Análise Bidimensional

Associação entre variáveis qualitativas

Tabelas de Contigência

Podemos construir tabelas de freqüências conjuntas (*tabelas de contingência*), relacionando duas variáveis qualitativas.

Exemplo 1(Pulse): Há indícios de associação entre Sexo e Hábito de fumar?

	Hábito de Fumar		
Sexo	Fuma	Não Fuma	Total
Masculino	20	37	57
Feminino	8	27	35
Total	28	64	92

Qual é o significado dos valores desta tabela?

Como concluir?

```
> dados<-read.csv("pulse.csv",dec=",")
```

ou

```
> dados<-read.csv("C:/WORKS/MAE116-Biology/pulse.csv",dec=",")
> names(dados)
[1] "Pulse1"   "Pulse2"   "Ran"      "Smokes"   "Sex"      "Height"   "Weight"
[8] "Activity"
> table(dados[,c(4,5)]) # ou pode ser # table(dados$Smokes,dados$Sex)

            Sex
Smokes    1    2
        1  20   8
        2  37   27
```

Verificar associação através da:

- porcentagem segundo as colunas, ou

- porcentagem segundo as linhas.

	Hábito de Fumar		
Sexo	Fuma	Não Fuma	Total
Masculino	71,43%	57,81%	61,96%
Feminino	28,57%	42,19%	38,04%
Total	100%	100%	100%

Qual é o significado dos valores desta tabela?

Como concluir? Será que saber se uma pessoa é do sexo masculino ou feminino ajuda a prever se ela é fumante ? Ajuda muito ? Não ajuda muito ? E vice-versa ?

Tabela de porcentagem segundo as colunas

```
> prop.table(table(dados[,c(4,5)]),2)
            Sex
Smokes           1         2
     1   0.3508772 0.2285714
     2   0.6491228 0.7714286
```

Tabela de porcentagem segundo as linhas

```
> prop.table(table(dados[,c(4,5)]),1)
            Sex
Smokes           1         2
     1   0.7142857 0.2857143
     2   0.5781250 0.4218750
```

Associação entre variáveis quantitativas

Correlação e Regressão

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

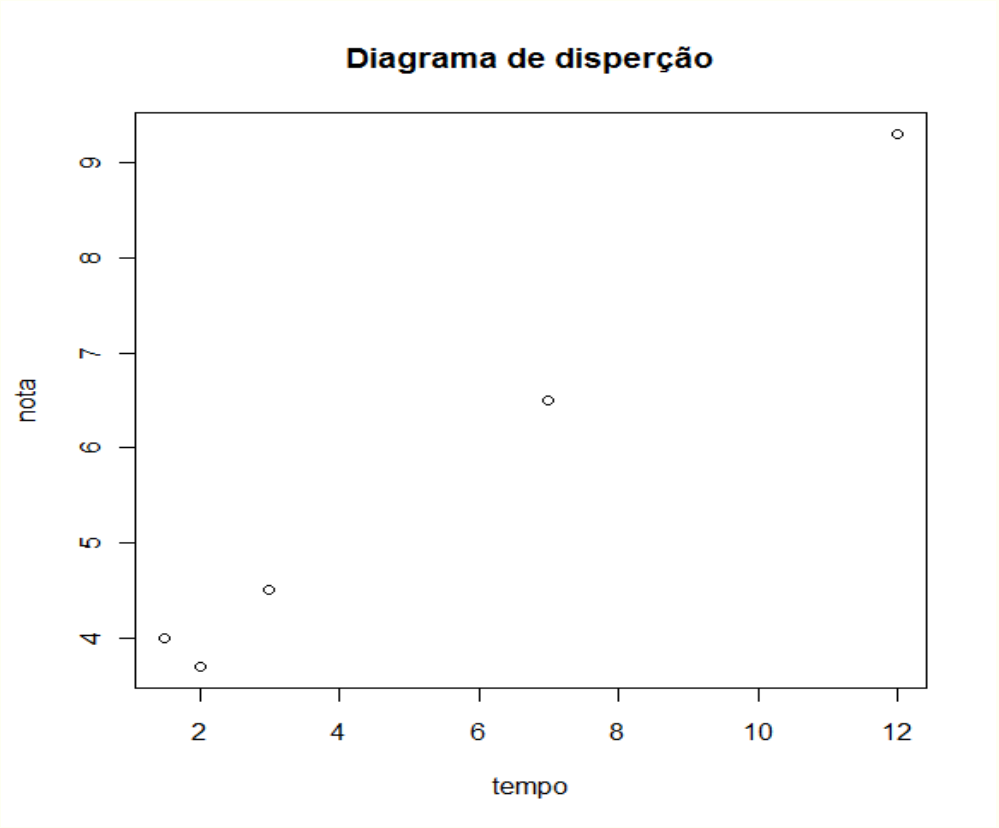
Exemplo 1: nota da prova e tempo de estudo

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

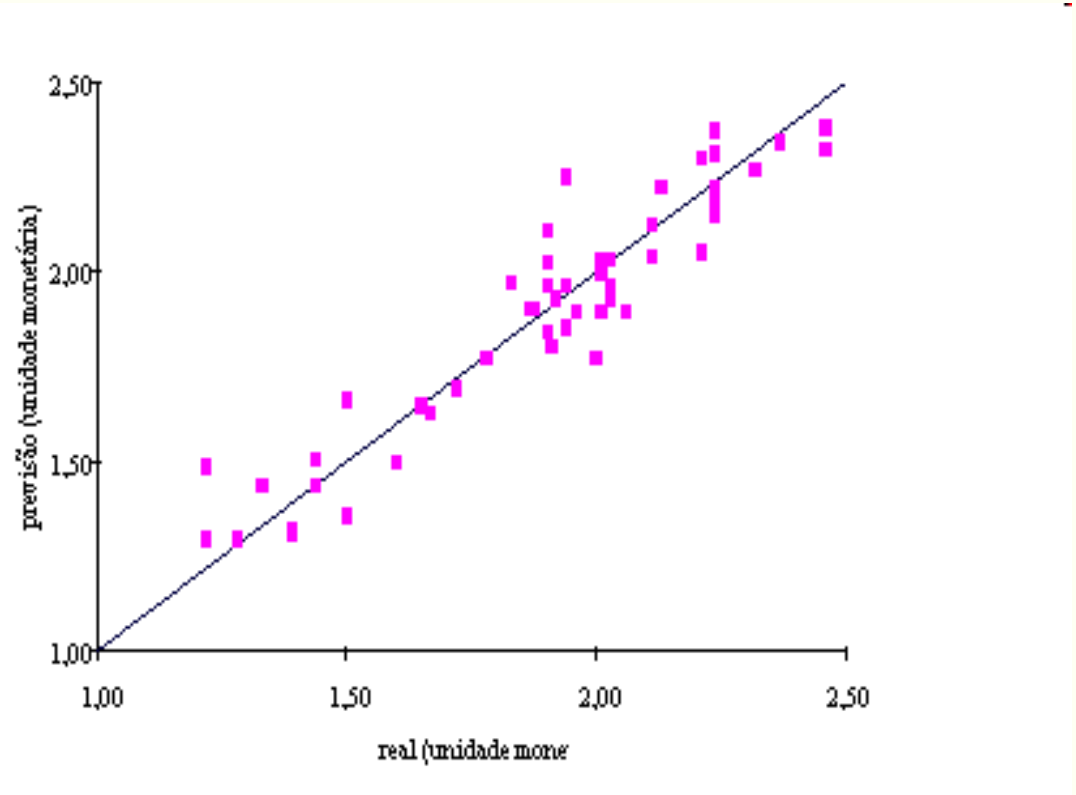


No R

```
> tempo<-c(3,7,2,1.5,12)
> nota<-c(4.5,6.5,3.7,4.0,9.3)
> plot(tempo,nota,main="Diagrama de disperção")
```

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:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_X s_Y}$$

sendo que

$\bar{x}$, $\bar{y}$ são as médias amostrais de X e Y, respectivamente,

s_X e s_Y são os desvios padrão de X e Y, respectivamente.

Fórmula alternativa:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)s_X s_Y}$$

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

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

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

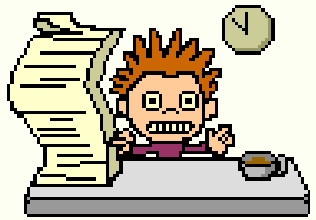
$$S_y^2 = \frac{(-1,1)^2 + \ldots + (3,7)^2}{4} = \frac{21,9}{4} = 5,47 \Rightarrow S_y = 2,34$$

