Estatística Descritiva III

Associação entre Variáveis

Associação entre variáveis qualitativas

Tabelas de Contingência

Podemos construir tabelas de <u>frequências conjuntas</u> (*tabelas de contingência*), relacionando duas variáveis qualitativas.

Exemplo 1: Dados ***CEA06P24***, do projeto ***Caracterização Postural de Crianças de 7 e 8 anos das Escolas Municipais da Cidade de Amparo/SP***

- **Estudo realizado pelo Departamento de Fisioterapia, Fonoaudiologia e Terapia Ocupacional da Faculdade de Medicina da USP;**
- **Ano de realização: 2006;**
- **Finalidade: mestrado;**
- **Análise estatística: Centro de Estatística Aplicada (CEA), IME-USP.**

A) Há indícios de associação entre Lado da escoliose e Tipo de mochila?

	Lado da Escoliose			
Tipo de Mochila	Ausente	Direito	Esquerdo	Total
Carrinho	8	37	35	80
Escapular	16	35	72	123
Lateral	2	10	11	23
Total	26	82	118	226

Qual é o significado dos valores desta tabela?

No R:

- **Dados → Importar arquivos de dados →**

  → de conjunto de dados do Excel, Access ou dBase...

(Defina o nome do conjunto de dados: *dados*)

- **Estatísticas → Tabelas de Contingência → Tabelas de dupla entrada**

(Variável linha : *tipomochila* ; Variável coluna: *escollado*) – Saída editada do software R

	Lado da escoliose			
Tipo de mochila	**Ausente**	**Direito**	**Esquerdo**	**Total**
Carrinho	**8**	**37**	**35**	**80**
Escapular	**16**	**35**	**72**	**123**
Lateral	**2**	**10**	**11**	**23**
Total	**26**	**82**	**118**	**226**

Verificar associação através da:

- **porcentagem segundo as colunas, ou**
- **porcentagem segundo as linhas.**

	Lado da Escoliose		
Tipo de Mochila	Ausente	Direito	Esquerdo	Total
Carrinho	10,0%	46,2%	43,8%	100,0%
Escapular	13,0%	28,5%	58,5%	100,0%
Lateral	8,7%	43,5%	47,8%	100,0%
Total	11,5%	36,3%	52,2%	100,0%

Como concluir? Será que o Tipo de Mochila utilizada influencia o Lado da Escoliose (caso tenha) de uma criança?

Comparando as porcentagens de cada uma das linhas, observamos uma pequena diferença com relação à porcentagem total. Aparentemente, há pouca influência do tipo de mochila utilizada no lado de ocorrência da escoliose.

• **Estatísticas → Tabelas de Contingência → Tabelas de dupla entrada**

(Variável linha : *tipomochila* ; Variável coluna: *escollado*;

Marcar opção *Percentual nas linhas*) – Saída editada do software R

	Lado escoliose			
Tipo de mochila	**Ausente**	**Direito**	**Esquerdo**	**Total**
Carrinho	**10.0**	**46.2**	**43.8**	**100.0**
Escapular	**13.0**	**28.5**	**58.5**	**100.0**
Lateral	**8.7**	**43.5**	**47.8**	**100.0**
Total	**11.5**	**36.3**	**52.2**	**100.0**

B) Será que existe relação entre o Sexo das crianças e o Tipo de Mochila utilizada por elas?

	Tipo de Mochila			
Sexo	**Carrinho**	**Escapular**	**Lateral**	**Total**
Feminino	**53**	**59**	**16**	**128**
Masculino	**27**	**64**	**7**	**98**
Total	**80**	**123**	**23**	**226**

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

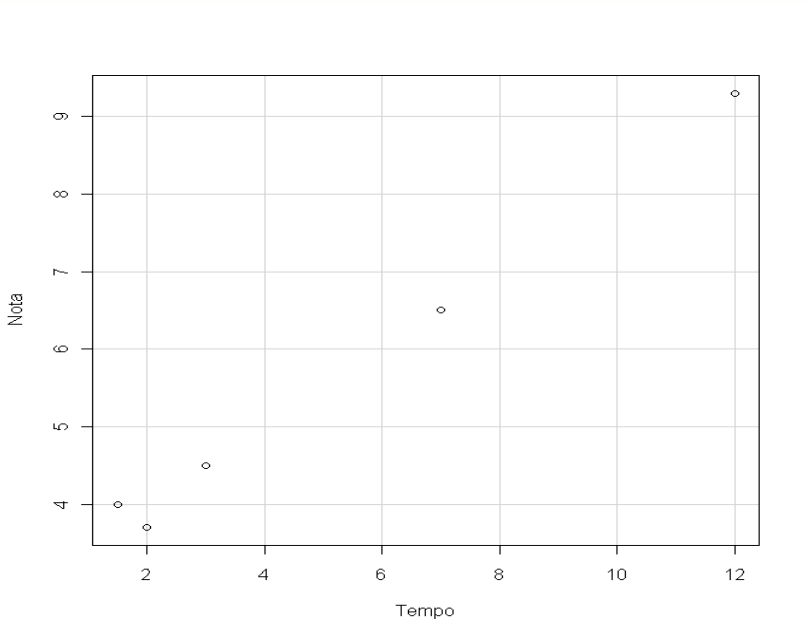
Exemplo 2: nota da prova e tempo de estudo

X : tempo de estudo (em horas)

Y : nota da prova

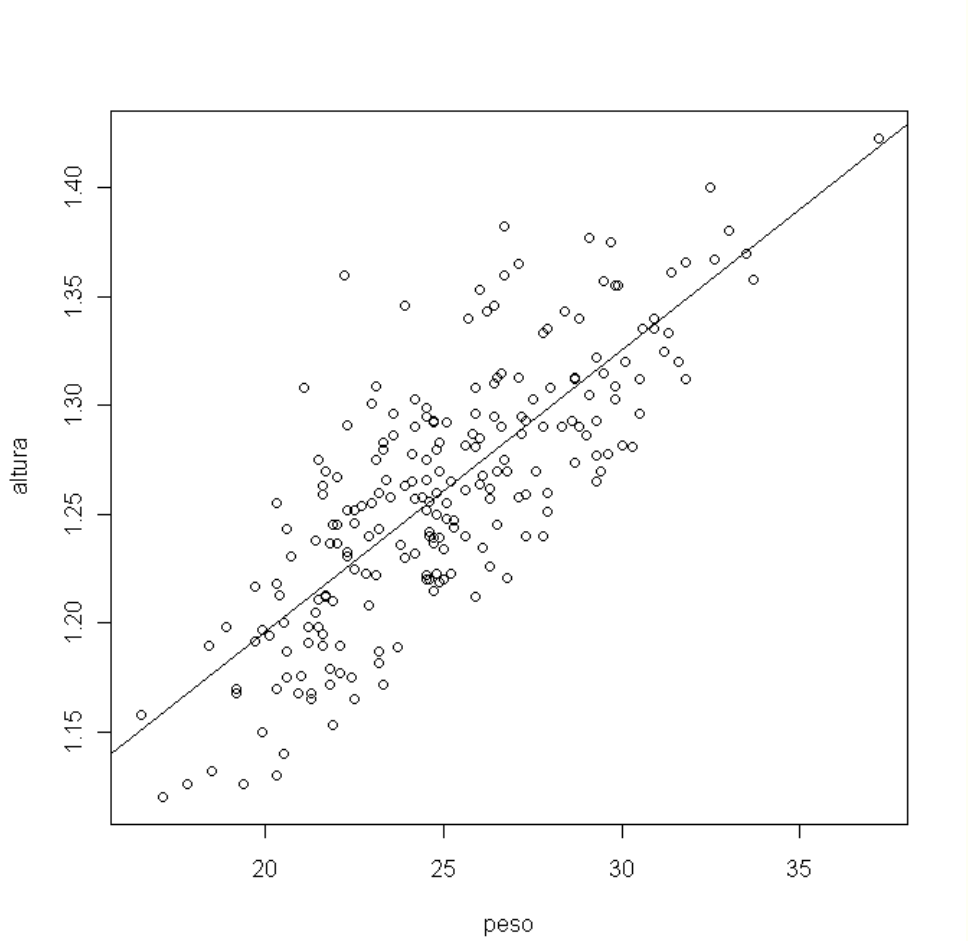
Pares de observações (X_i , Y_i) para cada estudante

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



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.

