Correlação e Regressão

Prof. Antonio Estanislau Sanches

2018

Objetivo

Estudar a relação entre duas variáveis quantitativas.

Exemplos:

Idade e altura das crianças

Tempo de prática de esportes e ritmo cardíaco

Tempo de estudo e nota na prova

Taxa de desemprego e taxa de criminalidade

Expectativa de vida e taxa de analfabetismo

Investigaremos a presença ou ausência de relação linear sob dois pontos de vista:

a) **Quantificando a força dessa relação: correlação.**
b) **Explicitando a forma dessa relação: regressão.**

Representação gráfica de duas variáveis quantitativas: Diagrama de dispersão

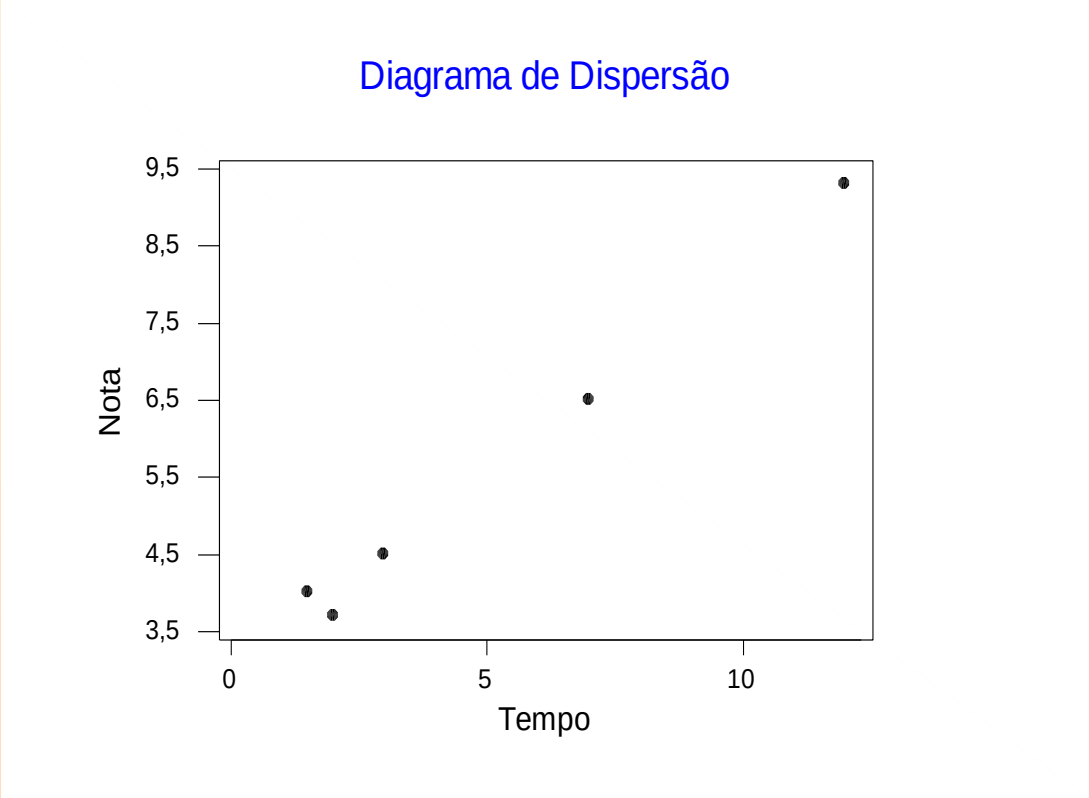
Exemplo 1: nota da prova e tempo de estudo

X : tempo de estudo (em horas)