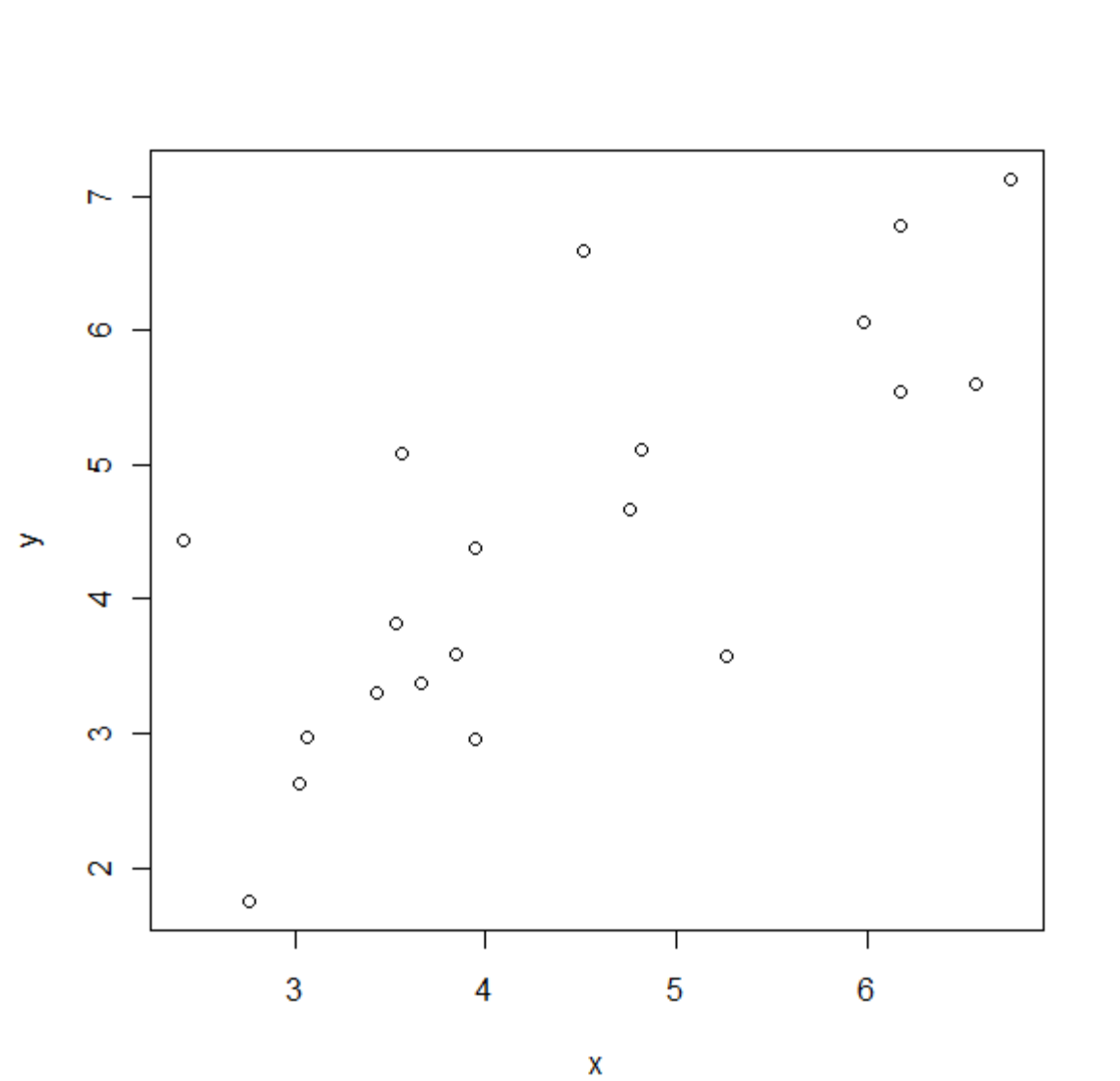
Então,

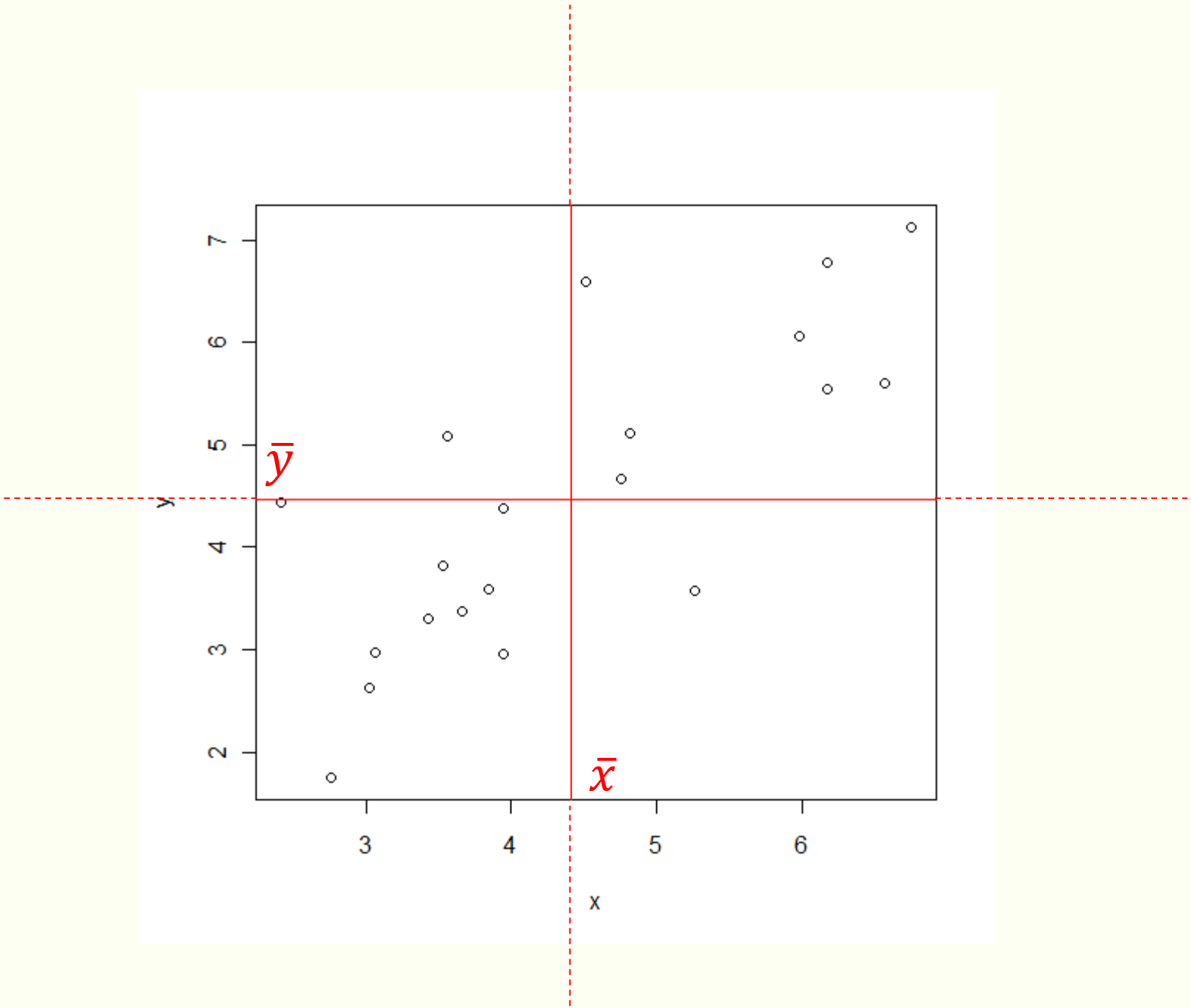
$$r = \frac{41,2}{4 \,.\, 4,42 \,.\, 2,34} = 0,9959$$

No R temos:

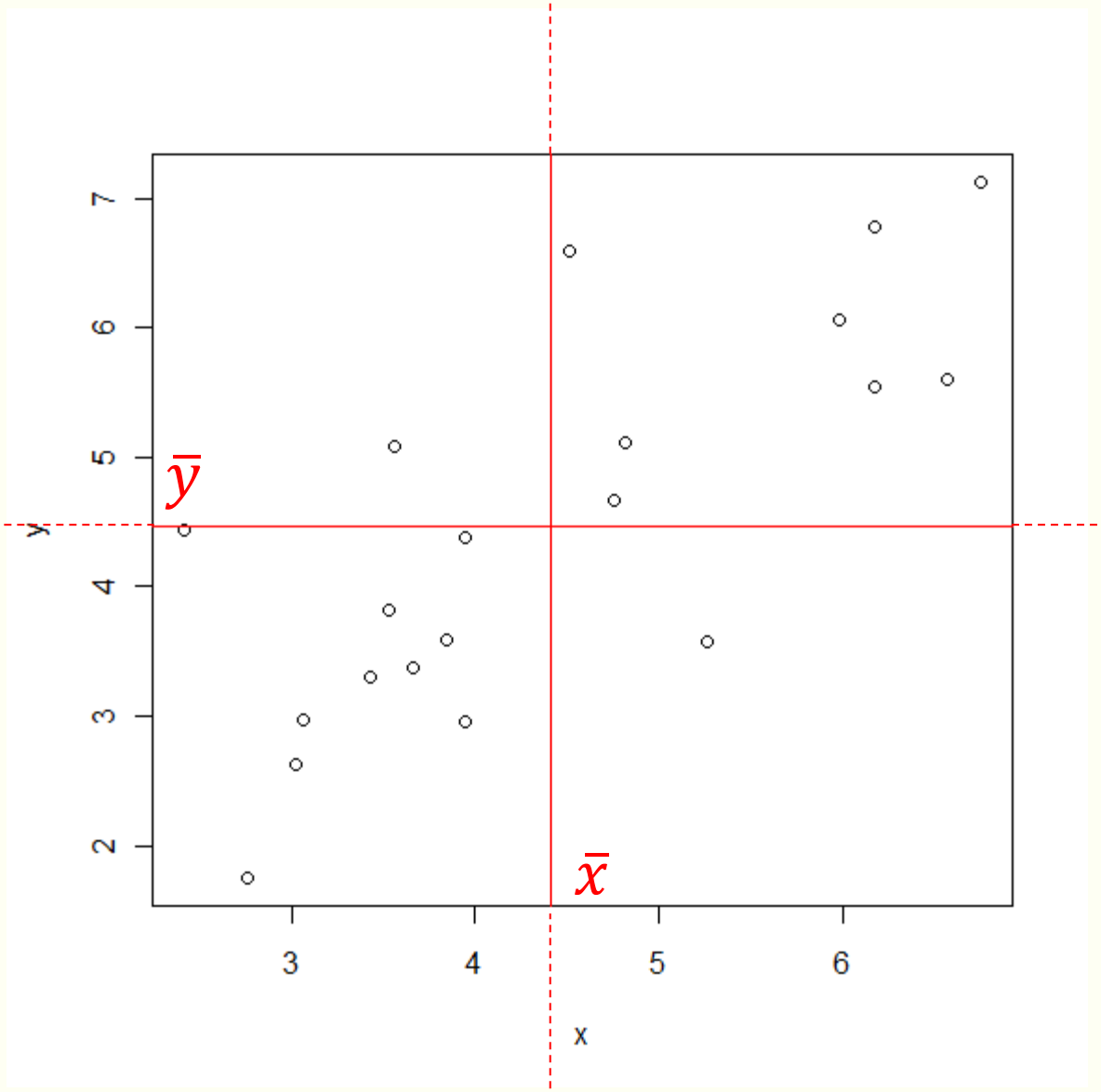
```
> cor(tempo,nota)
[1] 0.9960249
```



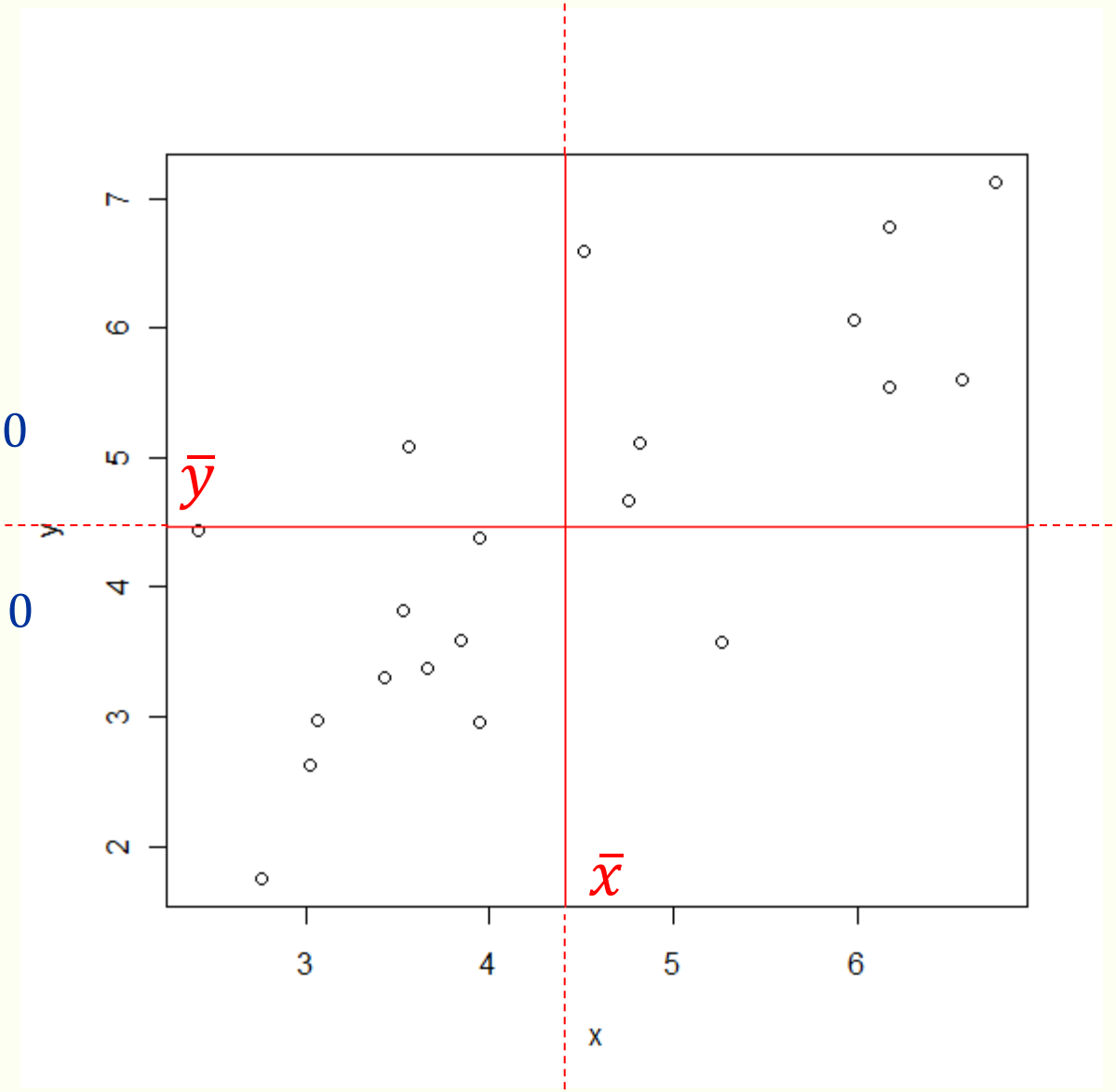

y
2
3
4
5
6
7
x
3
4
5
6


$\bar{y}$
$\bar{x}$
y
x
2
3
4
5
6
7
3
4
5
6

$(x_i - \bar{x}) < 0$ $(x_i - \bar{x}) > 0$

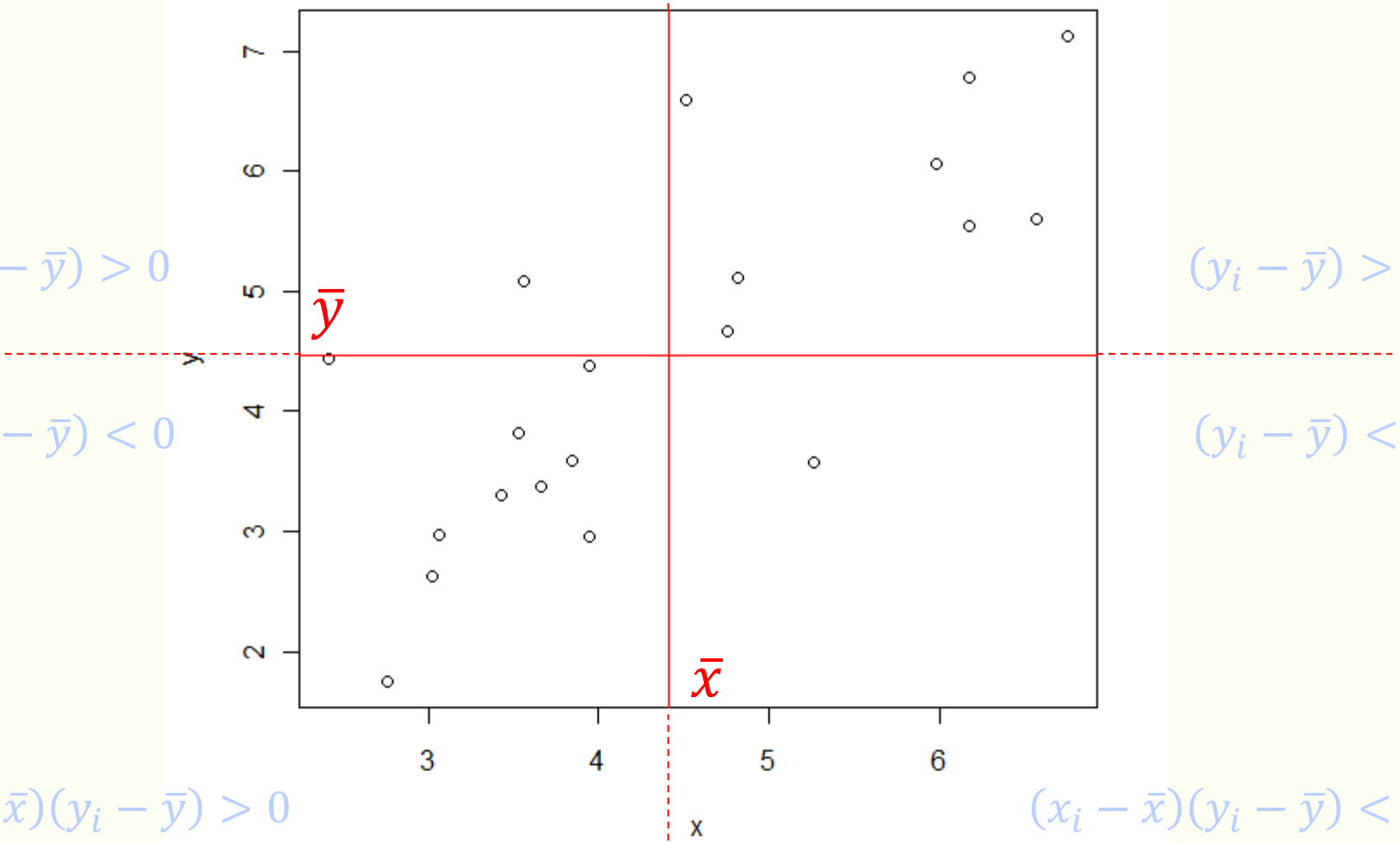


$(x_i - \bar{x}) < 0$ $(x_i - \bar{x}) > 0$


$(x_i - \bar{x}) < 0$
$(x_i - \bar{x}) > 0$
$(y_i - \bar{y}) > 0$
$(y_i - \bar{y}) > 0$
$\bar{y}$
$(y_i - \bar{y}) < 0$
$(y_i - \bar{y}) < 0$
$\bar{x}$
$(x_i - \bar{x}) < 0$
$(x_i - \bar{x}) > 0$
y
x
2
3
4
5
6
7
3
4
5
6

$(x_i - \bar{x}) < 0$ $(x_i - \bar{x}) > 0$

$(x_i - \bar{x})(y_i - \bar{y}) < 0$ $r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_X s_Y} > 0$ $(x_i - \bar{x})(y_i - \bar{y}) > 0$



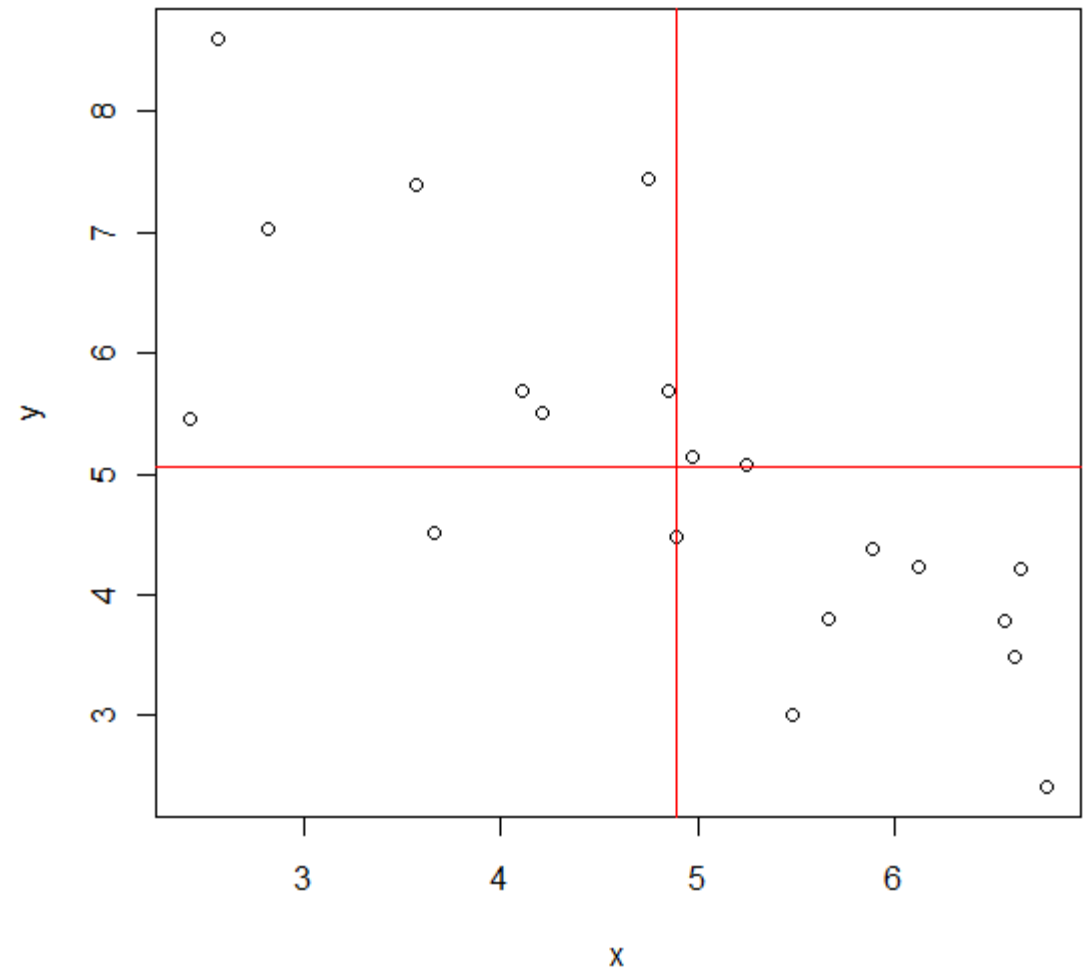
$(y_i - \bar{y}) > 0$ $(y_i - \bar{y}) > 0$