Fórmula alternativa:

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

$$S_X^2 = \frac{\sum_{i=1}^{n} X_i^{\,2} - n\overline{X}^{\,2}}{n-1}$$

Voltando ao Exemplo 2:

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

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

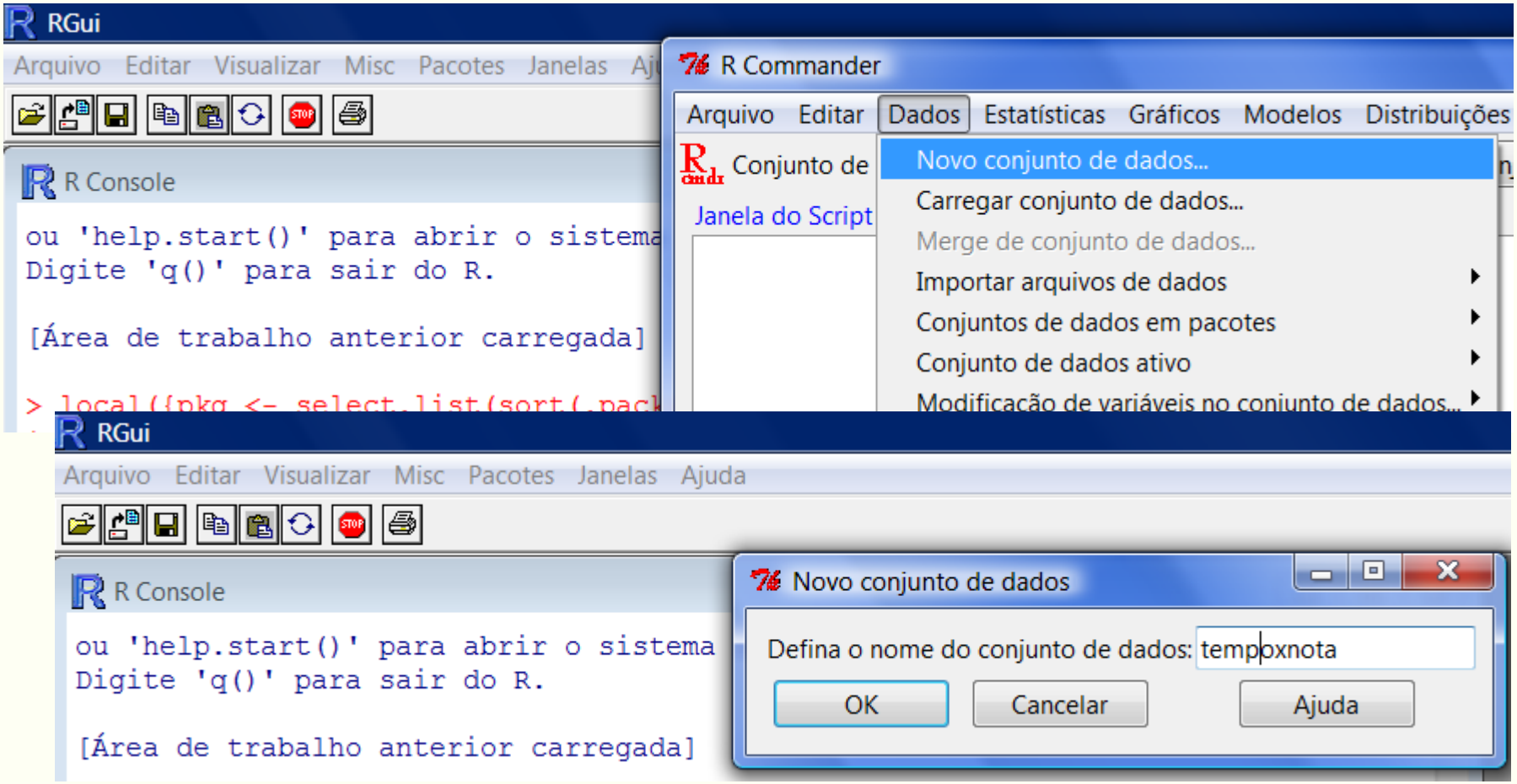
$$S_x^2 = \frac{(-2,1)^2 + \ldots + (6,9)^2}{4} = \frac{78,2}{4} = 19,55 \Rightarrow S_x = 4,42$$

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

Então,

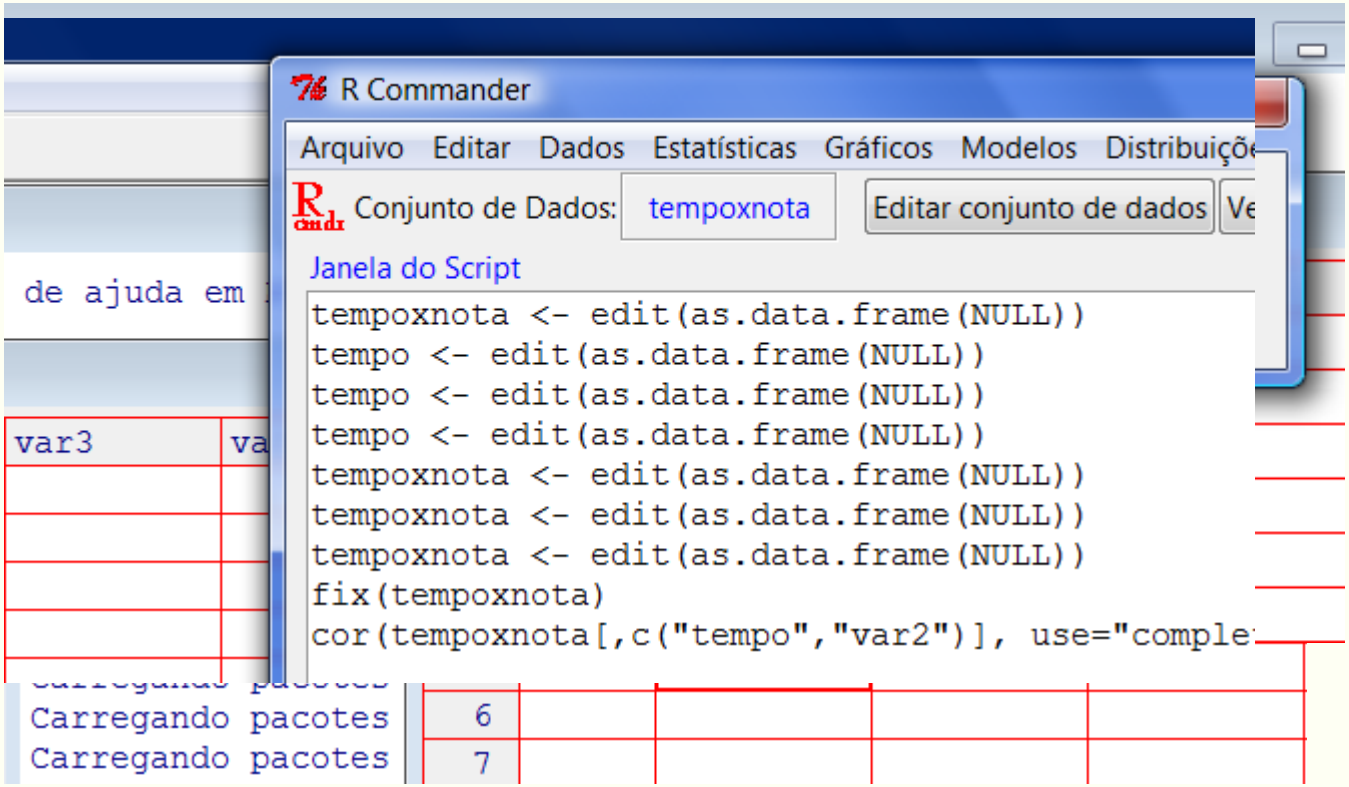
$$r = \frac{41,2}{4 \,.\, 4,42 \,.\, 2,34} = 0,9959$$

Criando arquivo de dados no R



Criando arquivo de dados no R

Digitar os dados na janela do editor e dar nomes (“Tempo” e “Nota”) às variáveis e fechar.