Y : nota da prova

Pares de observações (X_i , Y_i) para cada estudante

Tempo(X)	Nota(Y)
3,0	4,5
7,0	6,5
2,0	3,7
1,5	4,0
12,0	9,3

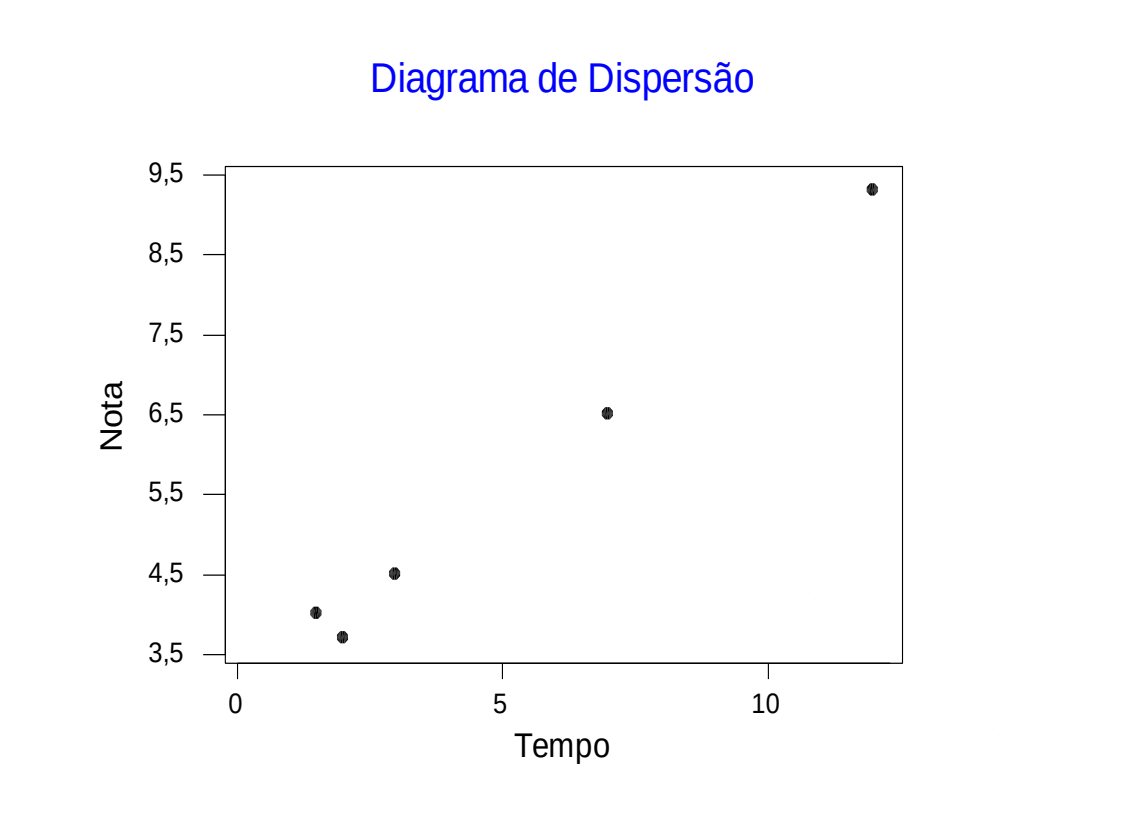


Exemplo 1: nota da prova e tempo de estudo

Tempo(X)	Nota(Y)
3,0	4,5
7,0	6,5
2,0	3,7
1,5	4,0
12,0	9,3

Criando o gráfico de dispersão da série:

1. Digitar os dados e os títulos em um arquivo excel (*pe: B1:C6*);
2. Selecionar os dados com os títulos (*B1:C6*);
3. Escolher a aba INSERIR e clicar no gráfico DISPERSÃO (*primeira opção*)
4. Verifique que o gráfico está quase pronto.... (*fazer alguns acertos finais*)



Coeficiente de correlação linear

É uma medida que avalia o quanto a "nuvem de pontos" no diagrama de dispersão aproxima-se de uma reta.

O coeficiente de correlação linear de Pearson é dado por:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)S_X S_Y},$$

sendo que,

$\overline{X}$ **e** $\overline{Y}$ **são as médias amostrais de X e Y, respectivamente,**

S_X **e** S_Y **são os desvios padrão de X e Y, respectivamente.**

Usar essa fórmula ou a

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)S_X S_Y},$$

Fórmula alternativa:

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{(n-1)S_X S_Y}$$

No exemplo:

Tempo (X)	Nota (Y)	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$
3,0	4,5	-2,1	-1,1	2,31
7,0	6,5	1,9	0,9	1,71
2,0	3,7	-3,1	-1,9	5,89
1,5	4,0	-3,6	-1,6	5,76
12,0	9,3	6,9	3,7	25,53
25,5	28,0	0	0	41,2

$\bar{X} = 5,1$ $\bar{Y} = 5,6$

$$S_x^2 = \frac{(-2,1)^2 + \ldots + (6,9)^2}{4} = \frac{78,2}{4} = 19,55 \Rightarrow S_x = 4,42$$

$$S_y^2 = \frac{(-1,1)^2 + \ldots + (3,7)^2}{4} = \frac{21,9}{4} = 5,47 \Rightarrow S_y = 2,34$$

Então,

$$r = \frac{41,2}{4 \,.\, 4,42 \,.\, 2,34} = 0,9959$$

Propriedade: $-1 \leq r \leq 1$

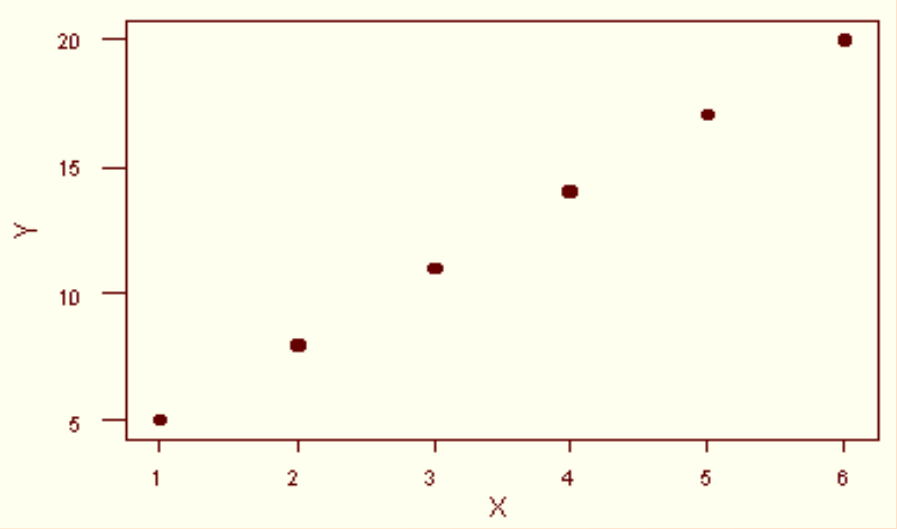
Casos particulares:

$r = 1 \Rightarrow$ correlação linear positiva e perfeita

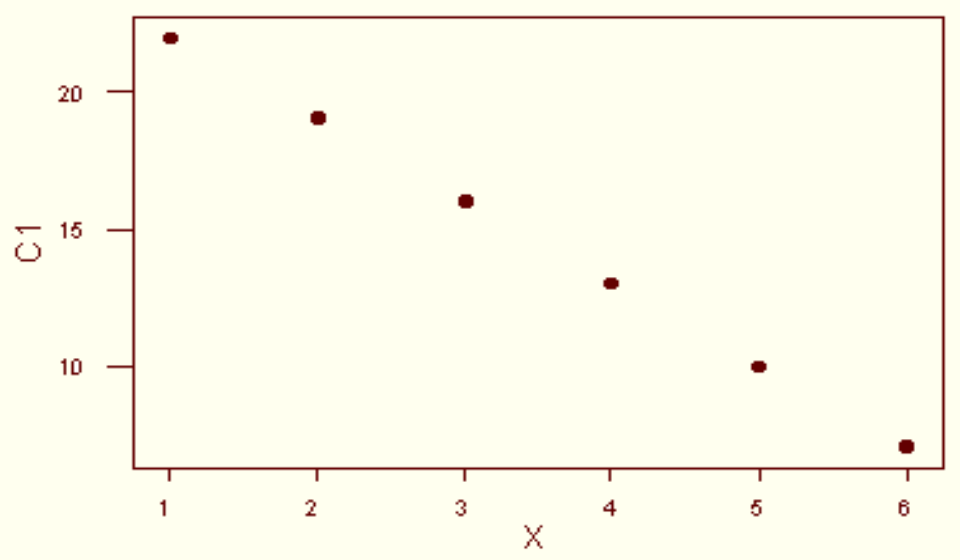
$r = -1 \Rightarrow$ correlação linear negativa e perfeita

$r = 0 \Rightarrow$ inexistência de correlação linear

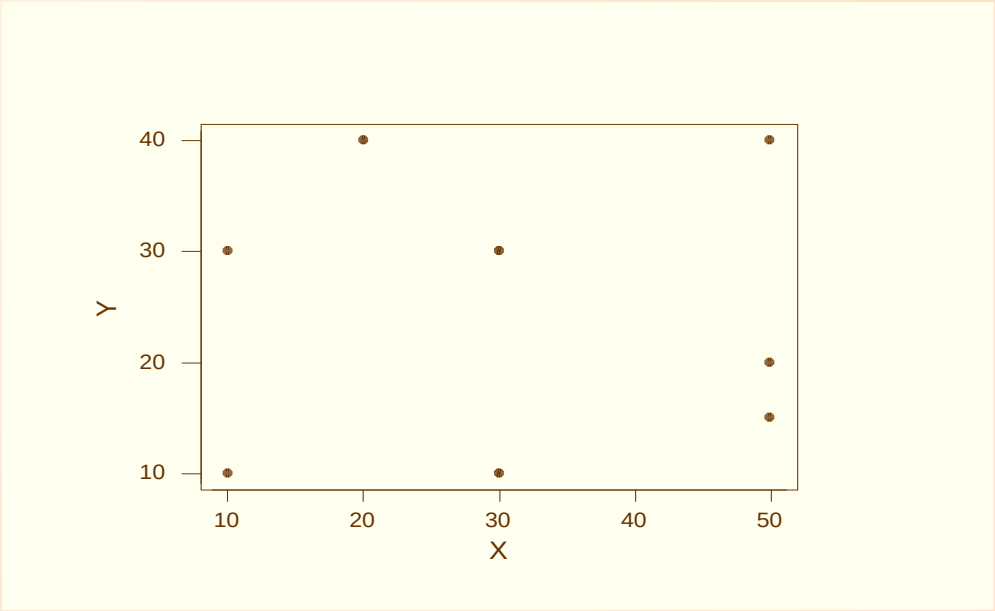
r = 1, correlação linear positiva e perfeita