$(y_i - \bar{y}) < 0$ $(y_i - \bar{y}) < 0$

$(x_i - \bar{x})(y_i - \bar{y}) > 0$ $(x_i - \bar{x})(y_i - \bar{y}) < 0$

$(x_i - \bar{x}) < 0$ $(x_i - \bar{x}) > 0$

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_X s_Y} < 0$$



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

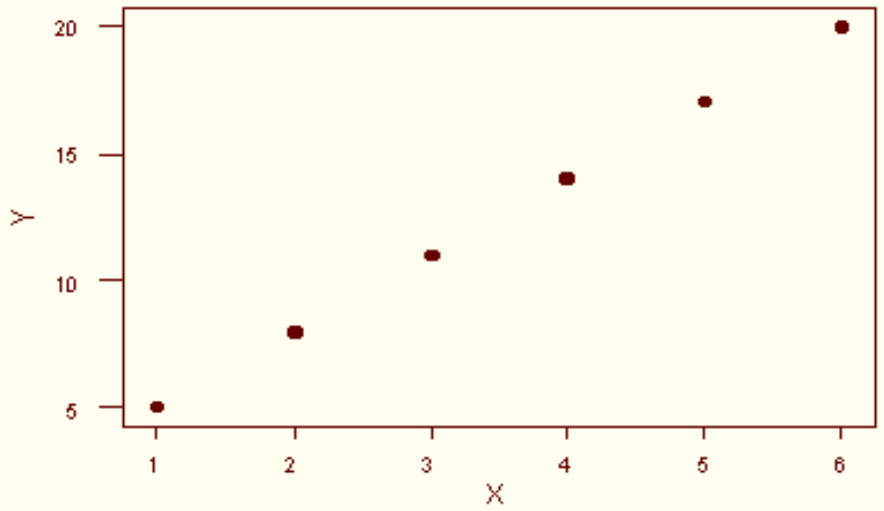
Casos particulares:

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

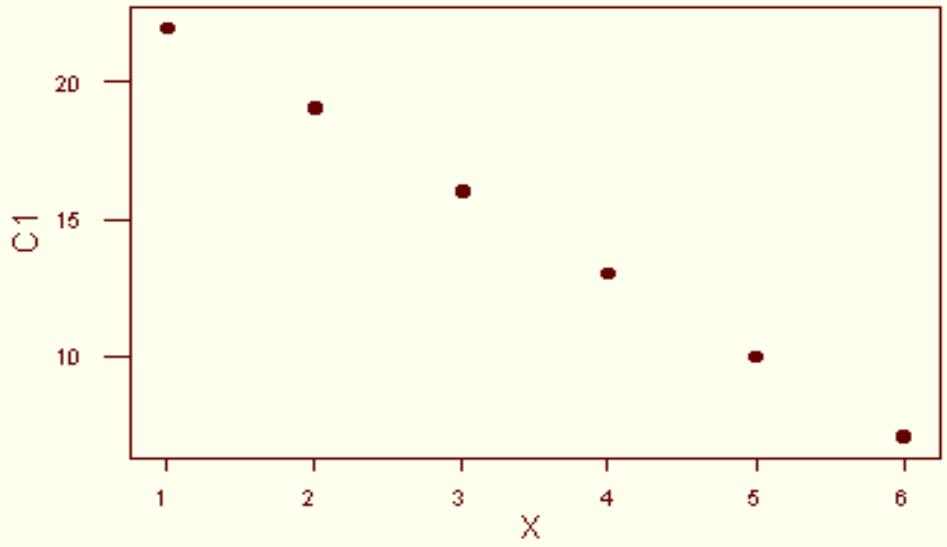
$r = -1 \Rightarrow$ correlação linear negativa e perfeita

$r = 0 \Rightarrow$ inexistência de correlação linear

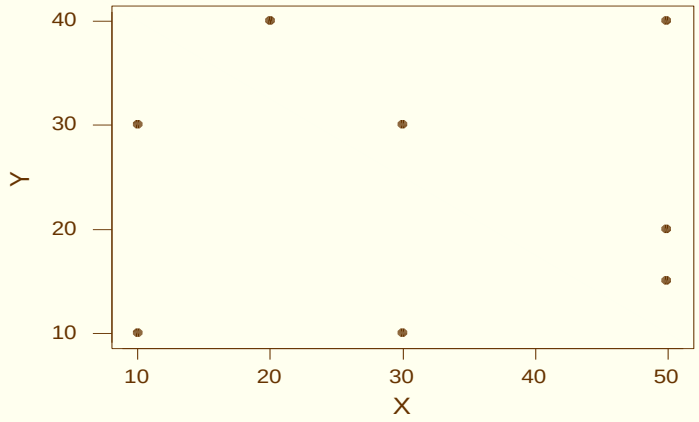
r = 1, correlação linear positiva e perfeita

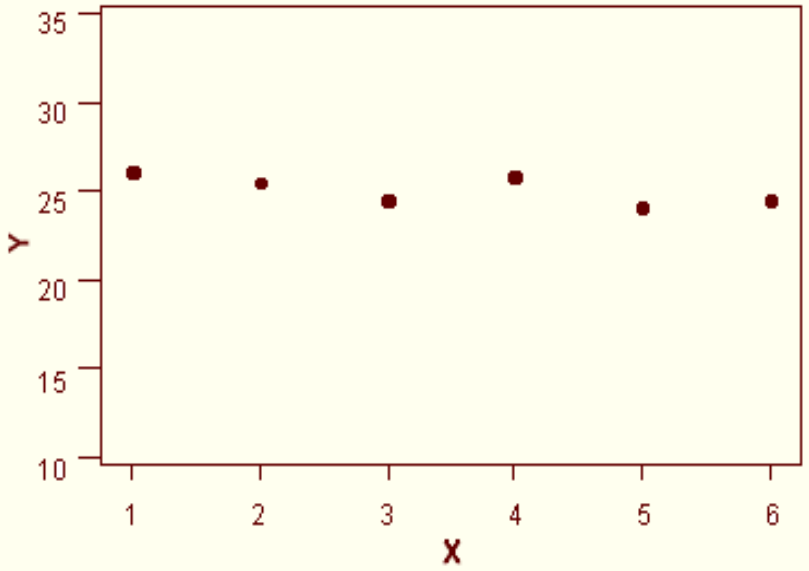


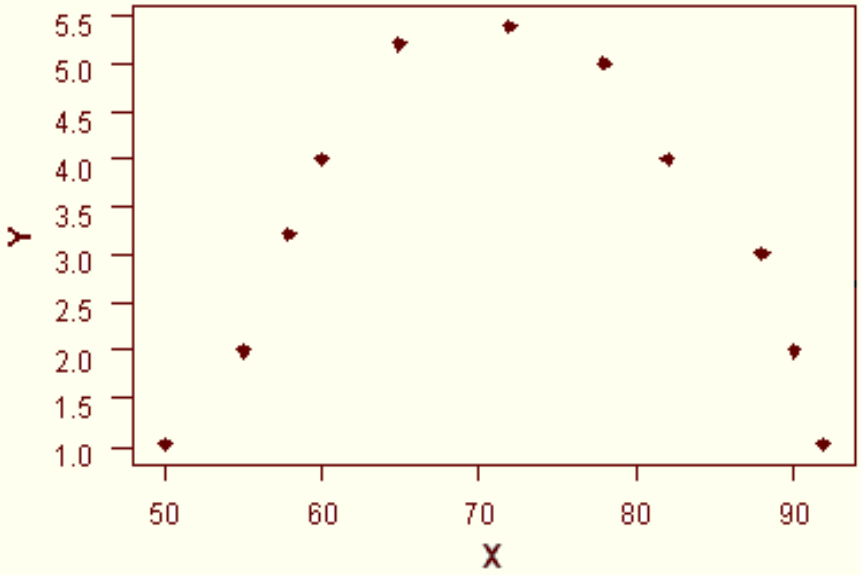
r = -1, correlação linear negativa e perfeita

