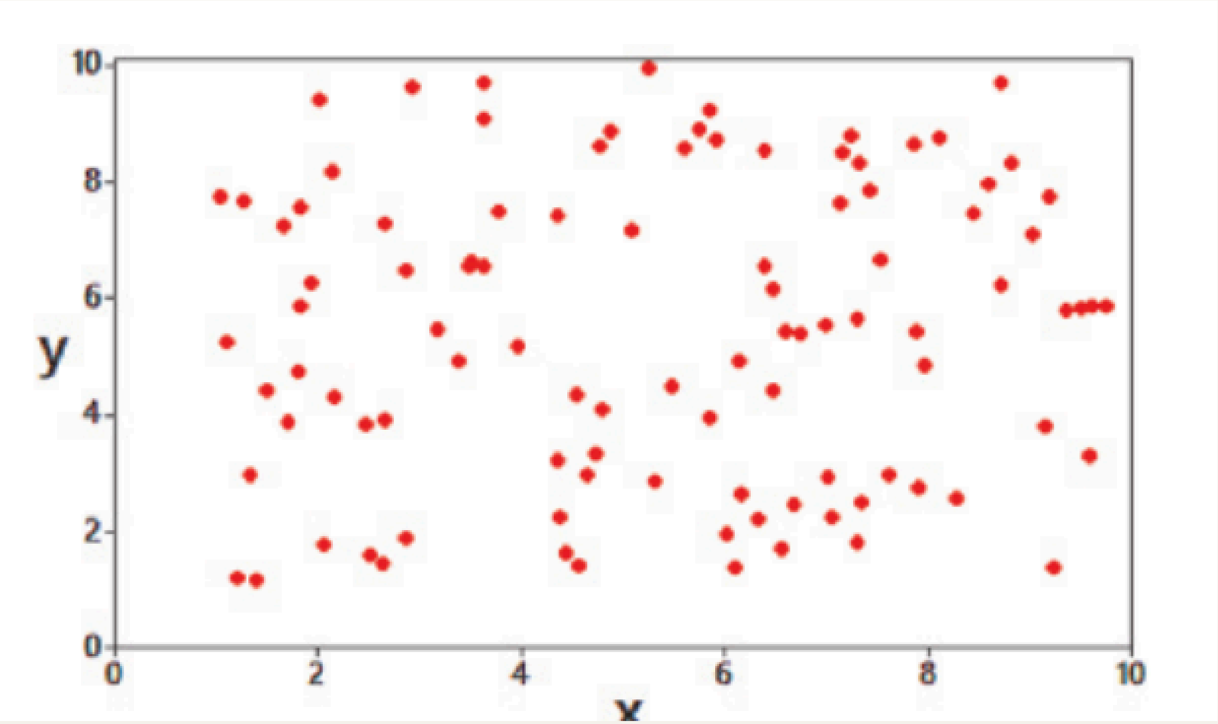
- Se $r = -1$, então há uma correlação negativa.



Coeficente de Correlação

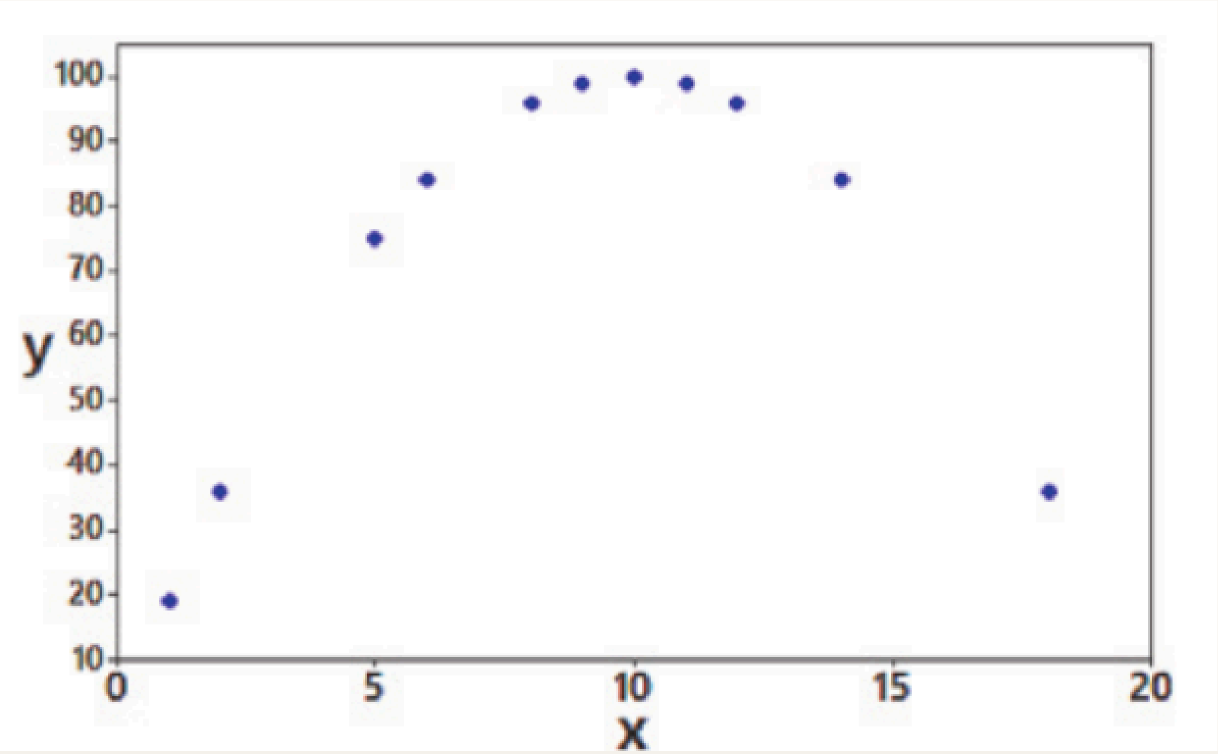
A correlação r é sempre um número entre -1 e 1.