No R temos:

```
> cor(tempoxnota$Tempo, tempoxnota$Nota)

[1] 0.9960249
```

Ou ainda

- **Estatísticas → Resumos → Matriz de Correlação**

(Selecione *Tempo* e *Nota* no conjunto de dados *tempoxnota*)

```
           Nota      Tempo
Nota   1.0000000 0.9960249
Tempo  0.9960249 1.0000000
```

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

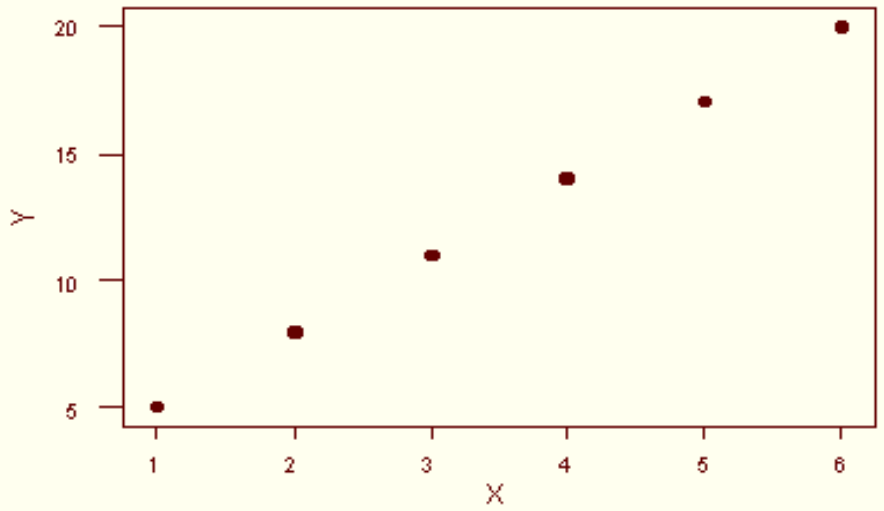
Casos particulares:

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

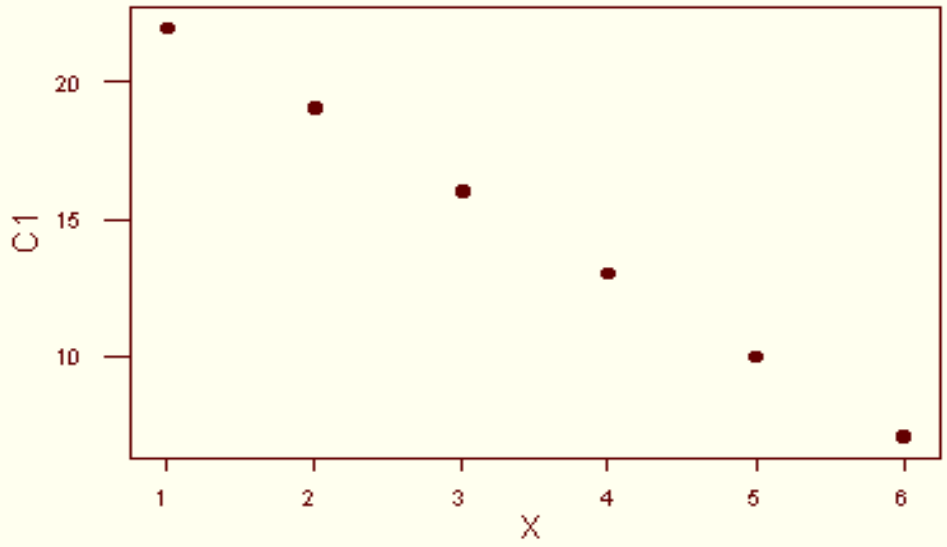
$r = -1 \Rightarrow$ correlação linear negativa e perfeita

$r = 0 \Rightarrow$ inexistência de correlação linear

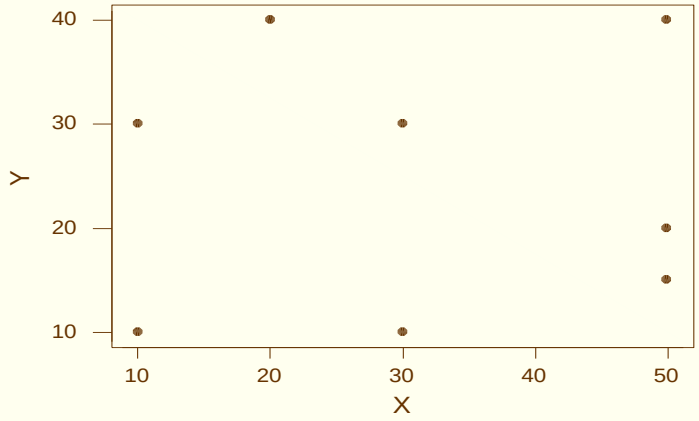
r = 1, correlação linear positiva e perfeita

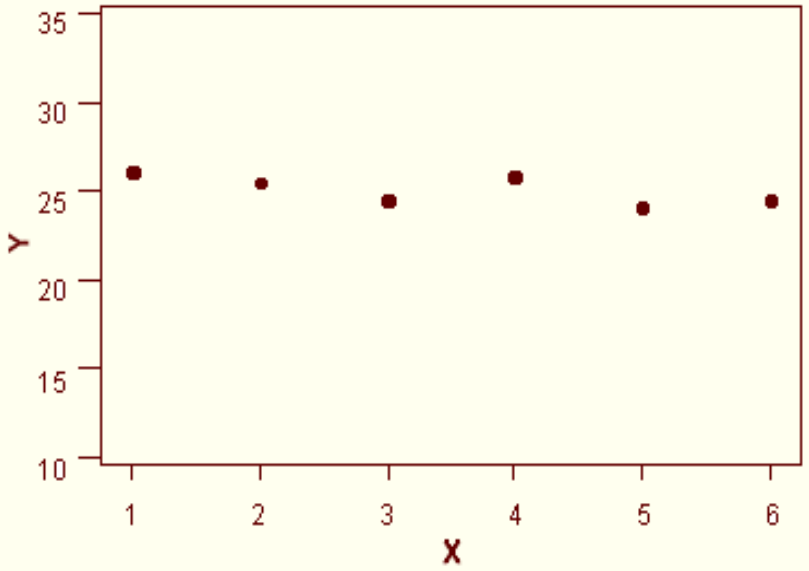


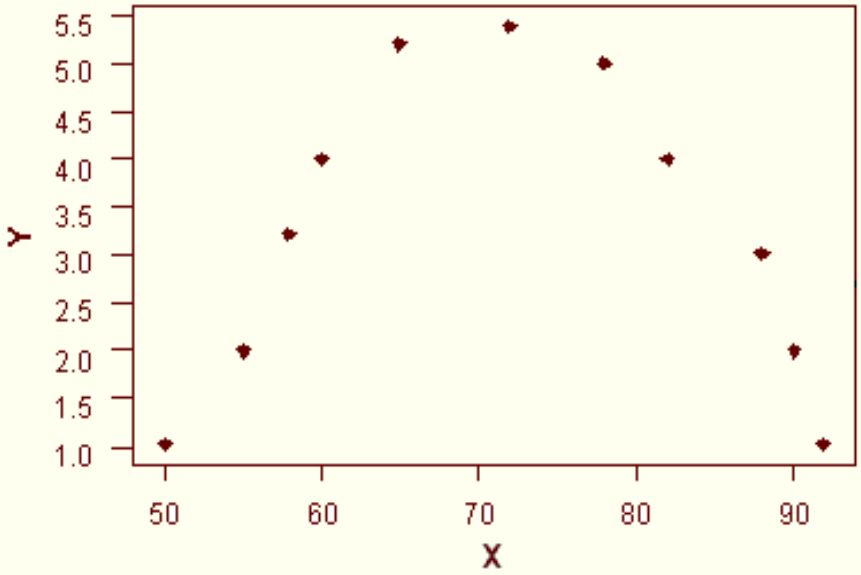
r = -1, correlação linear negativa e perfeita

