

Estatística Descritiva III

Análises de Correlação e Regressão

Associação entre variáveis qualitativas



Tabelas de Contingência

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

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

- Estudo realizado pelo Depto 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?

- **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.

- **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*;

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

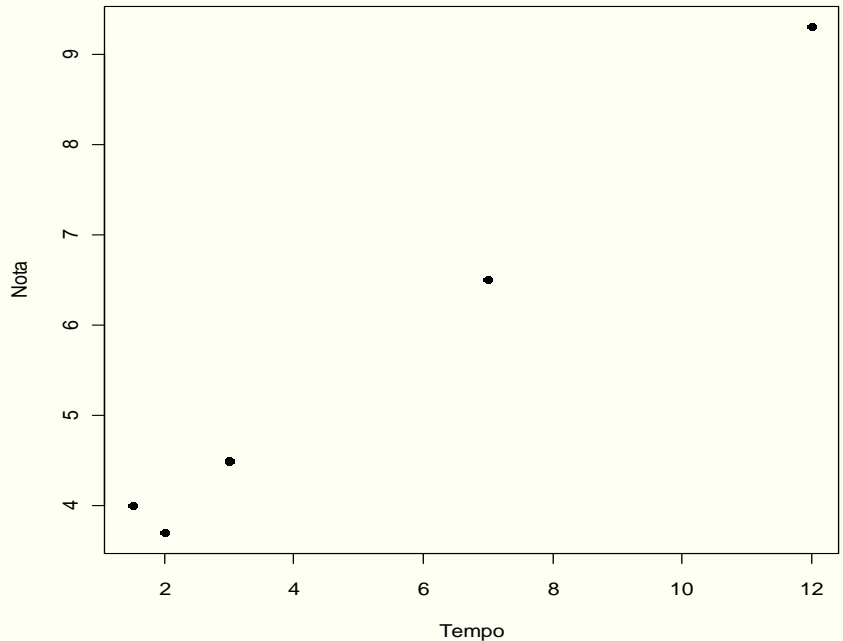
X : tempo de estudo (em horas)

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



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): é uma auto-avaliação do funcionário a respeito do seu estado de saúde antes do início das atividades (quanto maior o índice, melhor o EGS);
- Vitalidade antes (Vit_a): avalia a influência do cansaço e fadiga nos trabalhos diários antes do início das atividades (quanto maior o índice, melhor);