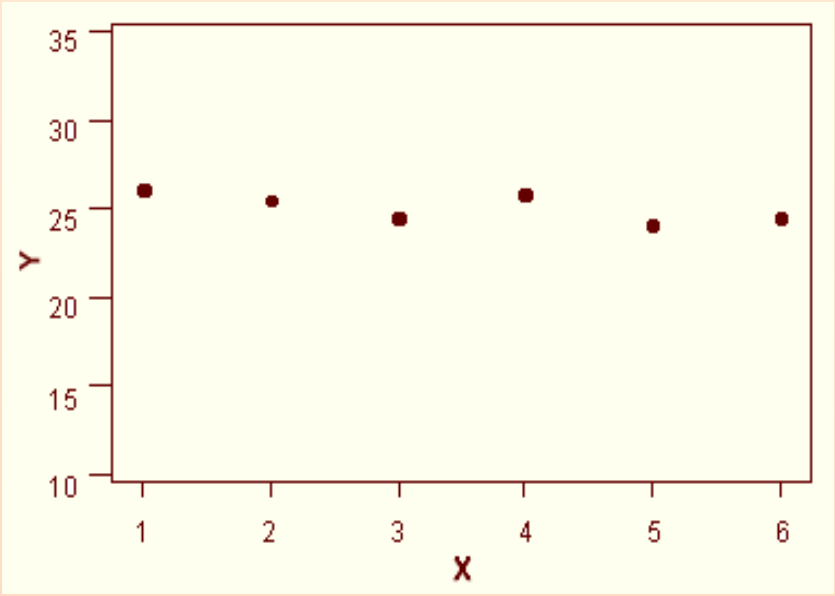


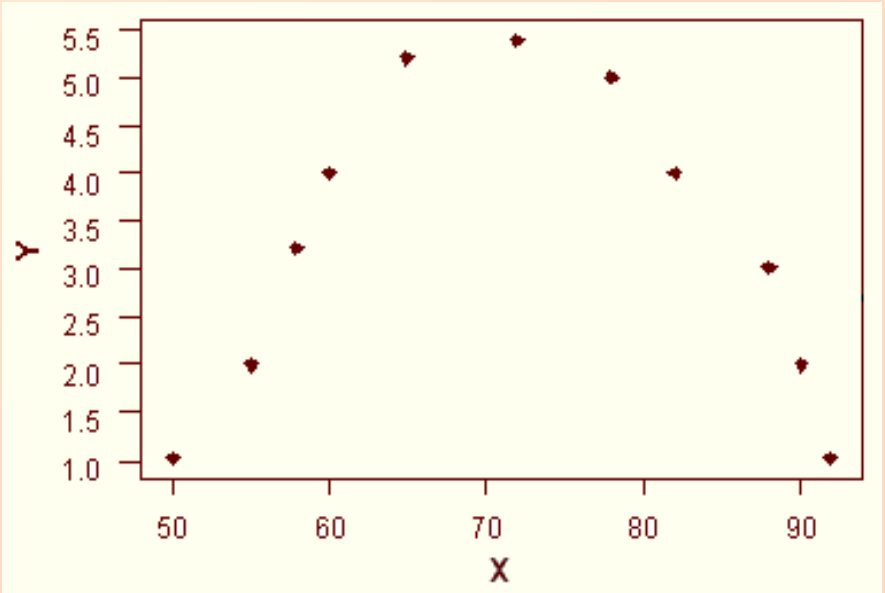
r = -1, correlação linear negativa e perfeita