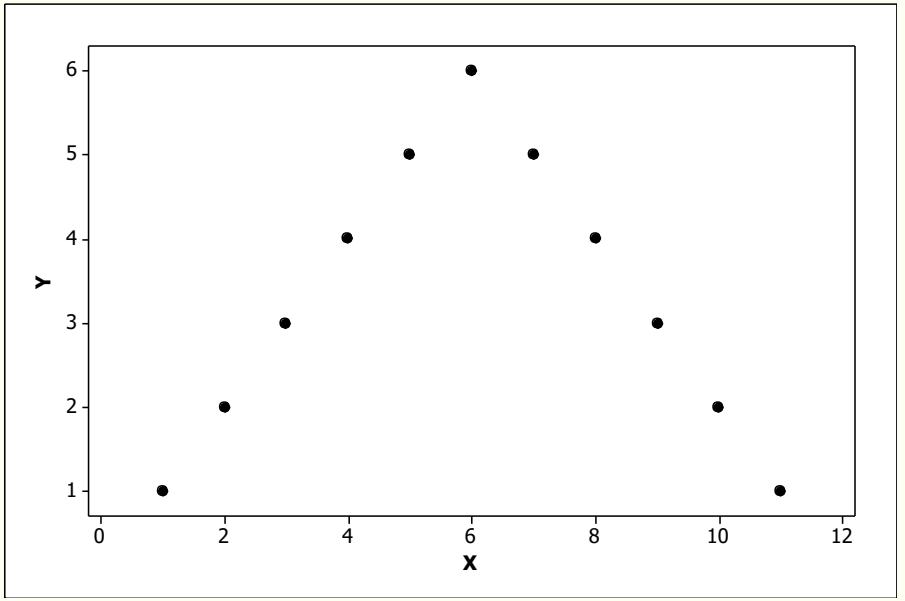


$r \cong 0$

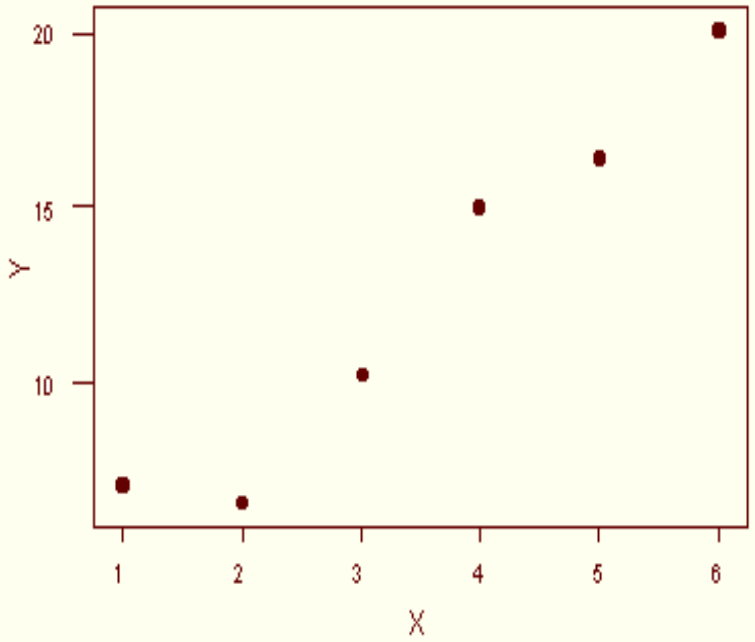

40
30
20
10
Y
10
20
30
40
50
X


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

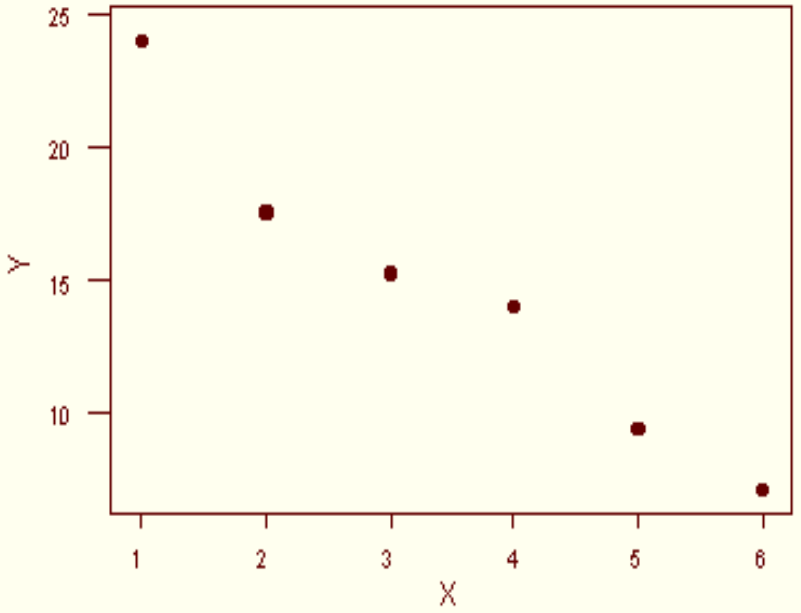

5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Y
50
60
70
80
90
X


6
5
4
3
2
1
Y
0
2
4
6
8
10
12
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**

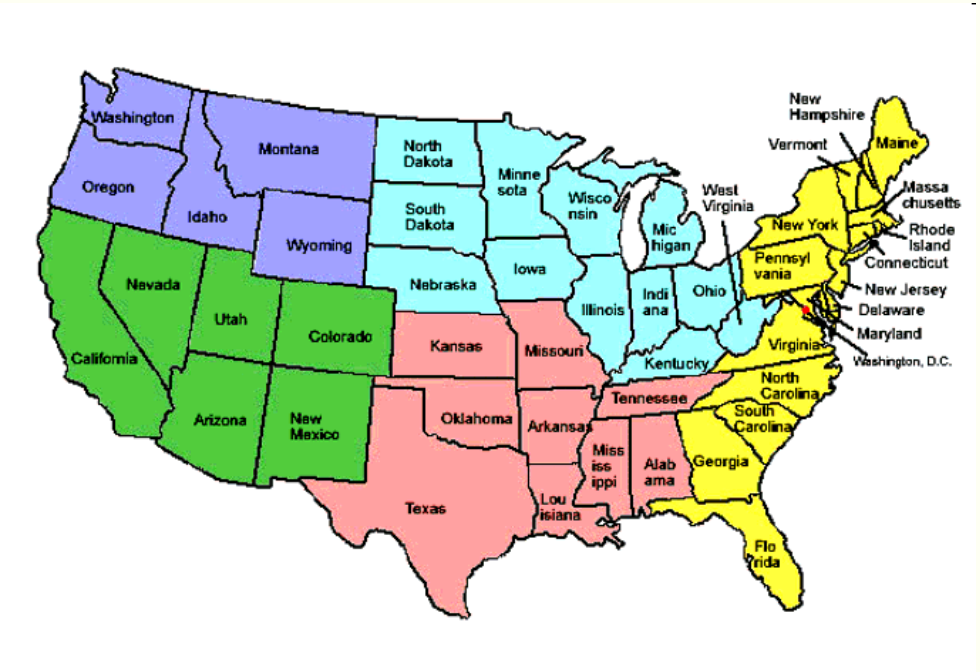
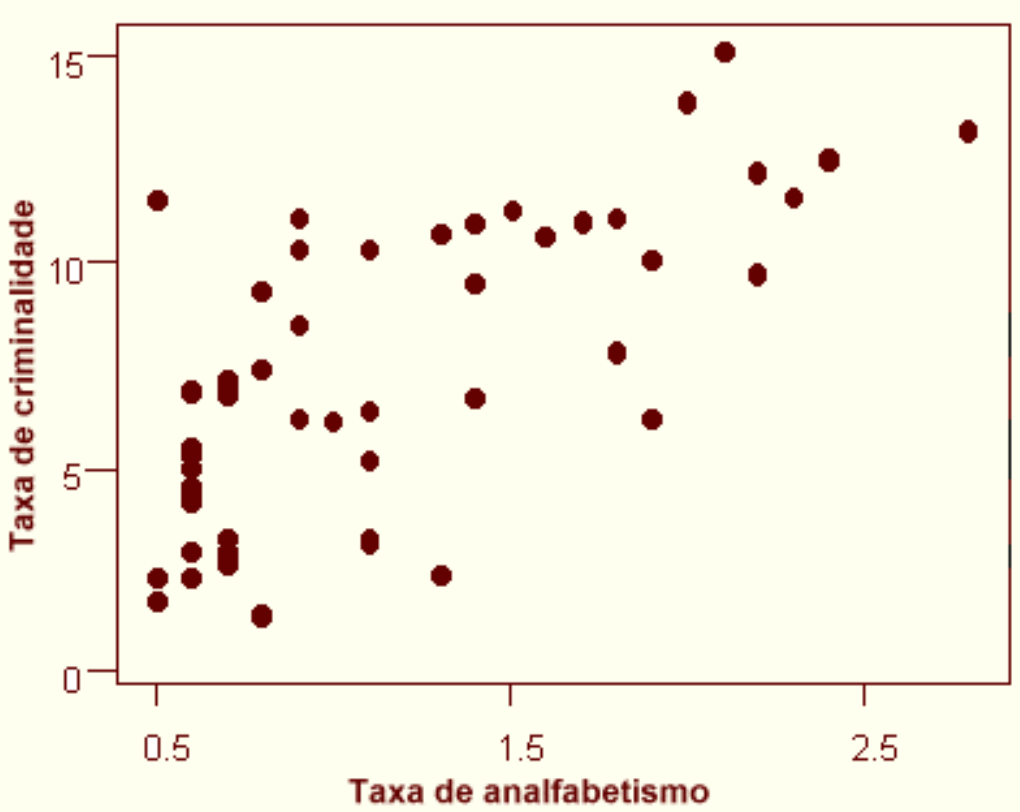


Diagrama de dispersão



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

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

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

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

$\Sigma X_i Y_i$ = 4122,8

Correlação entre X e Y:

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

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

Comentário:

- **Na interpretação do coeficiente de correlação é importante visualizar o diagrama de dispersão.**

ARQUIVO FA.MTW : 6 variáveis são medidas em 11 indivíduos

Row	X	Y1	Y2	Y3	X4	Y4
1	10	8,04	9,14	7,46	8	6,58
2	8	6,95	8,14	6,77	8	5,76
3	13	7,58	8,74	12,74	8	7,71
4	9	8,81	8,77	7,11	8	8,84
5	11	8,33	9,26	7,81	8	8,47
6	14	9,96	8,10	8,84	8	7,04
7	6	7,24	6,13	6,08	8	5,25
8	4	4,26	3,10	5,39	19	12,50
9	12	10,84	9,13	8,15	8	5,56
10	7	4,82	7,26	6,42	8	7,91
11	5	5,68	4,74	5,73	8	6,89

MTB > corr X Y1

Pearson correlation of X and Y1 = 0,816

Pearson correlation of X and Y2 = 0,816

Pearson correlation of X and Y3 = 0,816

Pearson correlation of X4 and Y4 = 0,817

⇒ Mesmos valores de correlação.

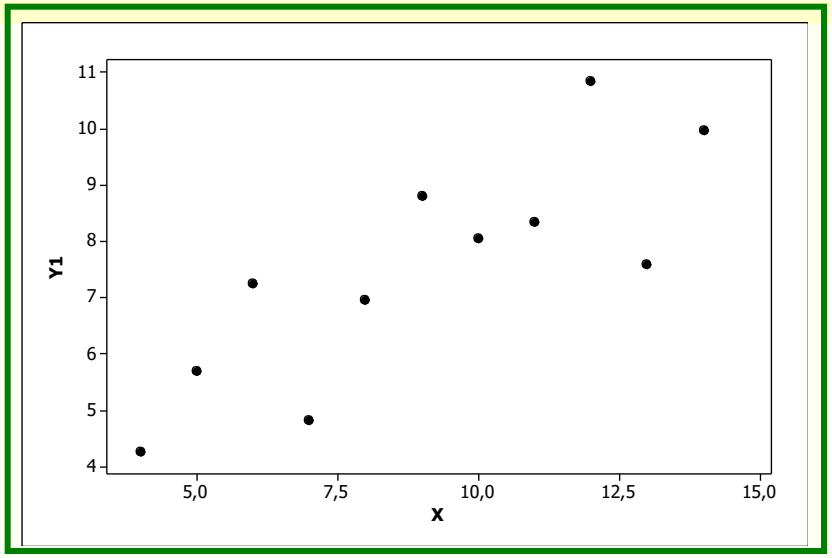
⇒ Qual a forma esperada da dispersão conjunta destas variáveis?