- Se $r = 0$, então não há correlação.



Coeficente de Correlação

Exemplo de correlação não-linear.



Regressão Linear

Dado um conjunto de dados amostrais em pares ordenados, a *reta de regressão* (ou reta de melhor ajuste, ou reta dos mínimos quadrados) é a reta que "melhor" se ajusta ao gráfico de dispersão dos dados.

A equação de regressão $\hat{y} = b_0 + b_1 x$ descreve algebricamente a reta de regressão.

A equação de regressão expressa uma relação entre x (chamada de variável explicativa, ou variável preditora, ou variável independente) e $\hat{y}$ (chamada de variável de resposta ou variável dependente).

Regressão Linear

Para calcular a expressão $\hat{y} = b_0 + b_1 x$, usaremos as expressões:

- Inclinação da reta:

$$b_1 = r \cdot \frac{s_y}{s_x}$$

  em que r é o coeficiente de correlação, s_y é o desvio padrão da variável y e s_x é o desvio padrão da variável x.

- Intercepto:

$$b_0 = \overline{y} - b_1 \cdot \overline{x}$$

  em que $\overline{y}$ é a média da variável y e $\overline{x}$ é a média da variável x.

Exercícios

Exercício

1. A partir dos dados tabelados, obtenha o coeficiente de correlação linear e indique se a correlação é linear positiva, linear negativa ou nula.

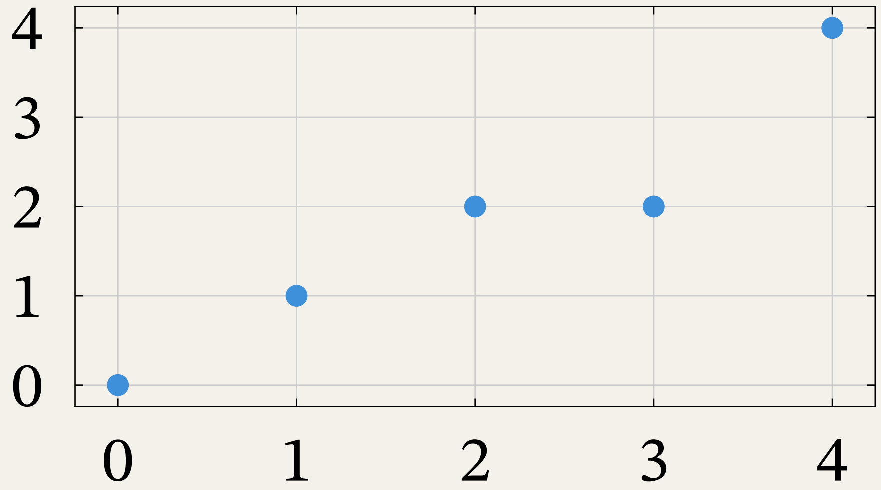
x = Número de leads

y = Número de vendas

x	0	1	2	3	4
y	0	1	2	2	4

Solução

Pelo diagrama de dispersão a correlação é linear:



Solução

Coeficiente de correlação:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}.$$

Vamos calcular separadamente cada somátorio em uma tabela.

Solução

x	y	x^2	y^2	$x \cdot y$
0	0	0	0	0
1	1	1	1	1
2	2	4	4	4
3	2	9	4	6
4	4	16	16	16
$\sum x = 10$	$\sum y = 9$	$\sum x^2 = 30$	$\sum y^2 = 25$	$\sum x \cdot y = 27$

Solução

$$
\begin{aligned}
r &= \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}} \\
&= \frac{5\cdot 27 - 10\cdot 9}{\sqrt{5\cdot 30 - (10)^2}\sqrt{5\cdot 25 - (9)^2}} \\
&= \frac{135 - 90}{\sqrt{150 - 100}\sqrt{125 - 81}} \\
&= \frac{45}{\sqrt{50}\sqrt{44}} \approx 0.959.
\end{aligned}
$$

A correlação é positiva.