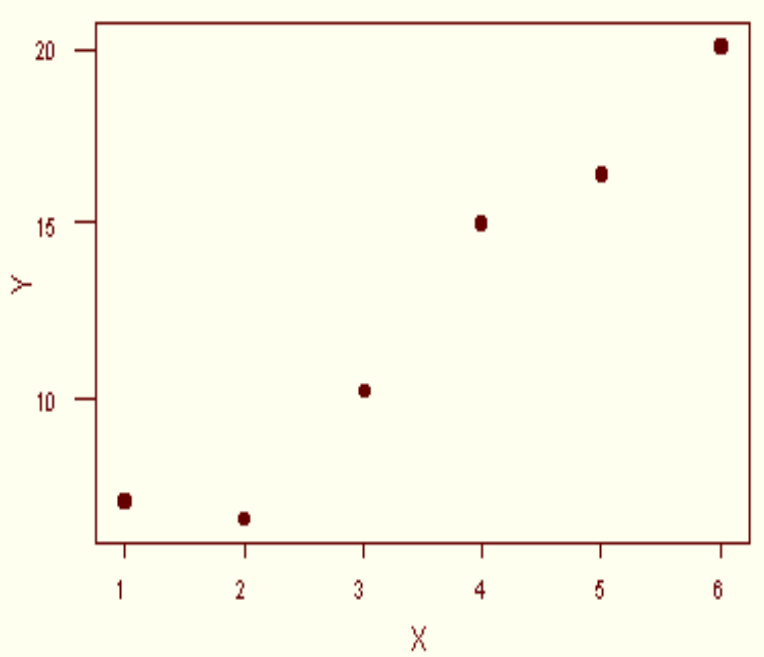
$r \cong 0$


40
30
20
10
Y
10
20
30
40
50
X

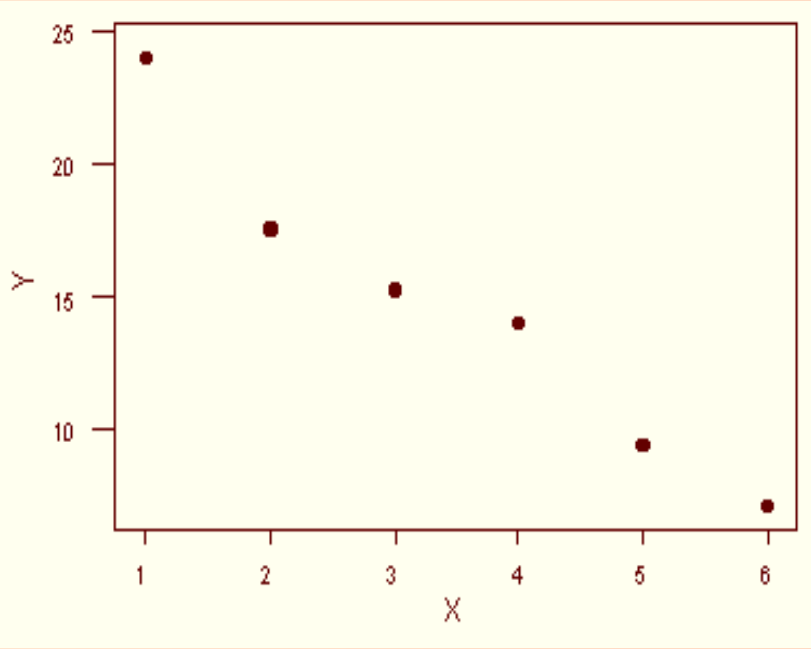

35
30
25
20
15
10
Y
1
2
3
4
5
6
X


5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Y
50
60
70
80
90
X

$r \cong 1$



$r \cong -1$



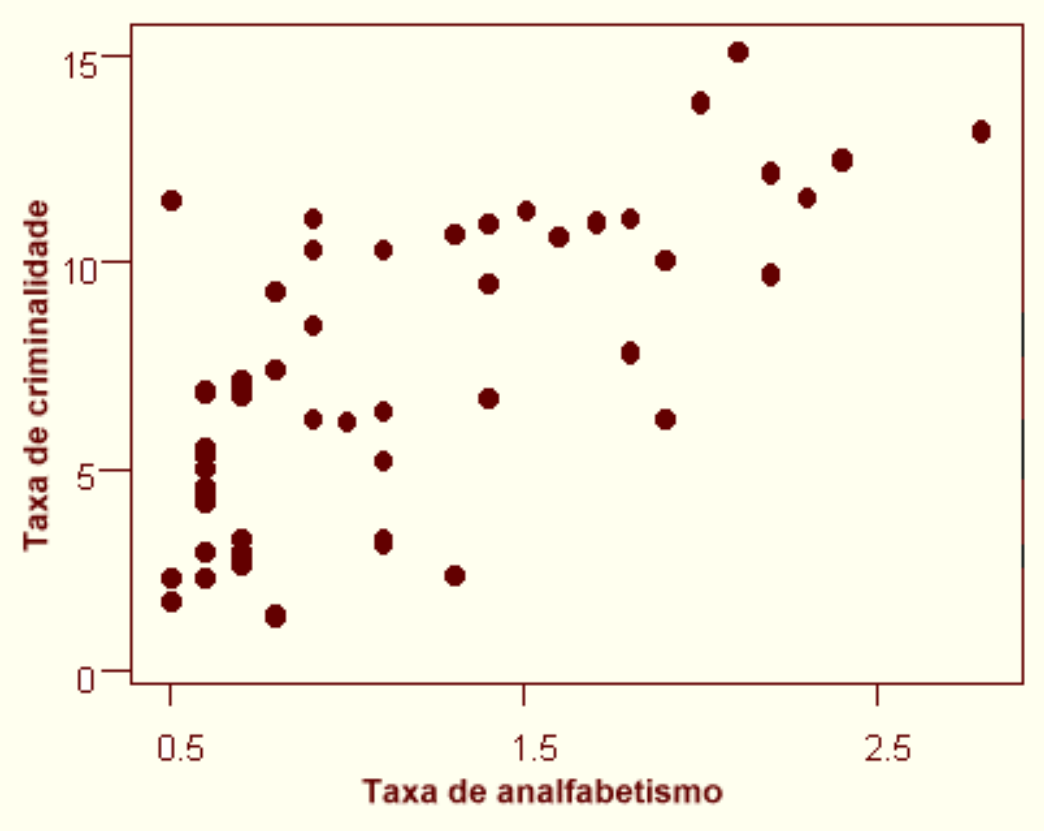
Exemplo 2: criminalidade e analfabetismo

Considere as duas variáveis observadas em 50 estados norte-americanos.

Y: taxa de criminalidade

X: taxa de analfabetismo

Diagrama de dispersão



Podemos notar que, conforme aumenta a taxa de analfabetismo (X), a taxa de criminalidade (Y) tende a aumentar. Nota-se também uma tendência linear.

Cálculo da correlação

$\bar{Y}$= 7,38 (média de Y) e S_Y = 3,692 (desvio padrão de Y)

$\bar{X}$= 1,17 (média de X) e S_x = 0,609 (desvio padrão de X)

$\Sigma X_i Y_i$ = 509,12 ; sendo, n = 50

Correlação entre X e Y:

$$r = \frac{\sum_{1}^{n} X_i\, Y_i - n\,\bar{X}\,\bar{Y}}{(n-1)\,S_x\,S_y}$$

$$r = \frac{509{,}12 - 50 \cdot 7{,}38 \cdot 1{,}17}{49 \cdot 3{,}692 \cdot 0{,}609} = \frac{77{,}39}{110{,}17} = 0{,}702$$

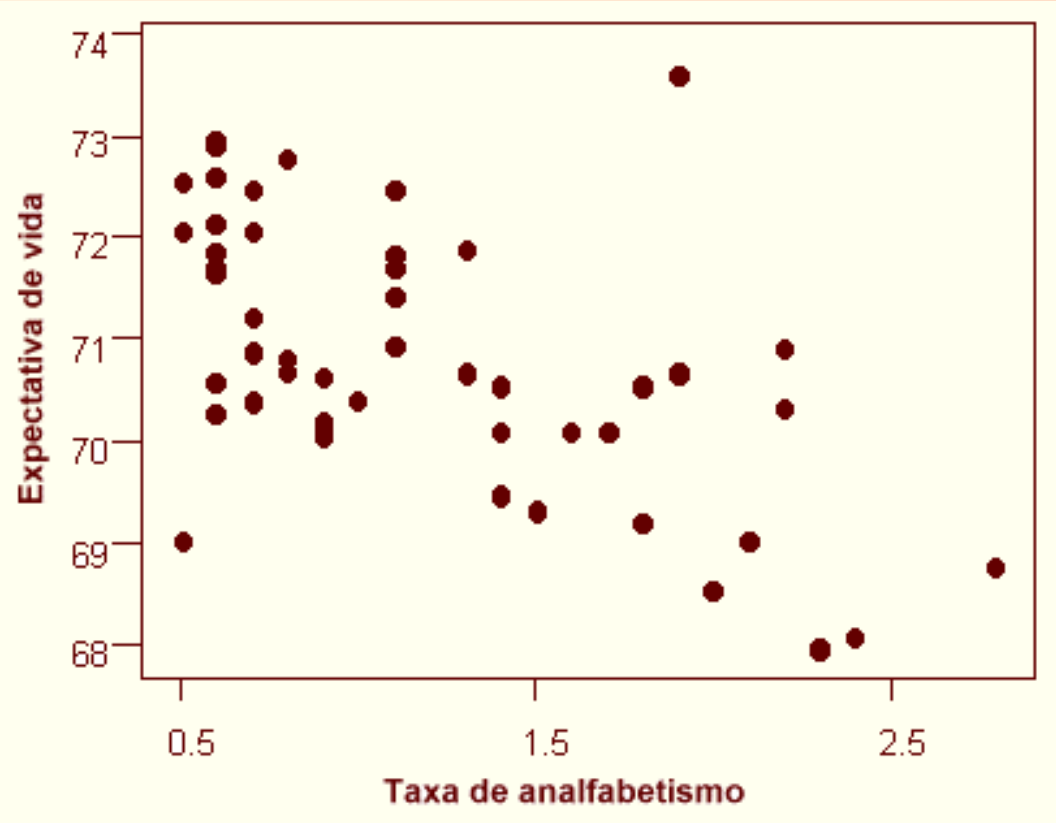
Exemplo 3: expectativa de vida e analfabetismo

Considere as duas variáveis observadas em 50 estados norte-americanos.