No R

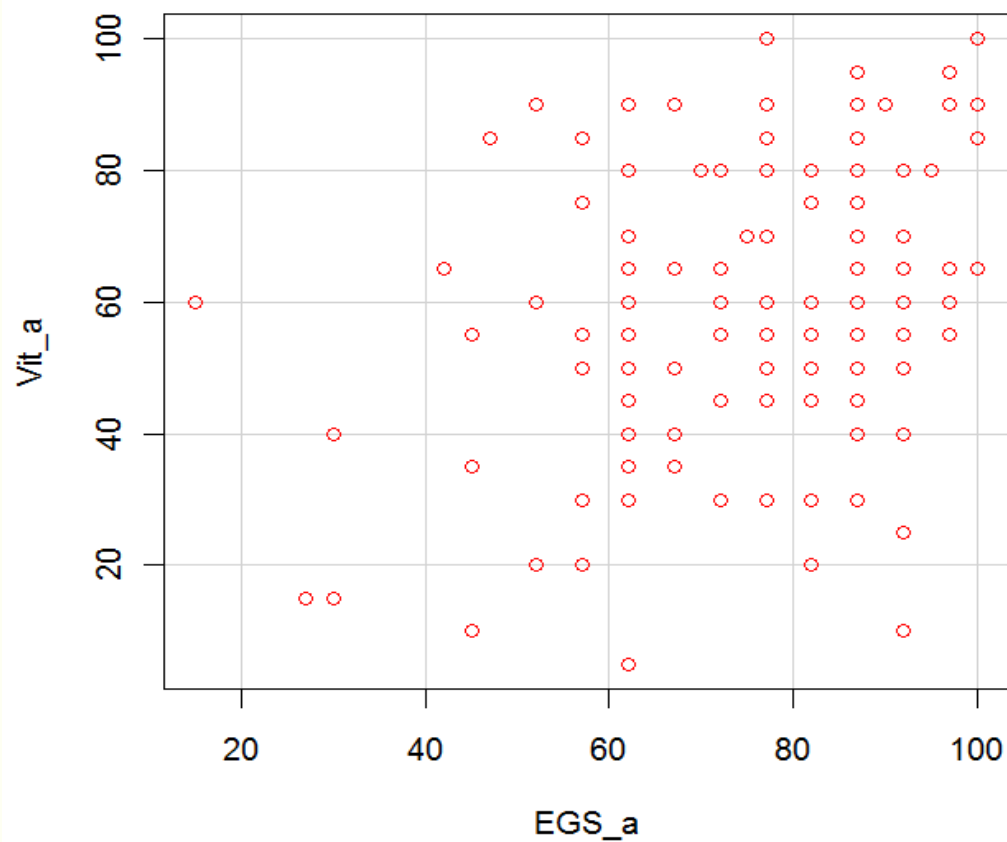
The image shows the R Commander interface. The 'Gráficos' menu is open, showing options like 'Gradiente de cores (color palette)', 'Gráfico por Ordem de Apresentação (Index Plot)...', 'Histograma...', 'Diagrama de ramo-e-folhas...', 'Boxplot', 'Gráfico de comparação de quantis...', and 'Diagrama de dispersão...'. The 'Diagrama de dispersão...' option is highlighted. Below it, the 'Gráfico de Dispersão' dialog box is open, showing a list of variables for 'variável-x (escolha uma)' and 'variável-y (escolha uma)'. The 'variável-x' list includes CF_d, Dor_a, Dor_d, and EGS_a. The 'variável-y' list includes SM_a, SM_d, Vit_a, and Vit_d. The 'Vit_a' variable is selected in the 'variável-y' list. The dialog also has an 'Options' section with checkboxes for 'Identifique pontos', 'Deslocamentos (Jitter) na variável-x', 'Deslocamentos (Jitter) na variável-y', 'Log eixo-x', 'Log eixo-y', 'Boxplots marginais', 'Linha de quadrados mínimos', 'Smooth line', and 'Mostre espalhamento (spread)'. The 'Parâmetros do gráfico' section includes 'Caracteres do gráfico' (set to '<auto>'), 'Tamanho do ponto' (set to 1.0), 'tamanho do texto no eixo' (set to 1.0), and 'tamanho do texto - rótulo do eixo' (set to 1.0). The 'Definição para a suavização (Smooth)' is set to 50. The 'rótulo do eixo-x' and 'rótulo do eixo-y' are both set to '<auto>'. In the background, the R Commander window shows the 'Conjunto de Dados: CEA05P11' and the 'Janela do Script' with the following code:

```
CEA05P11 <- sqlQuery(chan  
names(CEA05P11) <- make.n
```

The 'Janela de Resultados' shows the following code:

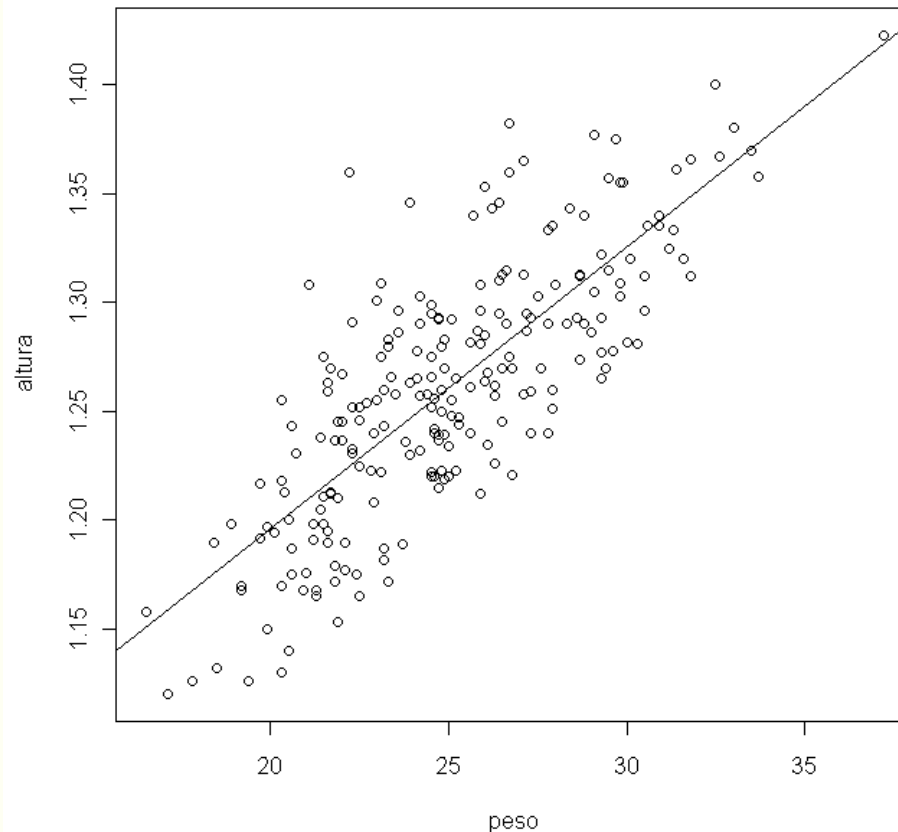
```
> CEA05P11 <- sqlQuery(ch  
> names(CEA05P11) <- make
```

Gráfico de Dispersão



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:

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

em que,

\bar{X} e \bar{Y} são as médias de X e Y, respectivamente;

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

Fórmula alternativa:

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

$$S^2 = \frac{\sum_{i=1}^n X_i^2 - n \bar{X}^2}{(n-1)}$$

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

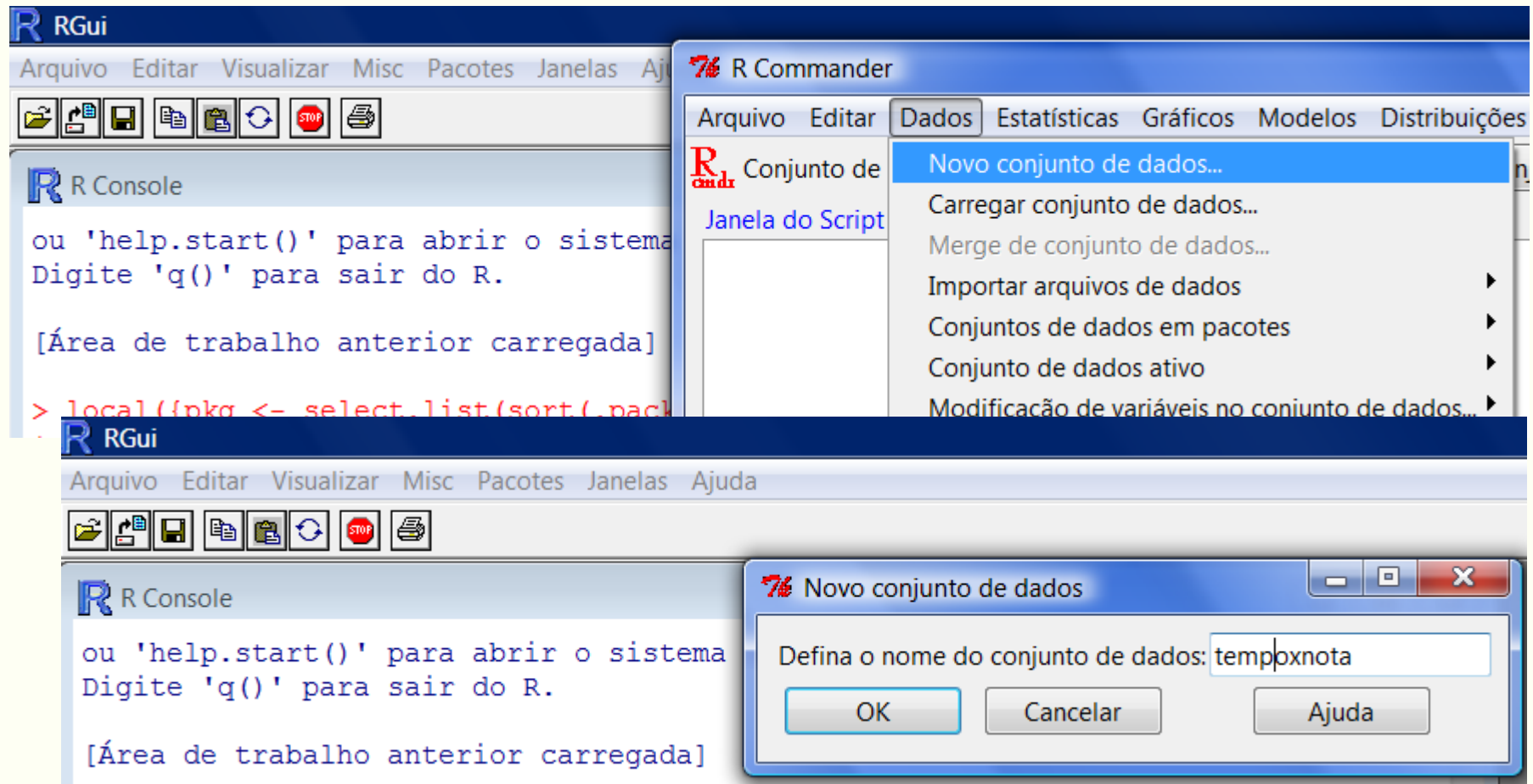
$$S_x^2 = \frac{(-2,1)^2 + \dots + (6,9)^2}{4} = \frac{78,2}{4} = 19,55 \Rightarrow S_x = 4,42$$