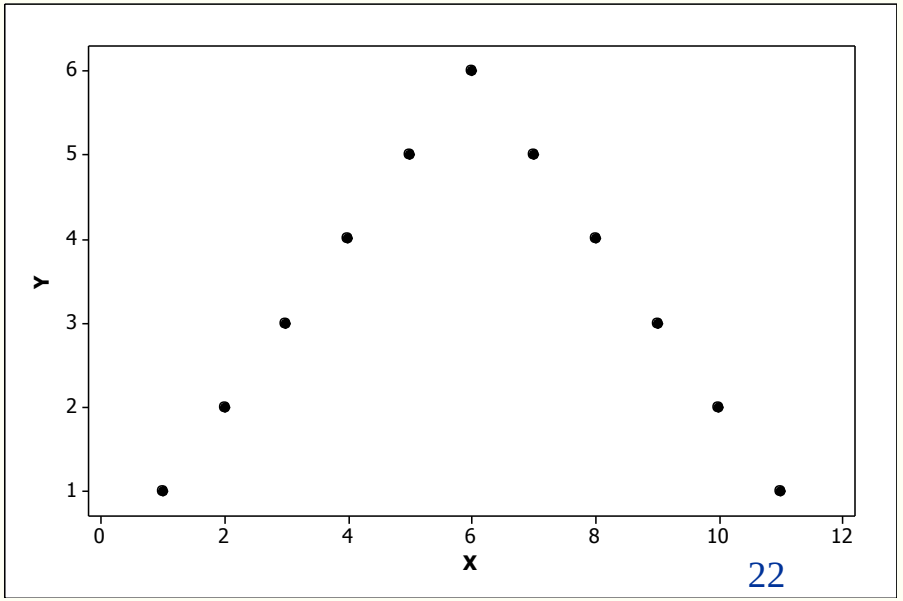


$r \cong 0$

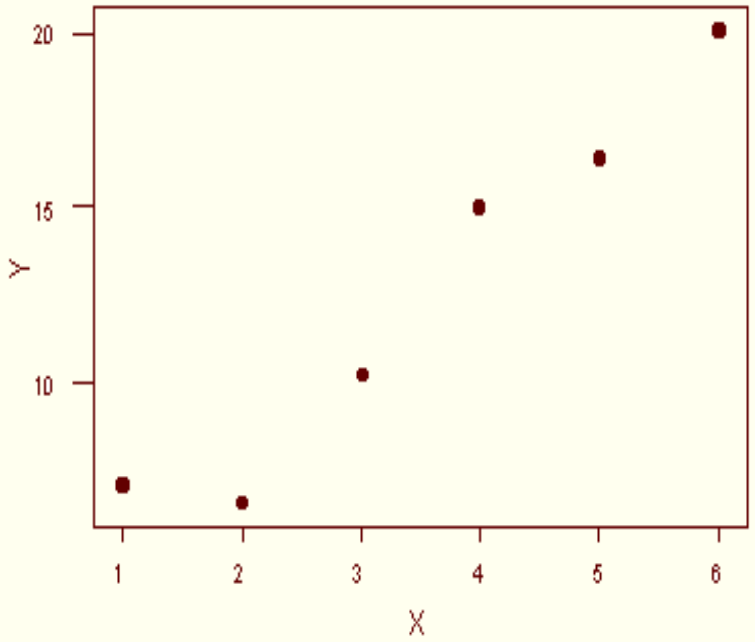

40
30
20
10
Y
10
20
30
40
50
X


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

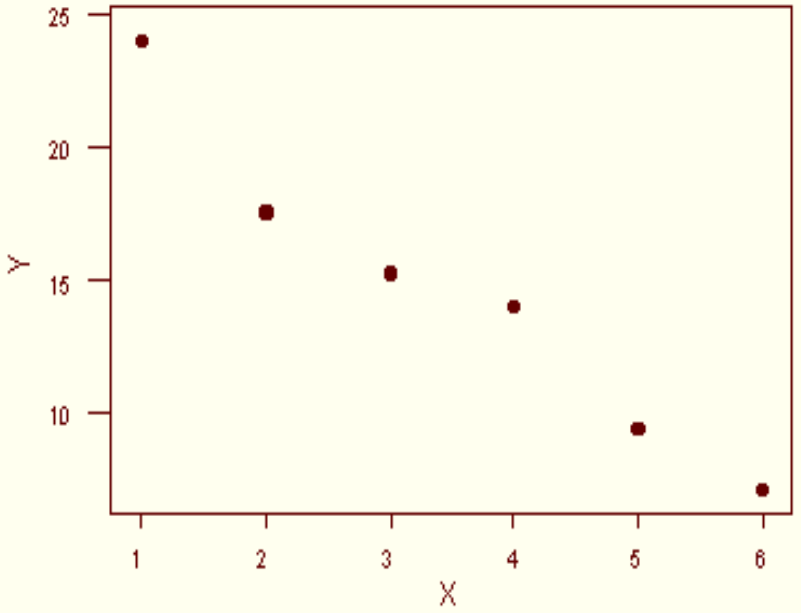

5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Y
50
60
70
80
90
X


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

$r \cong 1$



$r \cong -1$



Exemplo 3: criminalidade e analfabetismo

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

***Y*: taxa de criminalidade**

***X*: taxa de analfabetismo**

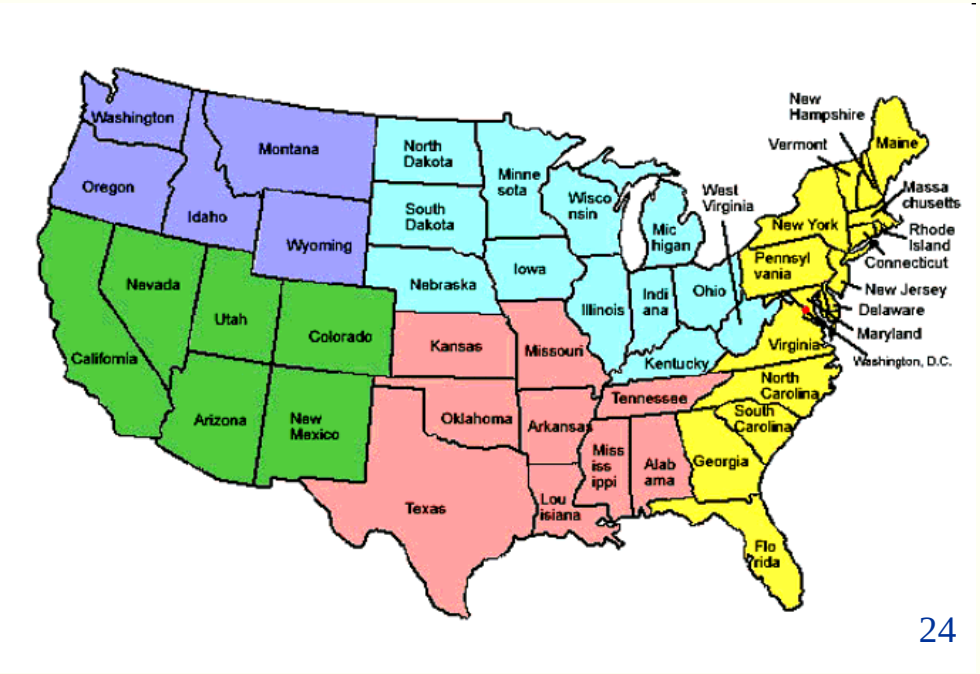


Diagrama de dispersão

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 4: 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)

$\Sigma X_i Y_i = 4122,8$

Correlação entre X e Y:

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

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

Comentário:

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

Considere o seguinte exemplo: 6 variáveis são medidas em 11 indivíduos.

	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

correlação linear de X e Y1 = 0,816

correlação linear de X e Y2 = 0,816

correlação linear de X e Y3 = 0,816

correlação linear de X4 e Y4 = 0,817

$\Rightarrow$ Mesmos valores de correlação.

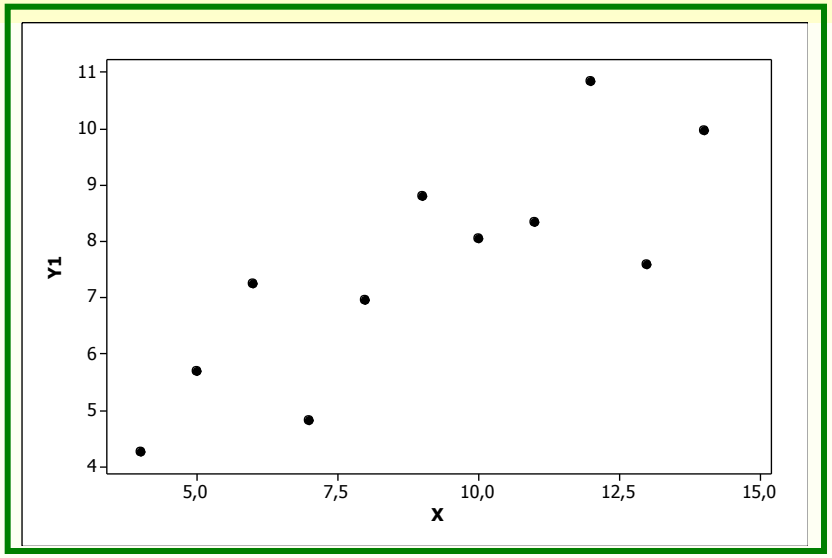
$\Rightarrow$ Qual a forma esperada da dispersão conjunta destas variáveis?