Y: expectativa de vida

X: taxa de analfabetismo

Diagrama de dispersão



Podemos notar que, conforme aumenta a taxa de analfabetismo (X), a expectativa de vida (Y) tende a diminuir. Nota-se também uma tendência linear.

Cálculo da correlação

$\bar{Y}$= 70,88 (média de Y) e S_Y = 1,342 (desvio padrão de Y)

$\bar{X}$= 1,17 (média de X) e S_x = 0,609 (desvio padrão de X)

$\Sigma X_i Y_i$ = 4122,8 ; sendo n= 50

Correlação entre X e Y:

$$r = \frac{\sum_{1}^{n} X_i\, Y_i - n\,\bar{X}\,\bar{Y}}{(n-1)\,S_x\,S_y}$$

$$r = \frac{4122{,}8 - 50 \cdot 70{,}88 \cdot 1{,}17}{49 \cdot 1{,}342 \cdot 0{,}609} = \frac{-23{,}68}{40{,}047} = -0{,}59$$

Reta ajustada:

$$\hat{Y} = a + bX$$

O que são a e b?

a: intercepto

b: inclinação

Interpretação de b:

Para cada aumento de uma unidade em X, temos um aumento médio de b unidades em Y.

Reta ajustada (método de mínimos quadrados)

Os coeficientes a e b são calculados da seguinte maneira:

$$b = \frac{\sum_{i=1}^{n} X_i Y_i - n.\overline{X}.\overline{Y}}{(n-1).S_X^2}$$

e

$$a = \overline{Y} - b.\overline{X}$$

Para os valores: $\overline{Y} = 7,38$ $\Sigma X_i Y_i = 509,12$

$\overline{X} = 1,17$ $n = 50$ $S_x = 0,609$

Calcular “a” ; “b” ; equação da reta e $\hat{y}$ p/ X=1,50:

a = 2,398 ; b = 4,258 ; $\hat{y}$=2,398 + 4,258 X e $\hat{y}_{X=1,50}$= 8,79

No exemplo 2,

a reta ajustada é:

$$\hat{y} = 2{,}398 + 4{,}258\,X$$

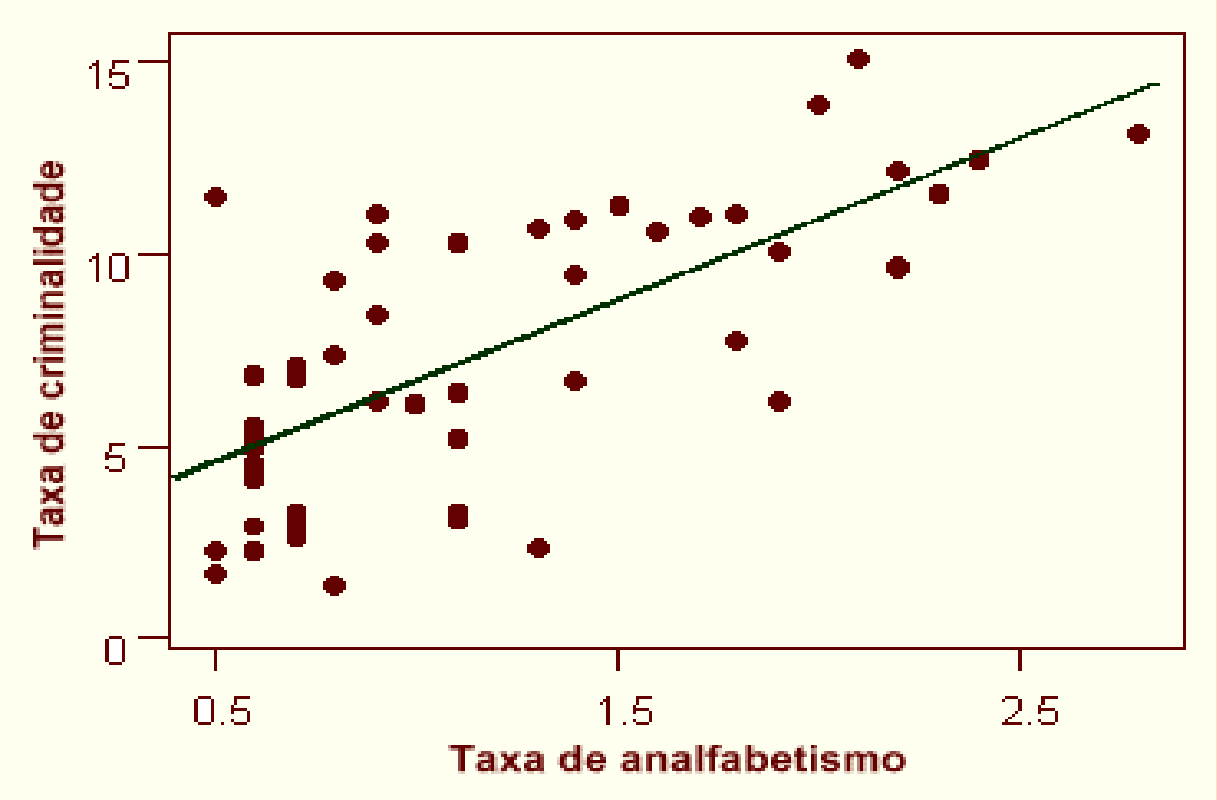
^
Y : valor predito para a taxa de criminalidade

X : taxa de analfabetismo

Interpretação de b:

Para um aumento de uma unidade na taxa do analfabetismo (X), a taxa de criminalidade (Y) aumenta, em média, 4,258 unidades.