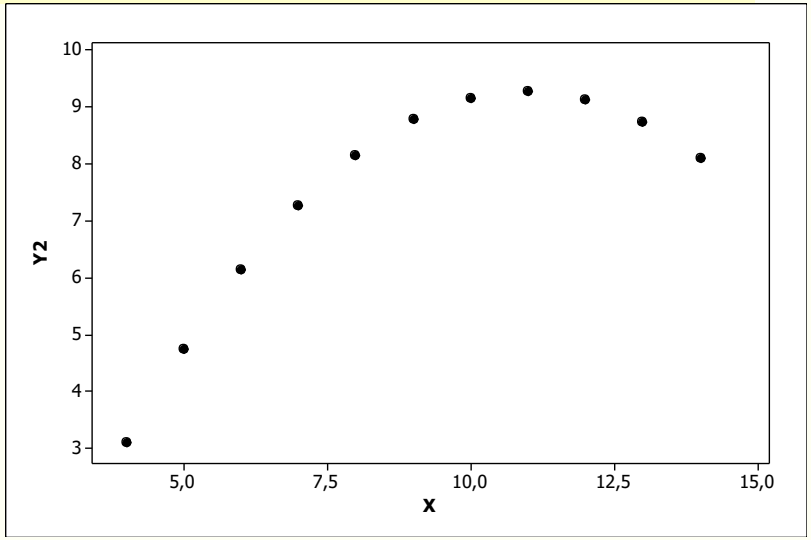
ARQUIVO FA.MTW

Diagramas de dispersão e Coeficientes de Correlação

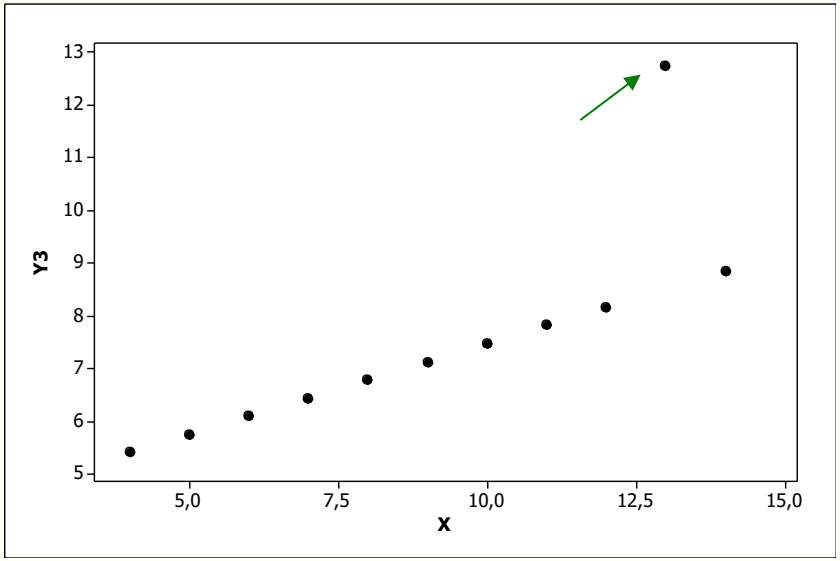
r = 0,816

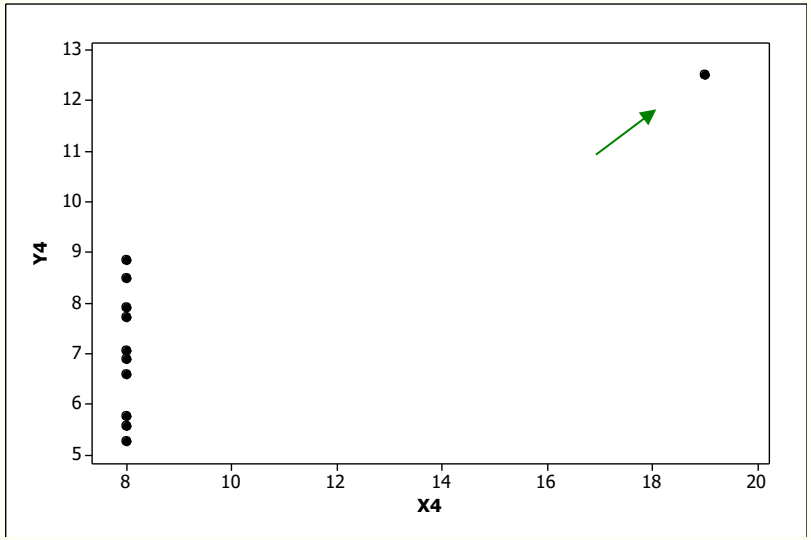
Dispersão esperada!





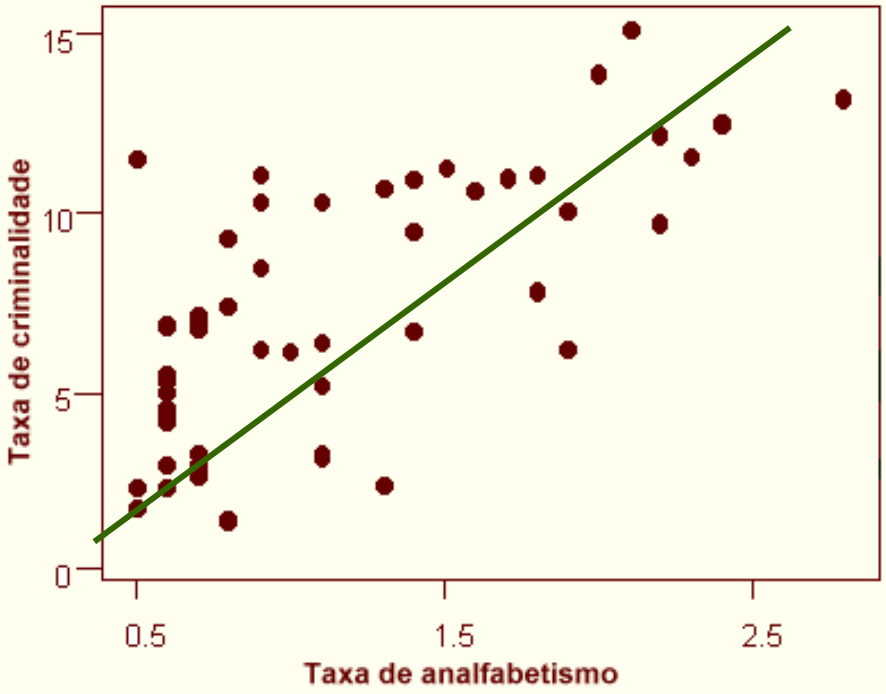
Pontos influentes!

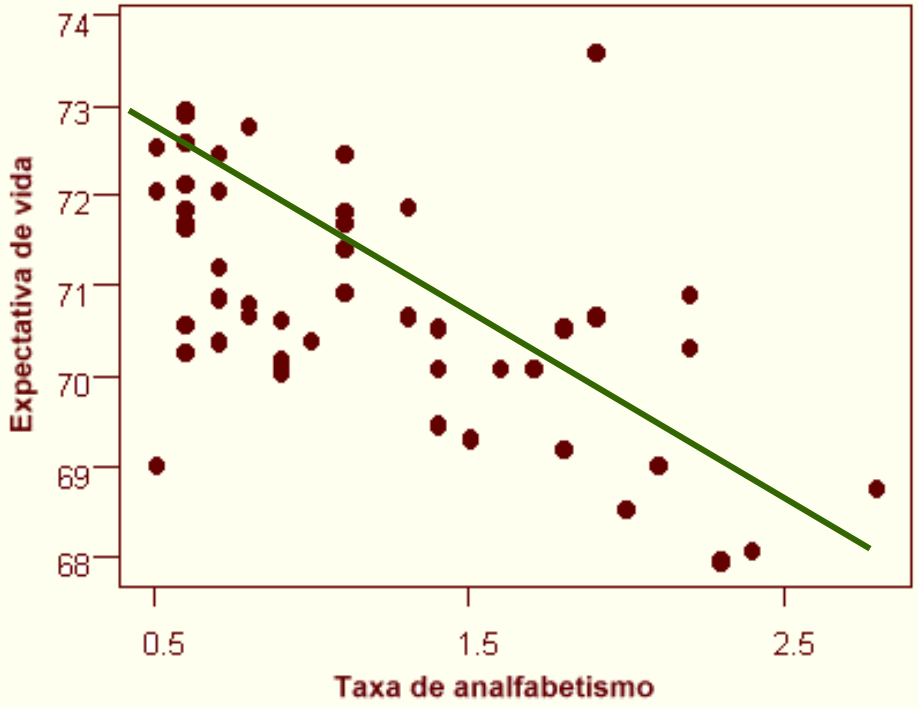




Análise de Regressão

Diagramas de Dispersão





⇒ **Explicar a forma da relação por meio de uma função matemática: $Y = a + bX$**

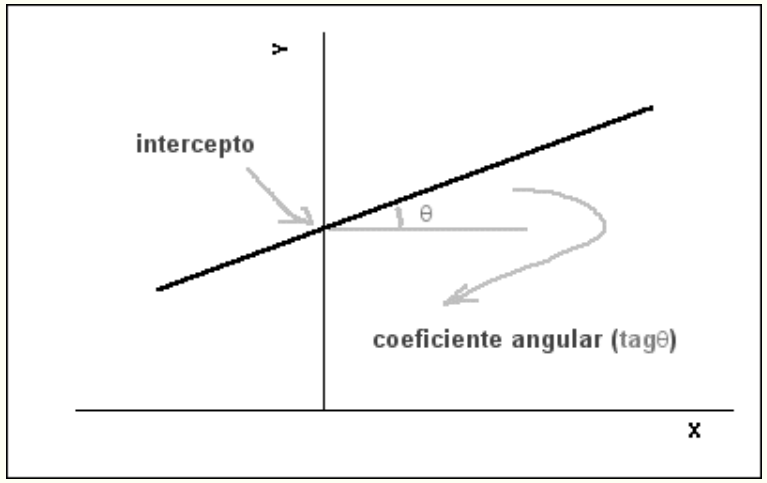
Análise de Regressão

Reta ajustada: $\hat{Y} = a + bX$

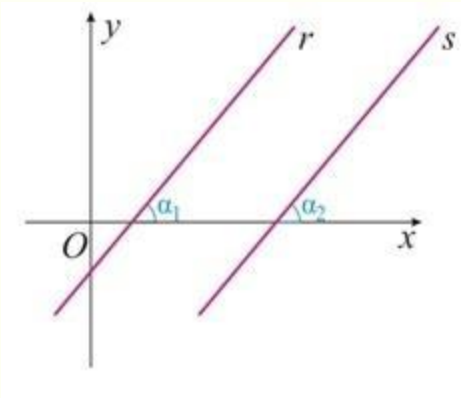
O que são *a* e *b*?