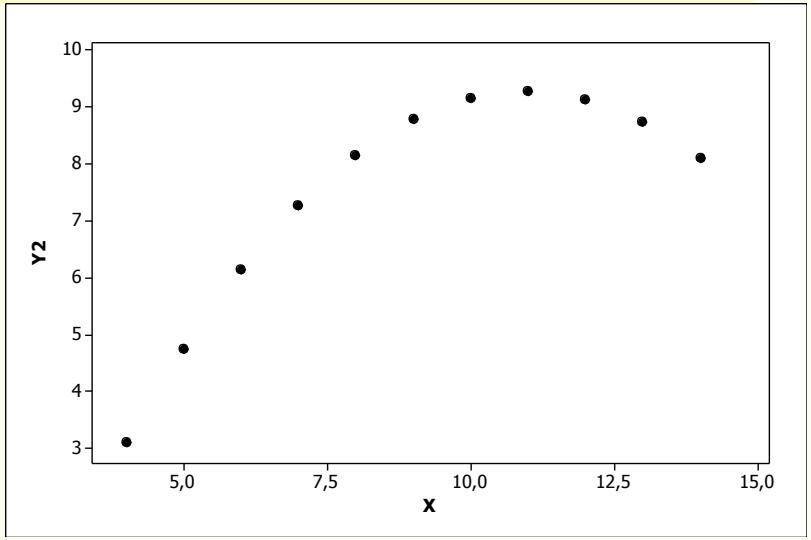
ARQUIVO FA.xls

Diagramas de dispersão e Coeficientes de Correlação

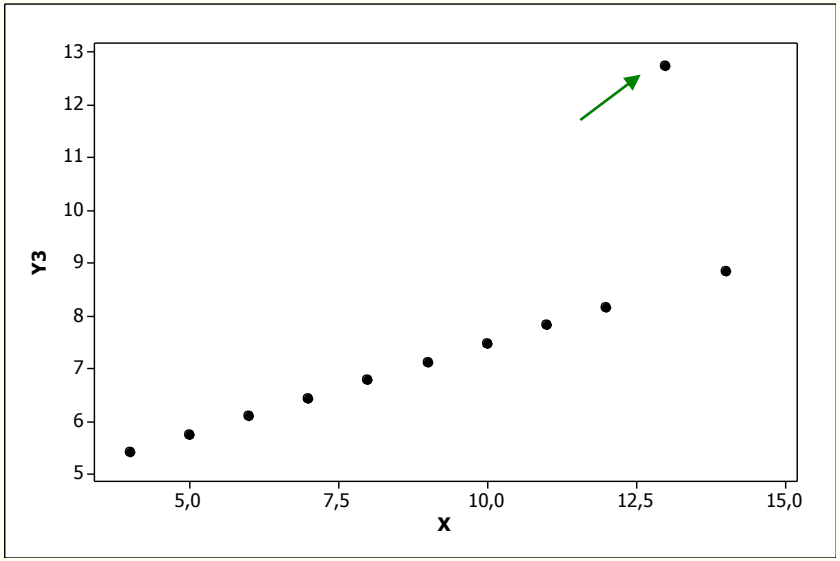
r = 0,816

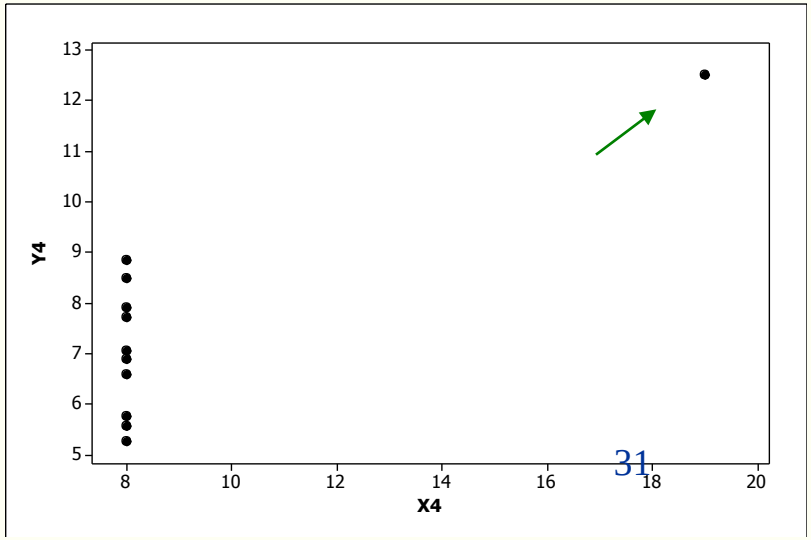
Dispersão esperada!





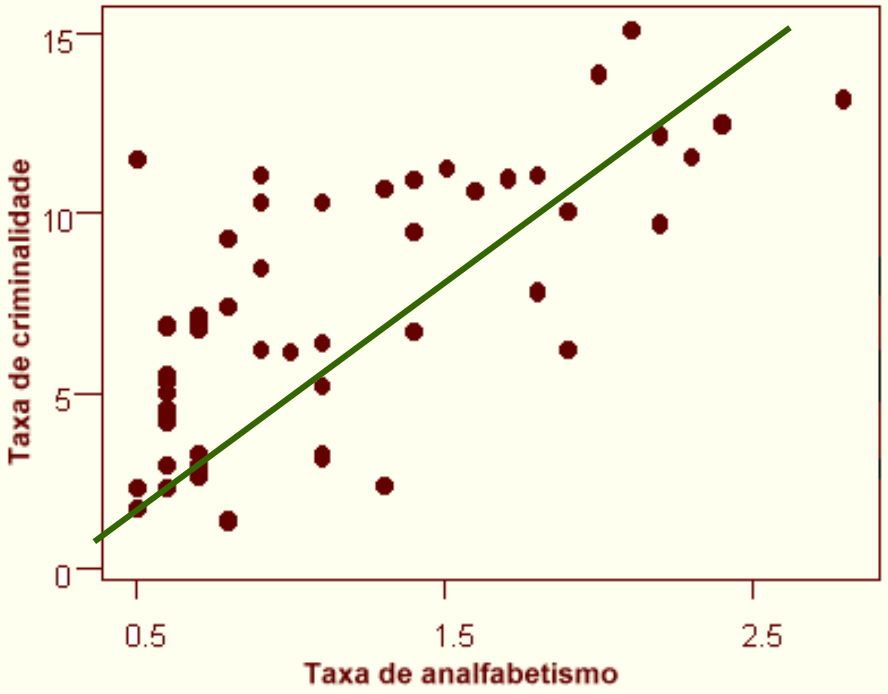
Pontos influentes!