Graficamente, temos



Como desenhar a reta no gráfico?

No exemplo 3,

Uma outra reta ajustada é:

$$\hat{Y} = 72{,}395 - 1{,}296\,X$$

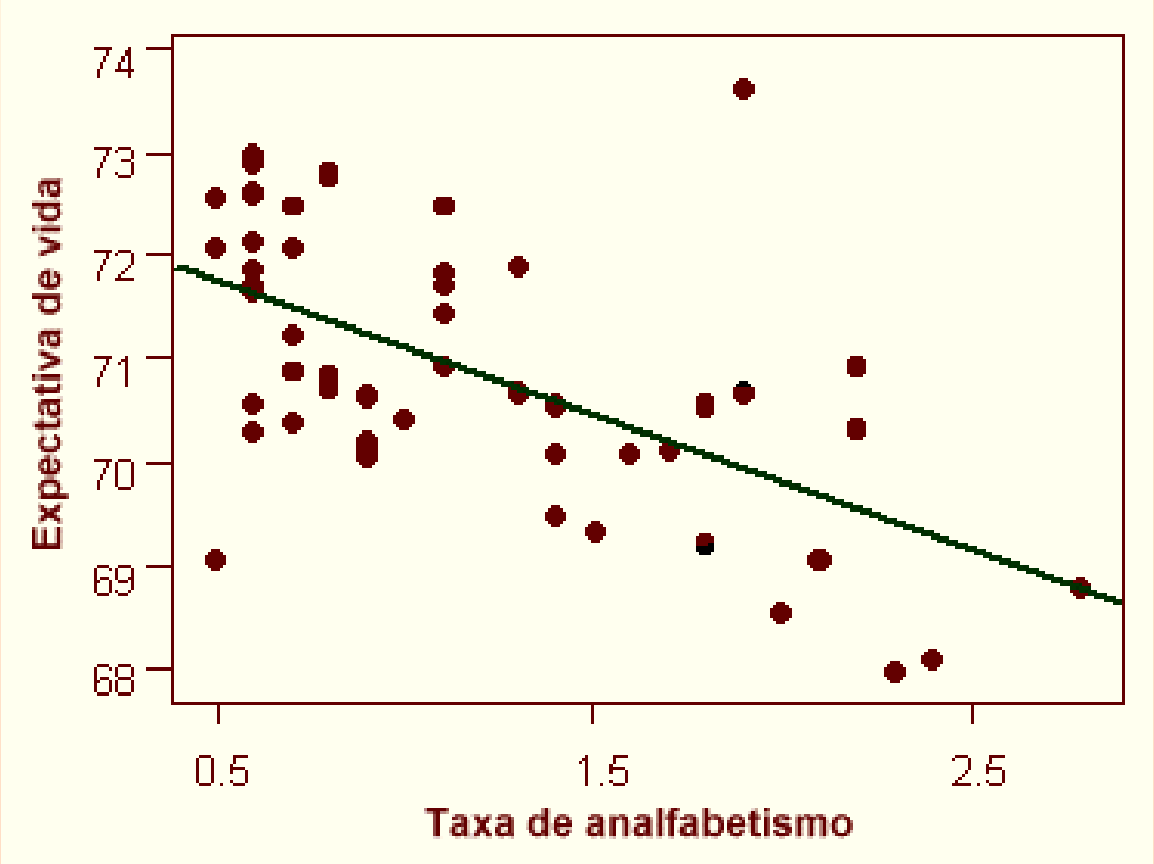
$\hat{Y}$: valor predito para a expectativa de vida

X : taxa de analfabetismo

Interpretação de b:

Para um aumento de uma unidade na taxa do analfabetismo (X), a expectativa de vida (Y) diminui, em média, 1,296 anos.

Graficamente, temos



Exemplo 4: consumo de cerveja e temperatura

Y: consumo de cerveja diário por mil habitantes, em litros.

X: temperatura máxima (em ºC).

As variáveis foram observadas em nove localidades com as mesmas características demográficas e sócio-econômicas.

Dados:

Localidade	Temperatura (X)	Consumo (Y)
1	16	290
2	31	374
3	38	393
4	39	425
5	37	406
6	36	370
7	36	365
8	22	320
9	10	269

Calcule:

a) *r* = Coef. Correl Person;
b) reta de regressão e
c) consumo previsto para uma temperatura de 25ºC