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

Então,

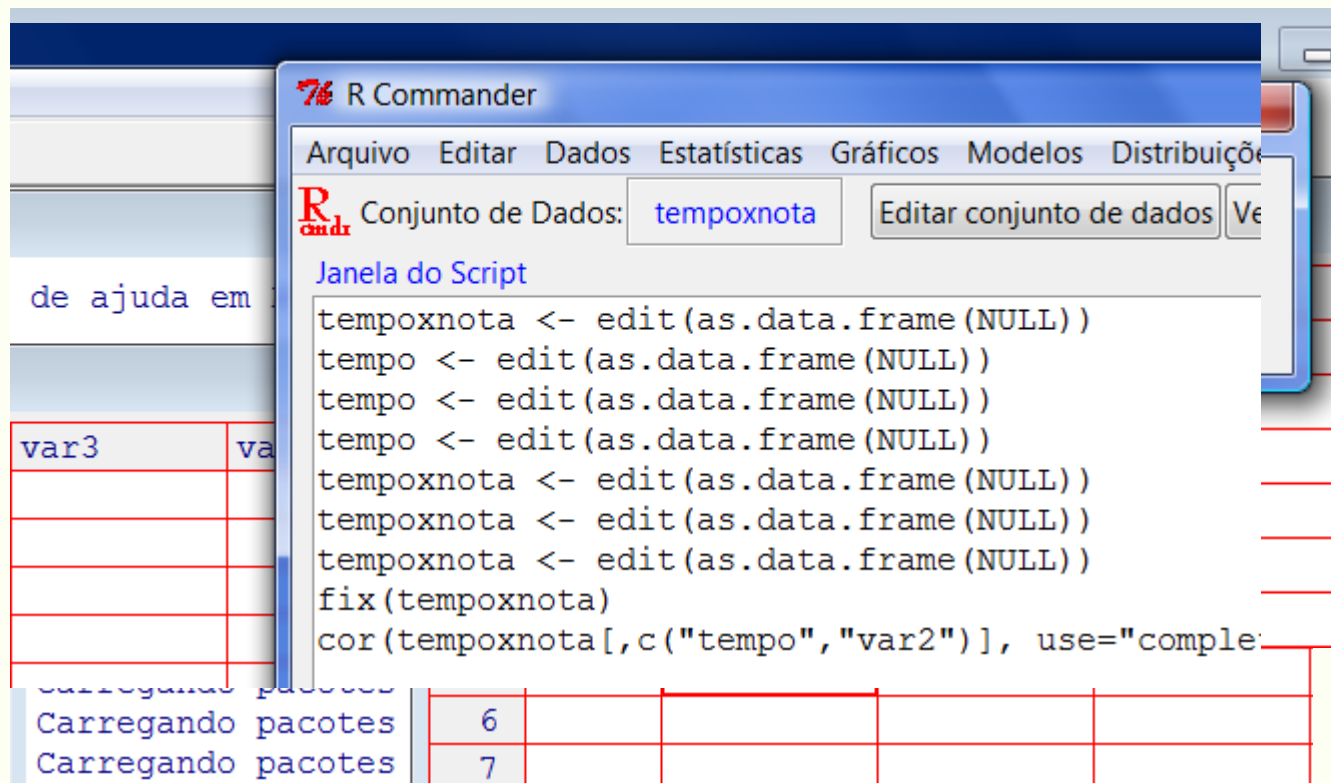
$$r = \frac{41,2}{4 \cdot 4,42 \cdot 2,34} = 0,9959$$

Criando arquivos no R



Criando arquivos no R

Digitar os dados na janela do editor e dar nomes às variáveis e fechar.



No R temos:

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

```
[1] 0.9960249
```

Em que `tempoxnota$Tempo` é a variável `Tempo` do conjunto de dados denominado `tempoxnota`.



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: $-1 \leq r \leq 1$

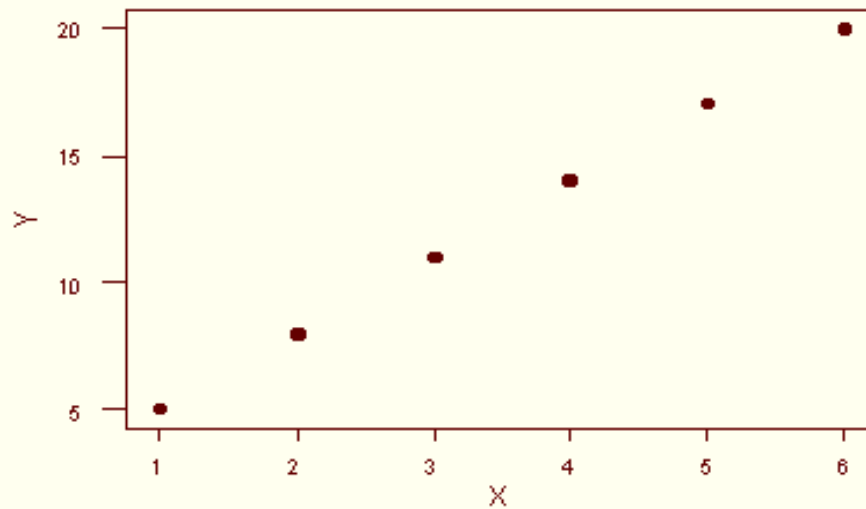
Casos particulares:

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

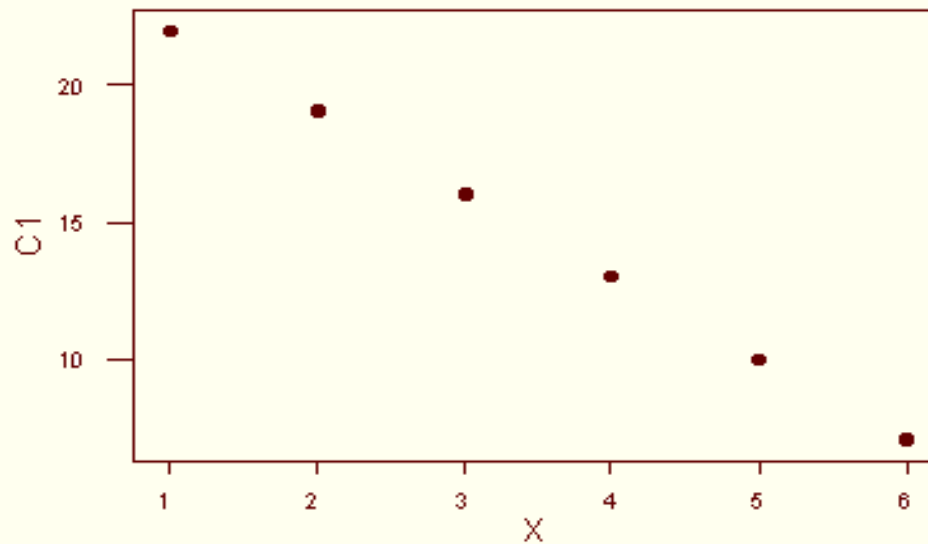
$r = -1 \Rightarrow$ correlação linear negativa e perfeita;

$r = 0 \Rightarrow$ inexistência de correlação linear.