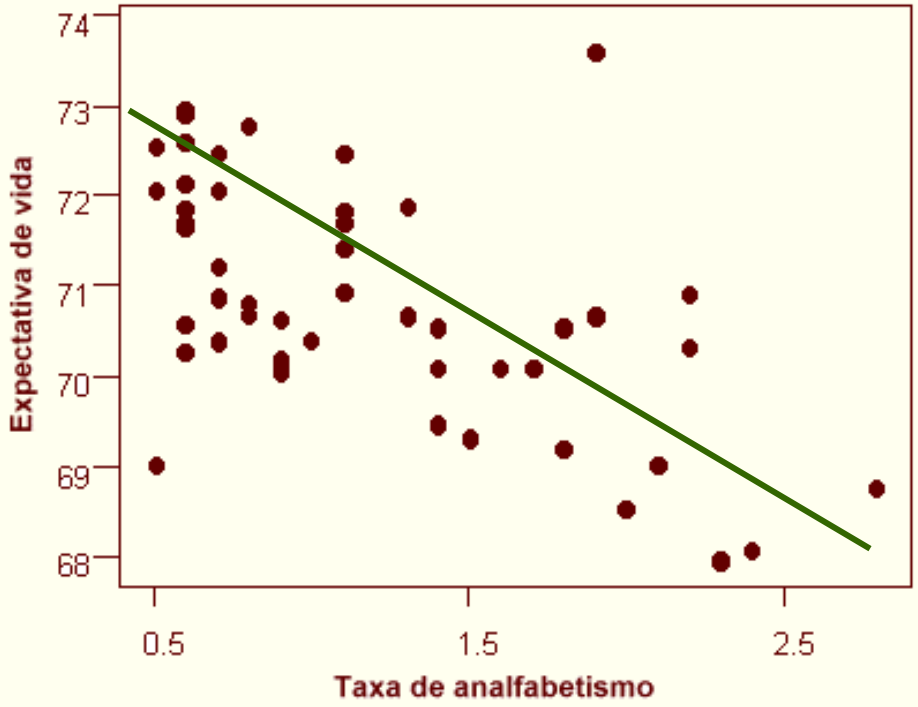




Análise de Regressão

Diagramas de Dispersão





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

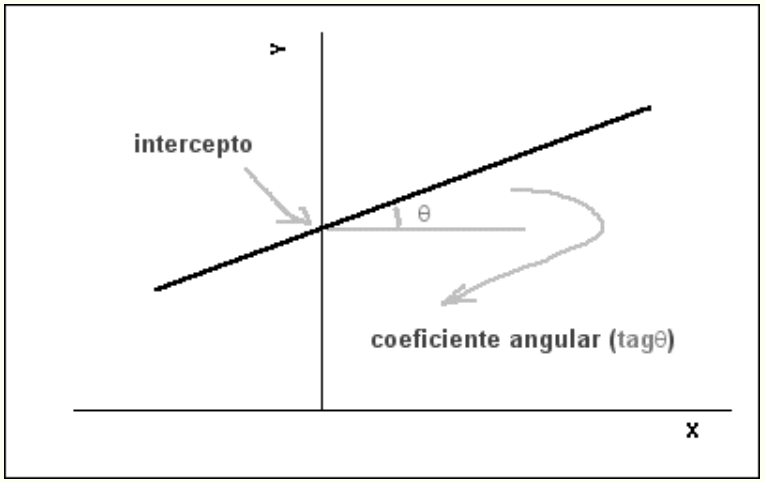
Análise de Regressão

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

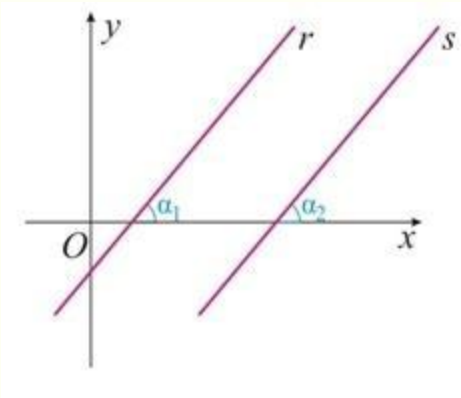
O que são *a* e *b*?

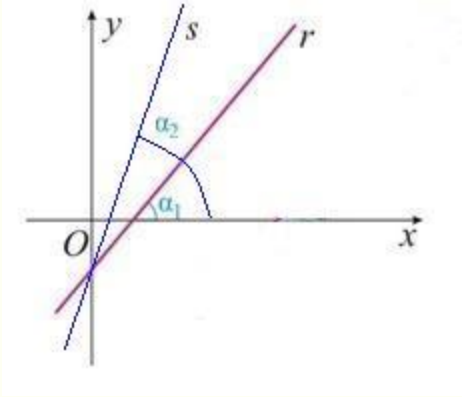
***a* : intercepto**

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



Análise de Regressão





- Iguais coeficientes angulares
- Diferentes interceptos

- Diferentes coeficientes angulares
- Iguais interceptos

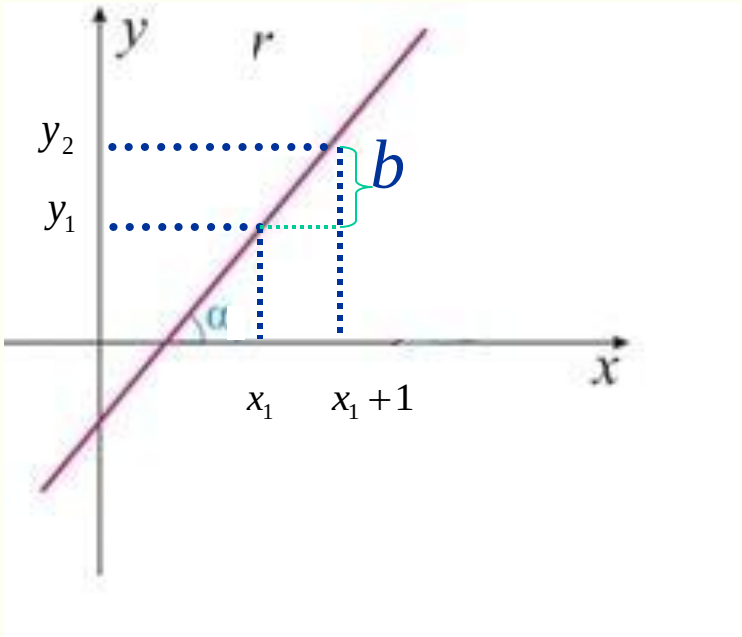
Reta ajustada:

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

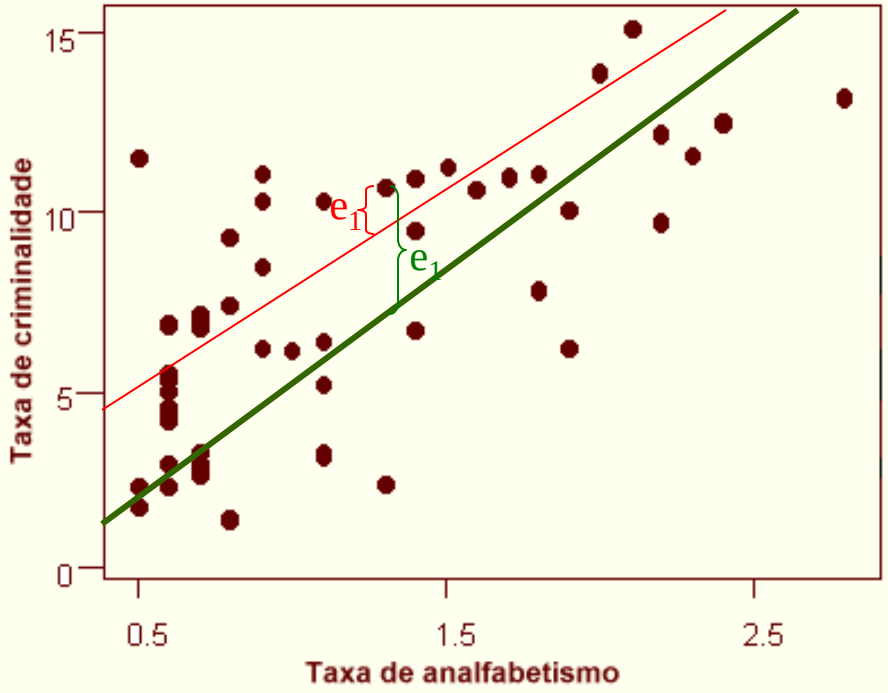
Interpretação de *b*:

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)



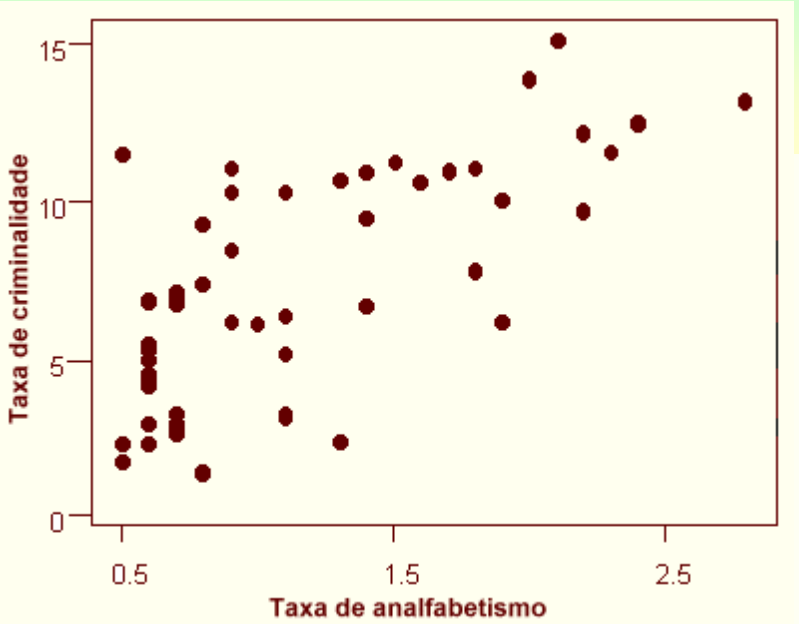
Reta ajustada
(método de mínimos quadrados)

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 3,



A reta ajustada é:

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

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

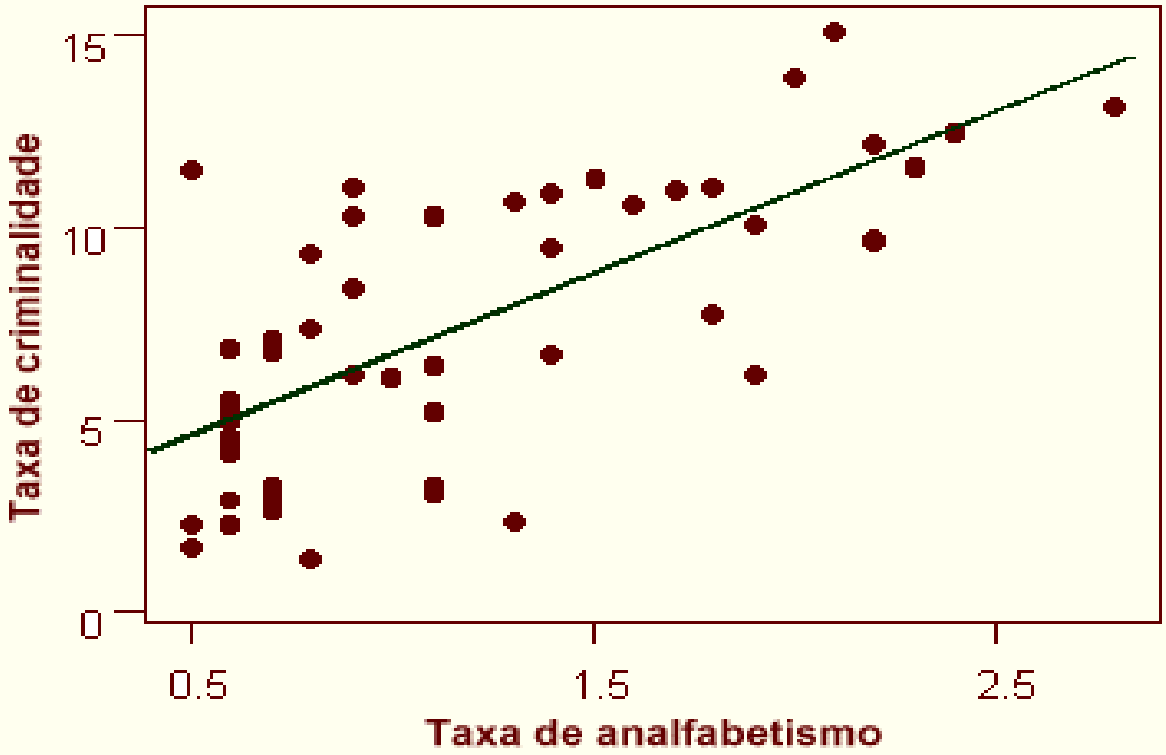
X : taxa de analfabetismo

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

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

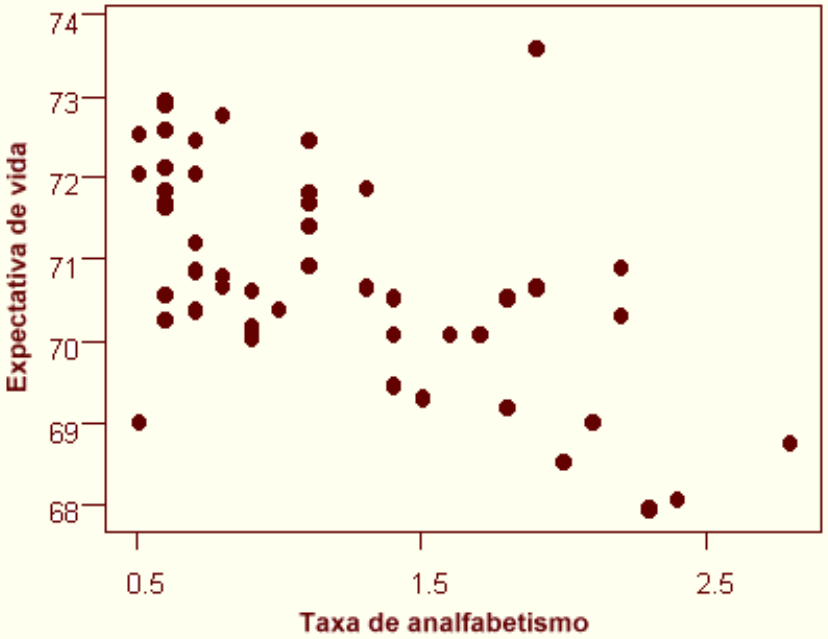


Como desenhar a reta no gráfico?

No exemplo 4,

A reta ajustada é:

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



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

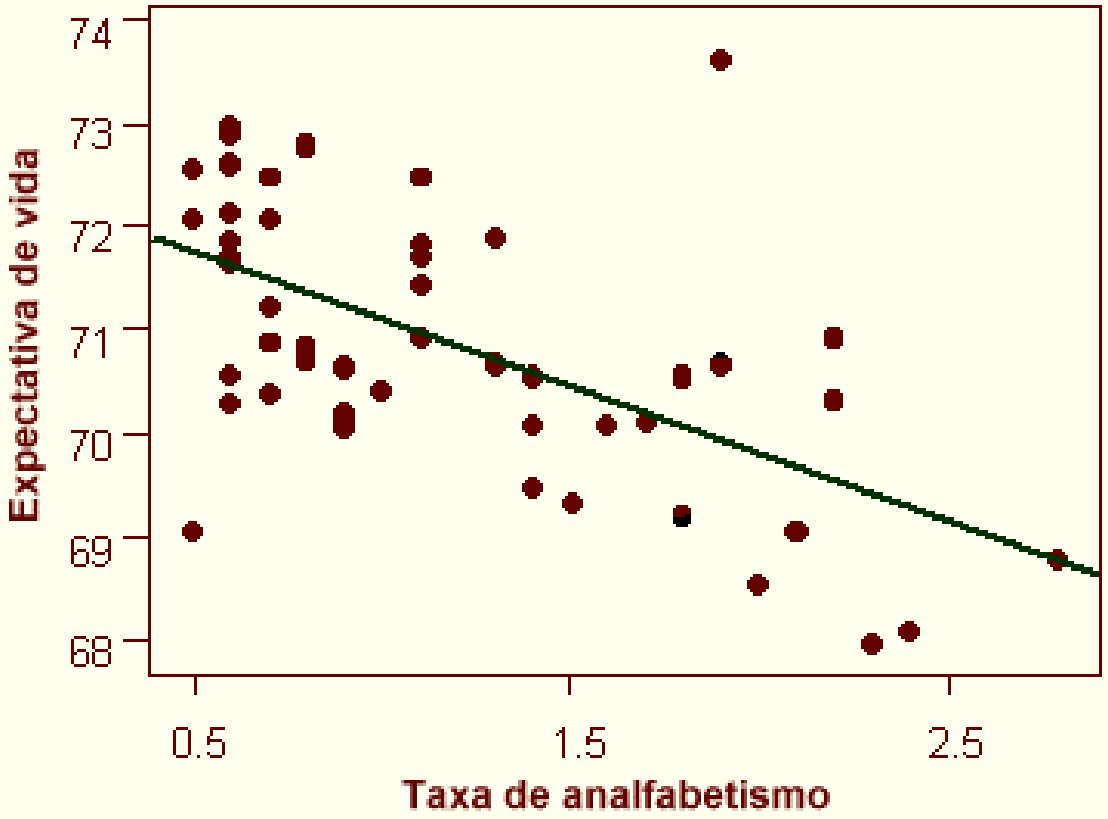
X : taxa de analfabetismo

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 5: consumo de cerveja e temperatura