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)S_X S_Y}$$

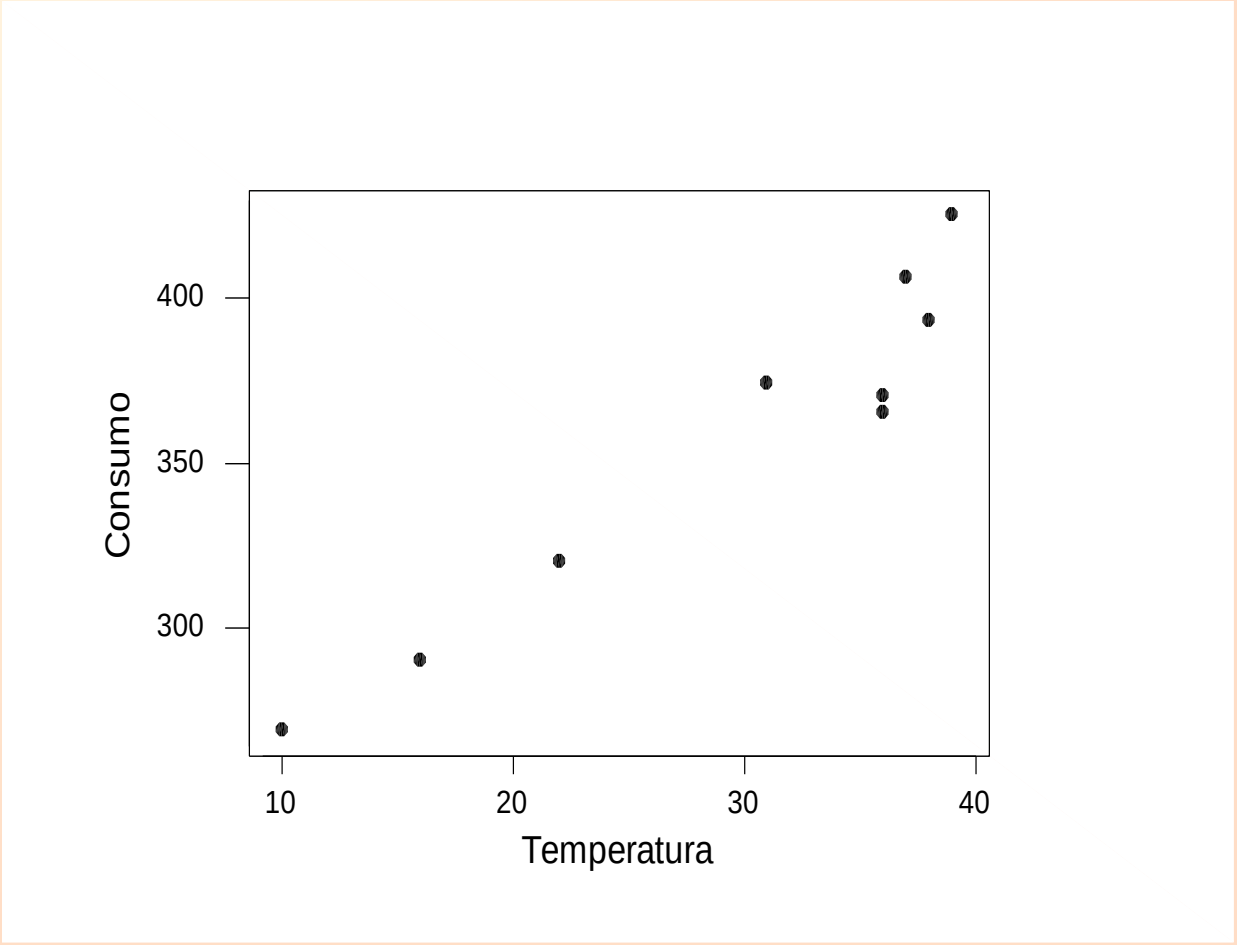
$$b = \frac{\sum_{i=1}^{n} X_i Y_i - n.\overline{X}.\overline{Y}}{(n-1).S_X^2}$$

$$a = \overline{Y} - b.\overline{X}$$

$$\hat{Y} = a + bX$$

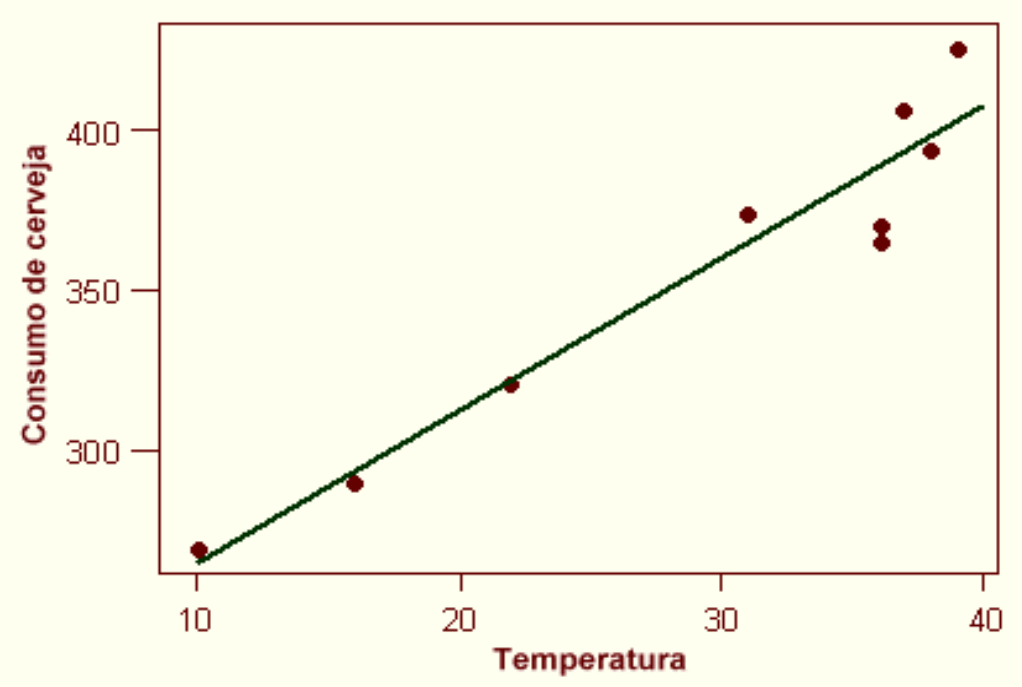
X	Y
°C (temp)	Consumo
16	290
31	374
38	393
39	425
37	406
36	370
36	365
22	320
10	269

Diagrama de dispersão



A correlação entre X e Y é r = 0,962.

A reta ajustada é: $\hat{Y} = 217{,}37 + 4{,}74\,X$



Qual a interpretação de b?

Aumentando-se um grau de temperatura (X), o consumo de cerveja (Y) aumenta, em média, 4,74 litros por mil habitantes.

Qual o consumo previsto para uma temperatura de 25ºC?

$$\hat{Y} = 217{,}37 + 4{,}74.25 = 335{,}83 \text{ litros}$$

F I M