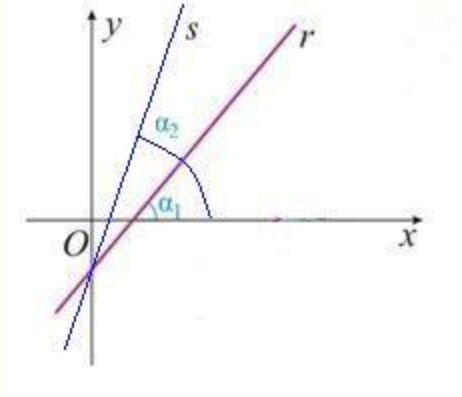
***a* : intercepto**

***b* : inclinação ou coeficiente angular**



Análise de Regressão





- Iguais coeficientes angulares
- Diferentes interceptos

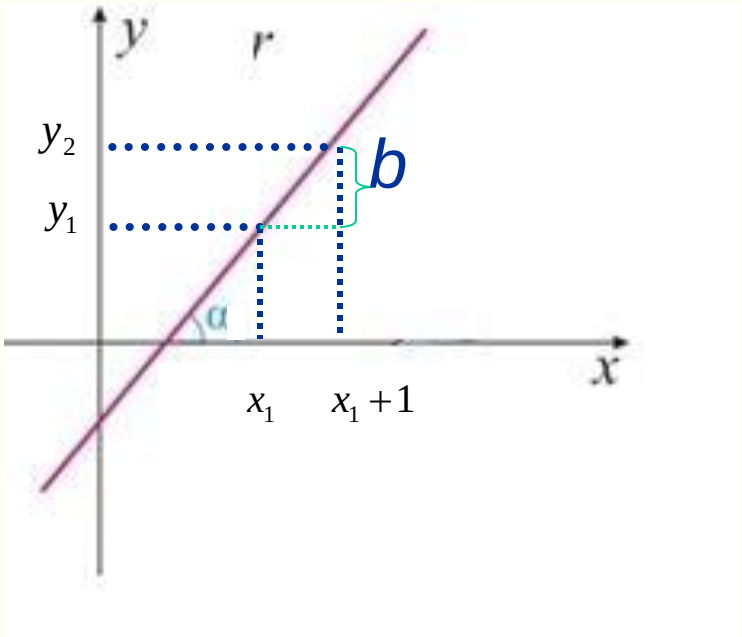
- Diferentes coeficientes angulares
- Iguais interceptos

Reta ajustada: $\hat{Y} = a + bX$

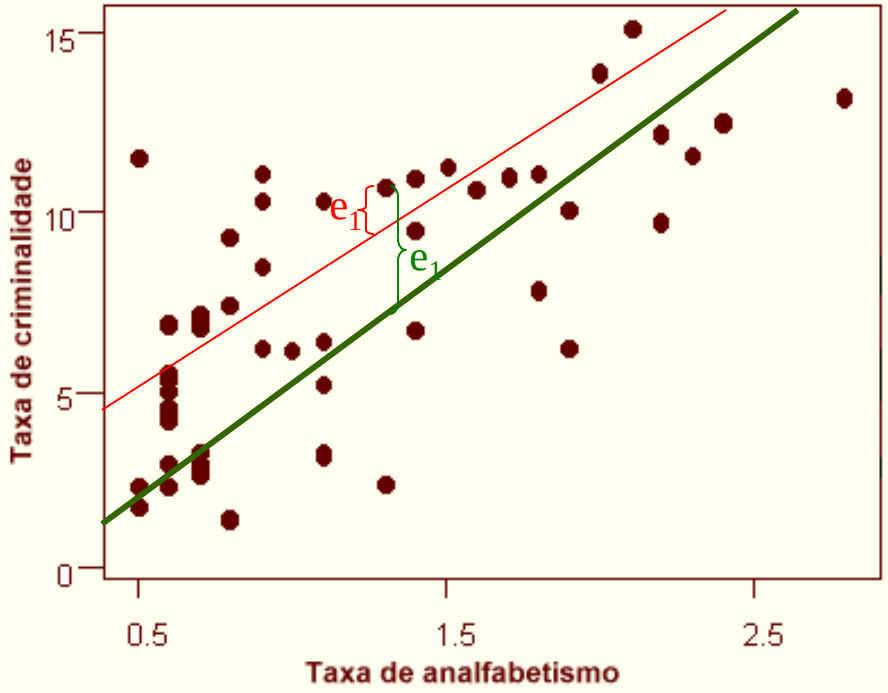
Interpretação de *b*:

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

$$tag(\alpha) = \frac{y_2 - y_1}{x_2 - x_1} = \frac{y_2 - y_1}{x_1 + 1 - x_1}$$
$$= y_2 - y_1 = b$$



Reta ajustada (método de mínimos quadrados)



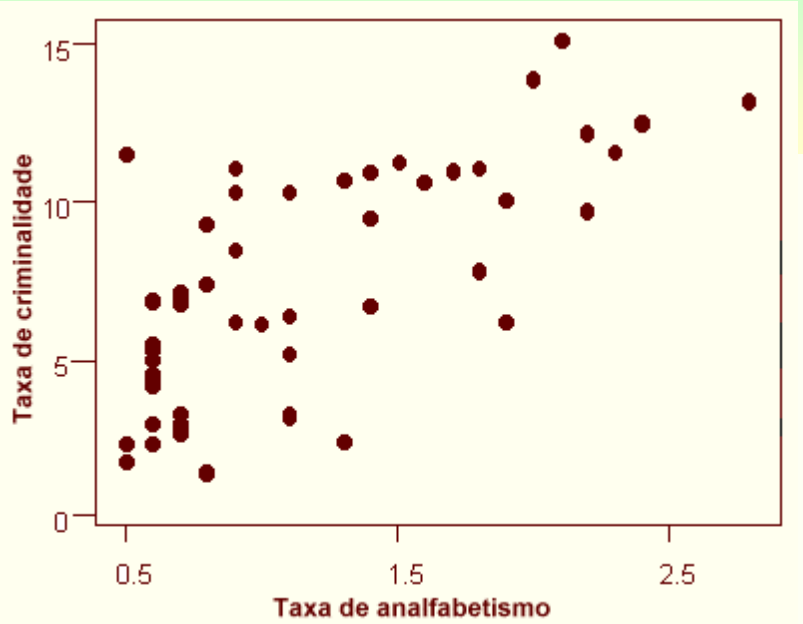
Reta ajustada (método de mínimos quadrados)

Os coeficientes *a* e *b* são calculados da seguinte maneira:

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

$$a = \bar{Y} - b\,\bar{X}$$

No Exemplo 2,



A reta ajustada é:

$$\hat{Y} = 2{,}397 + 4{,}257\,X$$

$\hat{Y}$: valor predito para a taxa de criminalidade

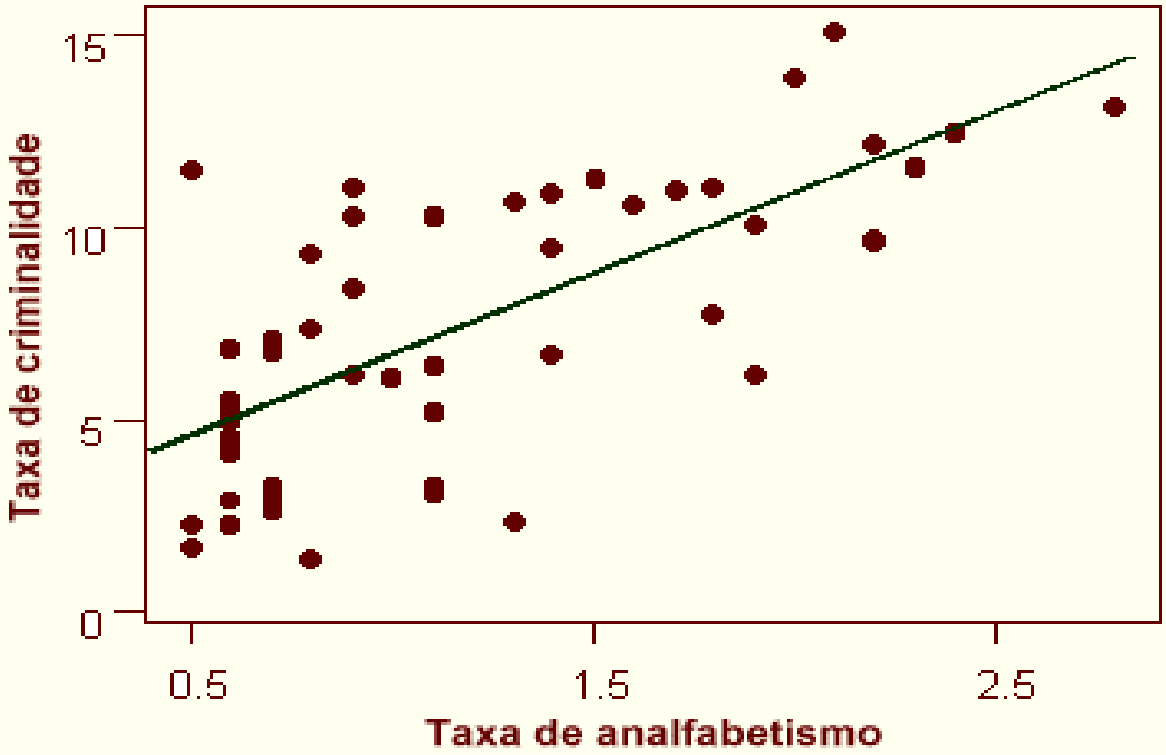
X : taxa de analfabetismo

Interpretação de *b*:

Para um aumento de uma unidade na taxa do analfabetismo (*X*), a taxa de criminalidade (*Y*) aumenta, em média, 4,257 unidades.