Exercício

2. A partir dos mesmos dados do exemplo anterior, calcule os desvios padrão amostrais s_x e s_y e obtenha a reta de regressão linear.

x	0	1	2	3	4
y	0	1	2	2	4

Solução

Médias:

$$\bar{x} = \frac{\sum x}{n} = \frac{10}{5} = 2$$

$$\bar{y} = \frac{\sum y}{n} = \frac{9}{5} = 1.8$$

Solução

Desvio padrão:

$$
\begin{aligned}
s_x &= \sqrt{\frac{\sum (x - \overline{x})^2}{n-1}} \\
&= \sqrt{\frac{(0-2)^2 + (1-2)^2 + (2-2)^2 + (3-2)^2 + (4-2)^2}{4}} \\
&= \sqrt{\frac{4+1+0+1+4}{4}} \\
&= \sqrt{\frac{10}{4}} = \sqrt{2.5} \approx 1.581
\end{aligned}
$$

Solução

Desvio padrão:

$$s_y = \sqrt{\frac{\sum\left(y - \bar{y}\right)^2}{n-1}}$$

$$= \sqrt{\frac{(0-1.8)^2 + (1-1.8)^2 + (2-1.8)^2 + (2-1.8)^2 + (4-1.8)^2}{4}}$$

$$= \sqrt{\frac{3.24 + 0.64 + 0.04 + 0.04 + 4.84}{4}}$$

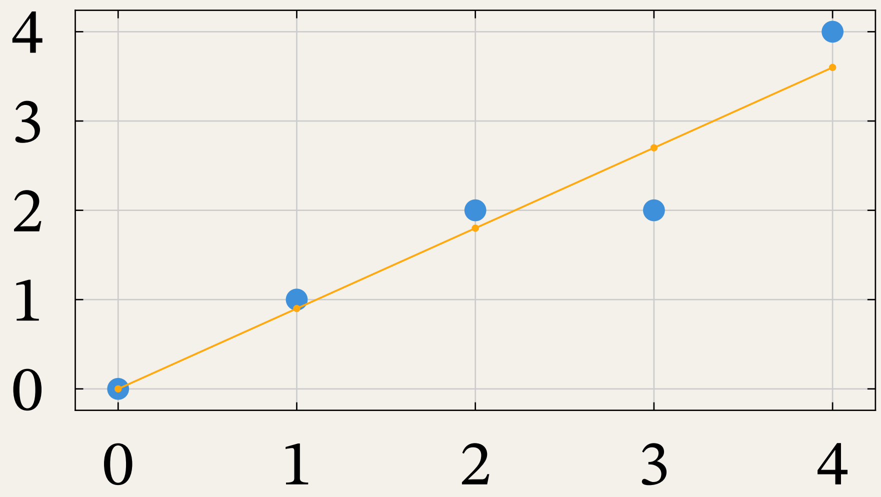
$$= \sqrt{\frac{8.80}{4}} = \sqrt{32.2} \approx 1.483$$

Solução

- $b_1 = r \cdot \frac{s_y}{s_x} = 0.959 \cdot \frac{1.483}{1.581} = 0.959 \cdot 0.938 \approx 0.9.$
- $b_0 = \overline{y} - b_1 \cdot \overline{x} = 1.8 - 0.9 \cdot 2 = 0.$

Reta de regressão $\hat{y} = b_0 + b_1 \cdot x$:

$$\hat{y} = 0.9 \cdot x$$



Exercício

3. Duas variáveis x e y possuem as seguintes propriedades:

- Correlação: $r = -0.8$
- Desvios padrão $s_x = 5.01,\ s_y = 3.21$
- Médias $\overline{x} = 6,\ \overline{y} = 8$.

a) Obtenha a reta de regressão.

b) Qual o valor de y esperado para $x = 10$?

Solução

- $b_1 = r \cdot \frac{s_y}{s_x} = -0.8 \cdot \frac{3.21}{5.01} = -0.8 \cdot 0.64 \approx -0.5$
- $b_0 = \overline{y} - b_1 \cdot \overline{x} = 8 - (-0.5) \cdot 6 = 8 + 3 = 11.$

Portanto a reta de regressão é

$$\hat{y} = 11 - 0.5x$$

Quando $x = 10$, o valor esperado para y é

$$y = 11 - 0.5 \cdot 10 = 11 - 5 = 6.$$

Exercício

4. Uma empresa avalia a projeção de vendas y (em milhares de unidades) em relação ao investimento em publicidade x (em milhares de reais). Os dados do último semestre apresentam média do investimento em publicidade $\overline{x} = 20$ e desvio padrão $s_x = 5$, média das vendas $\overline{y} = 50$ e desvio padrão $s_y = 8$, e correlação entre investimento e vendas é $r = 0.75$. Usando regressão linear, estime o número esperado de vendas se a empresa investir 30 mil reais em publicidade.

Bons Estudos!