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

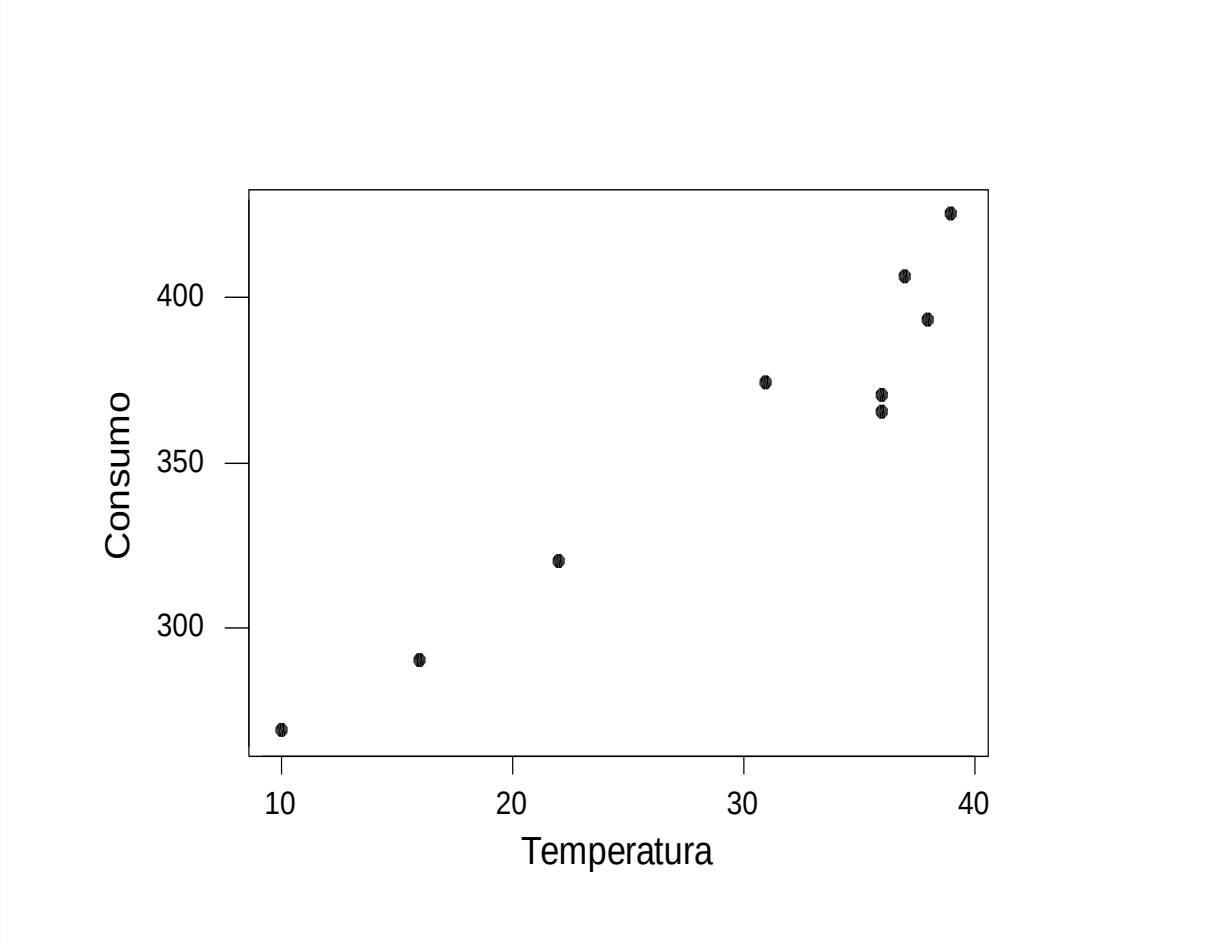
***X*: temperatura máxima (em ºC).**

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

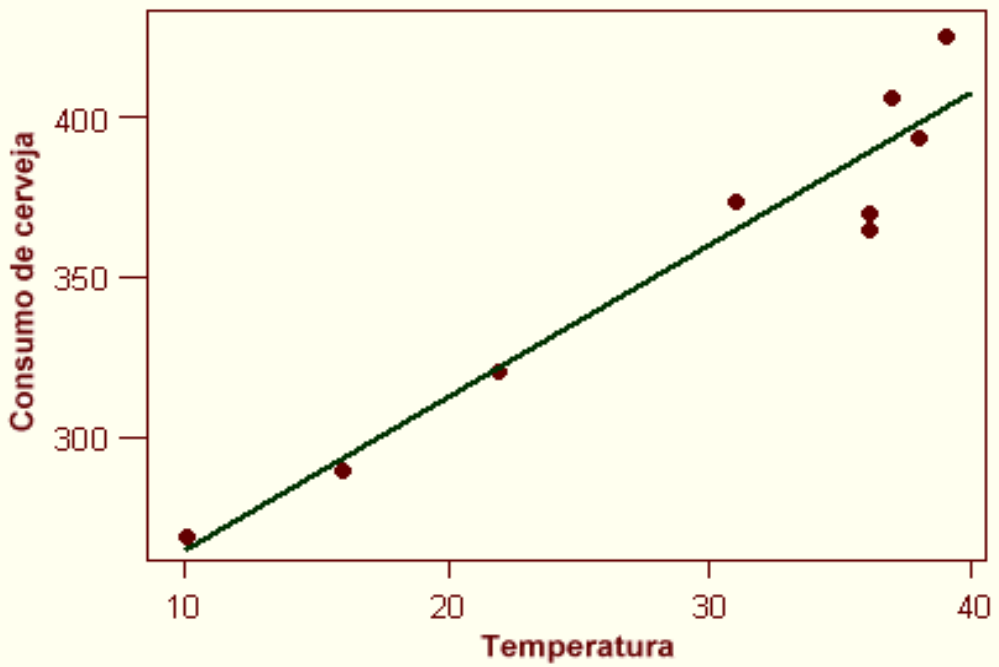
Diagrama de dispersão



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?

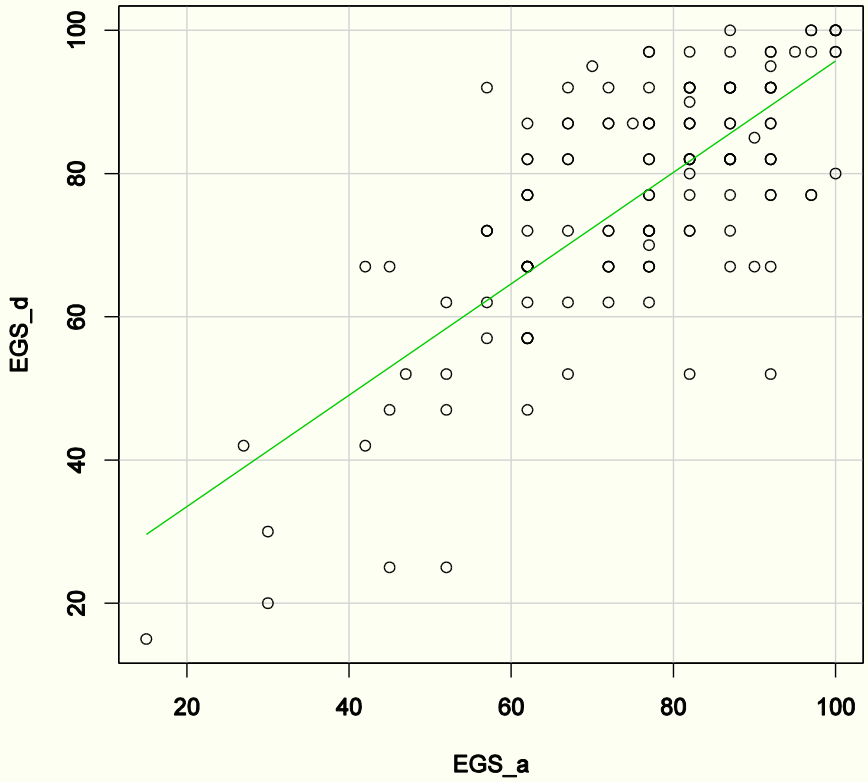
$$\hat{Y} = 217{,}37 + 4{,}74\,(25) = 335{,}87 \text{ litros}$$

Exemplo no R

O arquivo CEA05P11.xls contém dados sobre o projeto “Avaliação de um trabalho de Ginástica Laboral implantado em algumas unidades da USP”. Consideremos as variáveis:

Estado Geral de Saúde antes (EGS_a) e **Estado Geral de Saúde depois** (EGS_d): auto-avaliação do funcionário a respeito do seu estado de saúde antes e depois do início das atividades respectivamente. Quanto maior o índice, melhor a avaliação.

Gráficos → Diagrama de Dispersão (variável-x: EGS_a ; variável-y: EGS_d; marcar opção *Linha de quadrados mínimos*)



Estatísticas → Ajuste de Modelos → Regressão Linear (variável resposta: EGS_d ; variável explicativa: EGS_a)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.94397    4.54712   3.946 0.000125 ***
EGS_a        0.77791    0.05894  13.198  < 2e-16 ***
```

a=17,94397, b=0,77791

$$\hat{Y} = 17{,}94397 + 0{,}77791\, EGS\_a$$