Graficamente, temos

$$\hat{Y} = 2{,}397 + 4{,}257\ X$$

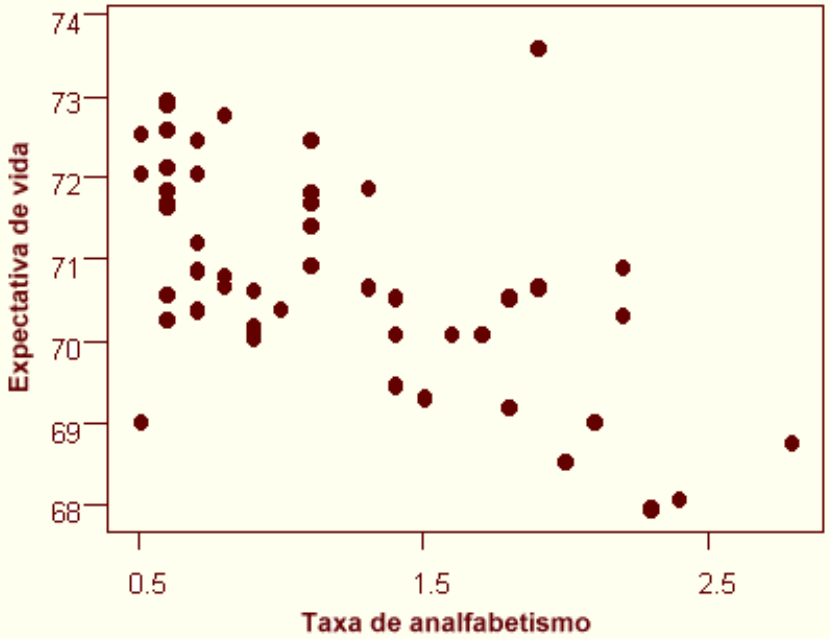


Como desenhar a reta no gráfico?

No exemplo 3,

A reta ajustada é:

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



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

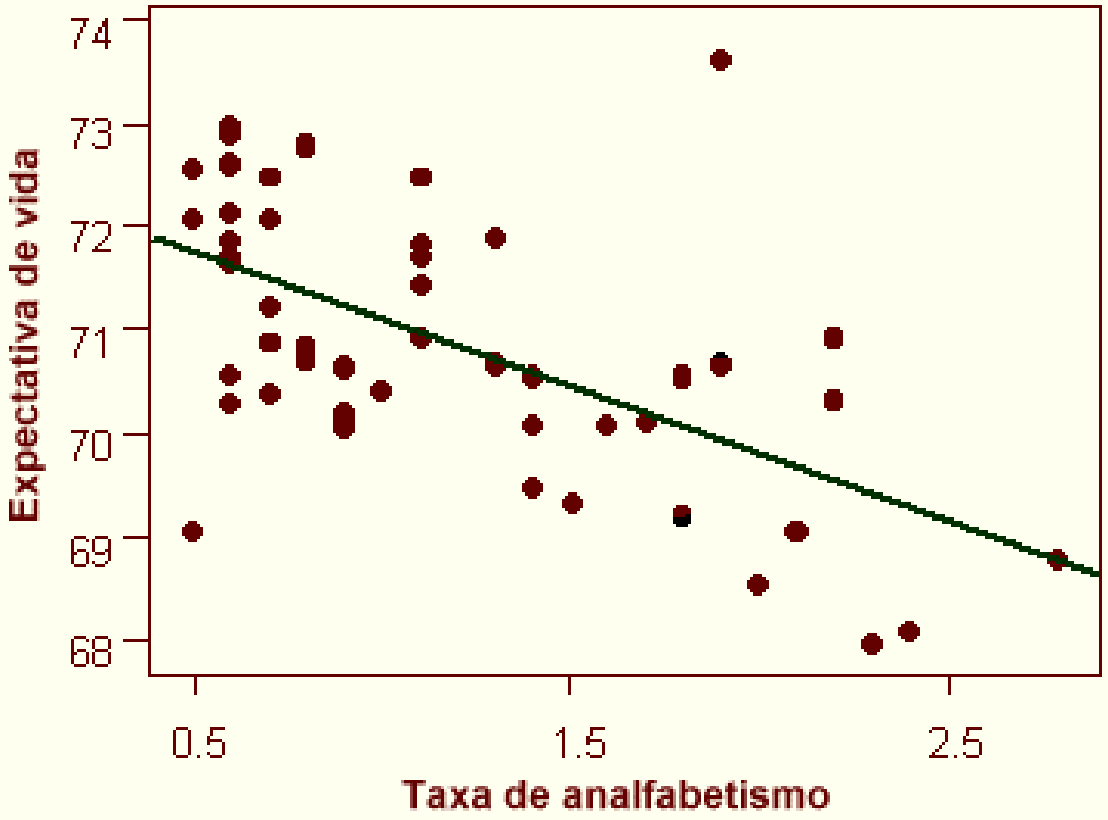
X : taxa de analfabetismo

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

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



Exemplo 4: consumo de cerveja e temperatura

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

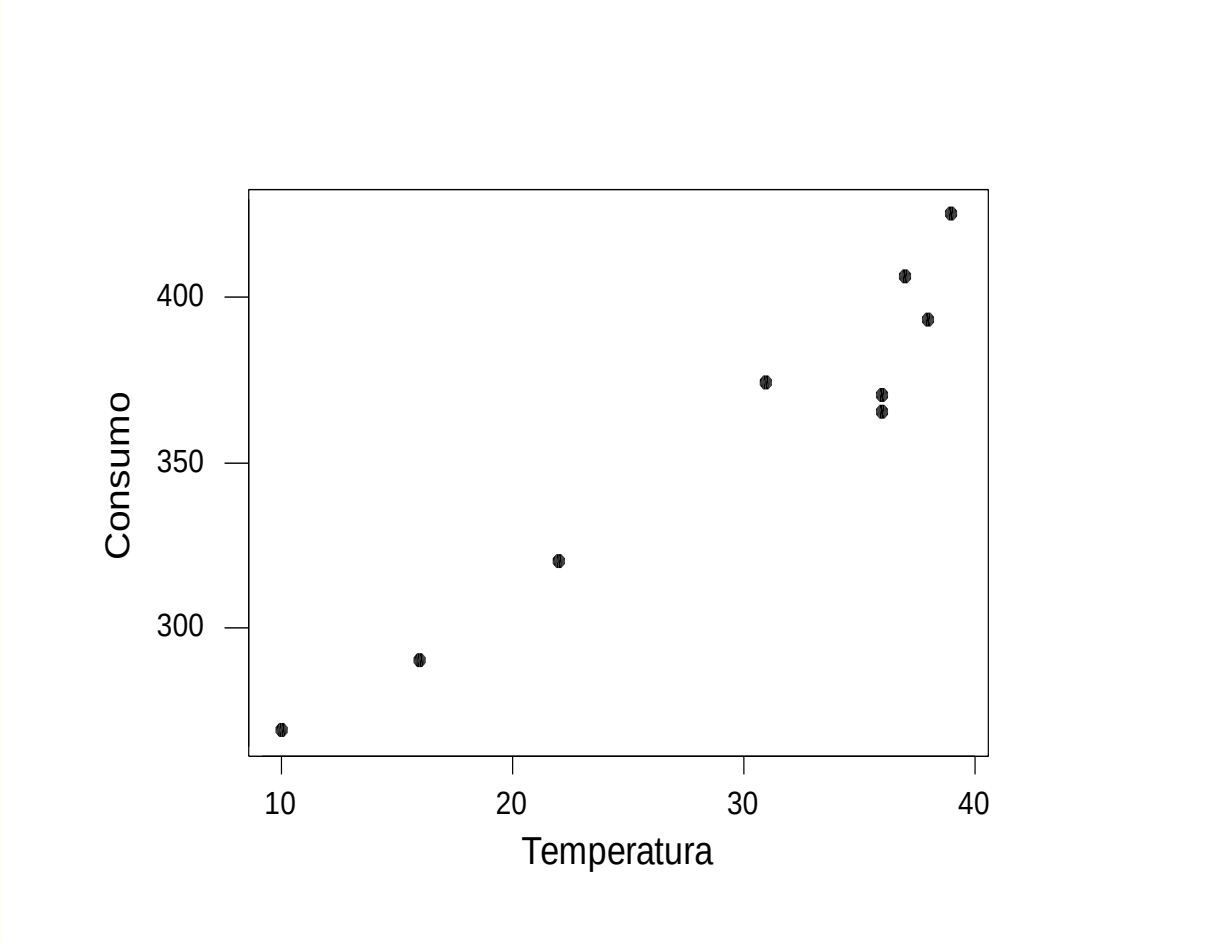
X: temperatura máxima (em ºC).

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

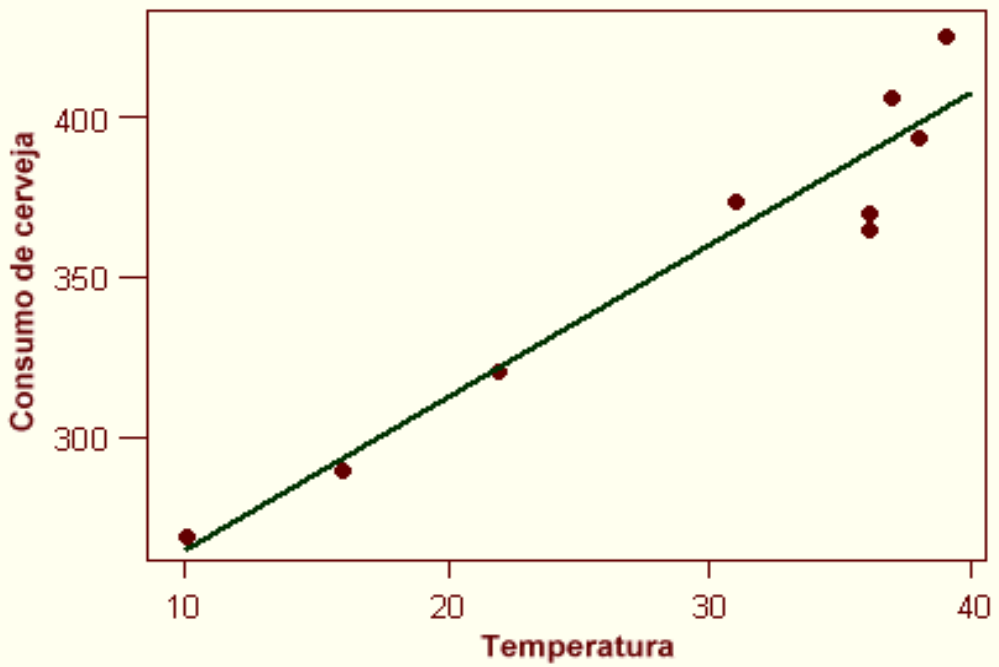
Diagrama de dispersão



A correlação entre *X* e *Y* é *r* = 0,962.

A reta ajustada é:

$$\hat{Y} = 217{,}37 + 4{,}74\,X$$



Qual é 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 é o consumo previsto para uma temperatura de 25ºC?

$$\hat{Y} = 217{,}37 + 4{,}74\,(25) = 335{,}87 \text{ litros}$$