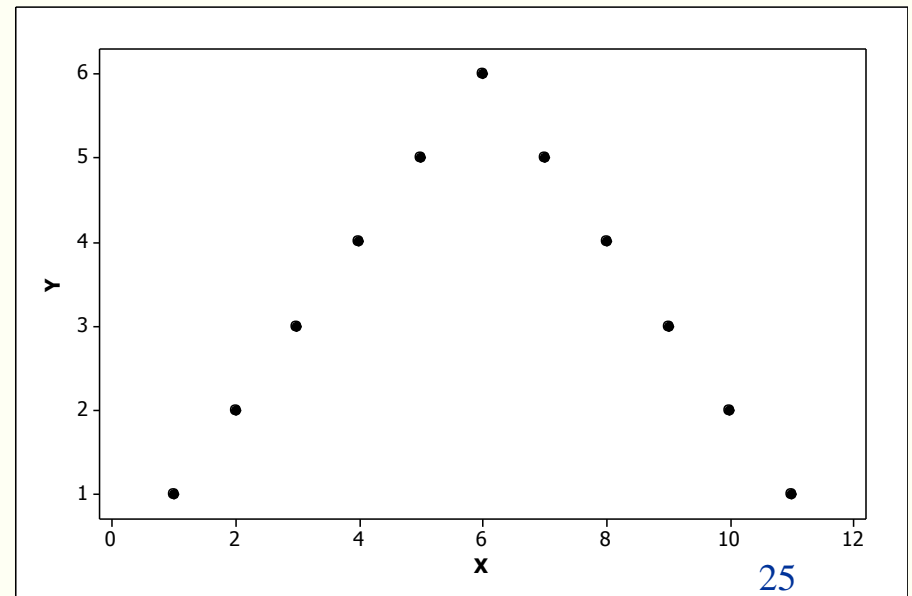
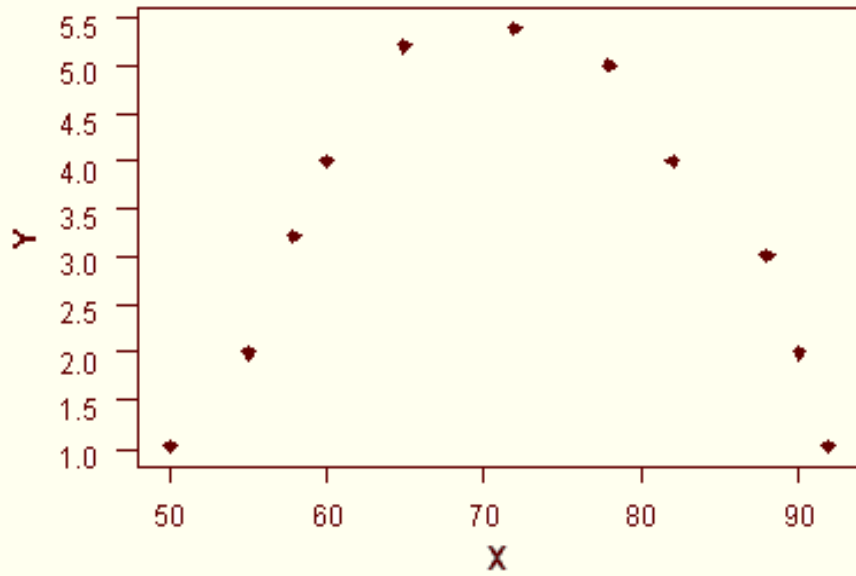
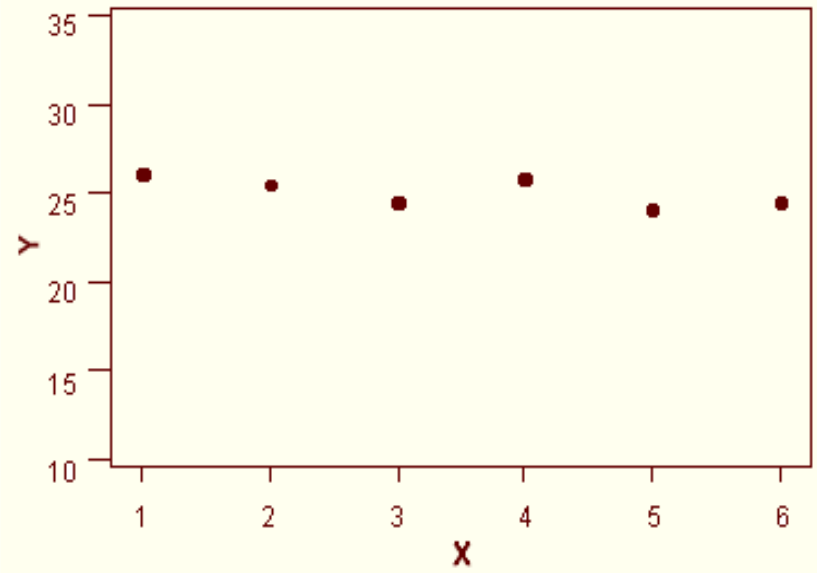
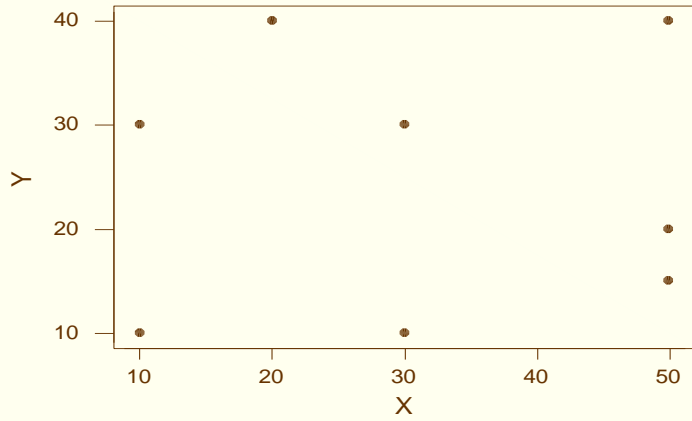
$r = 1$, correlação linear positiva e perfeita



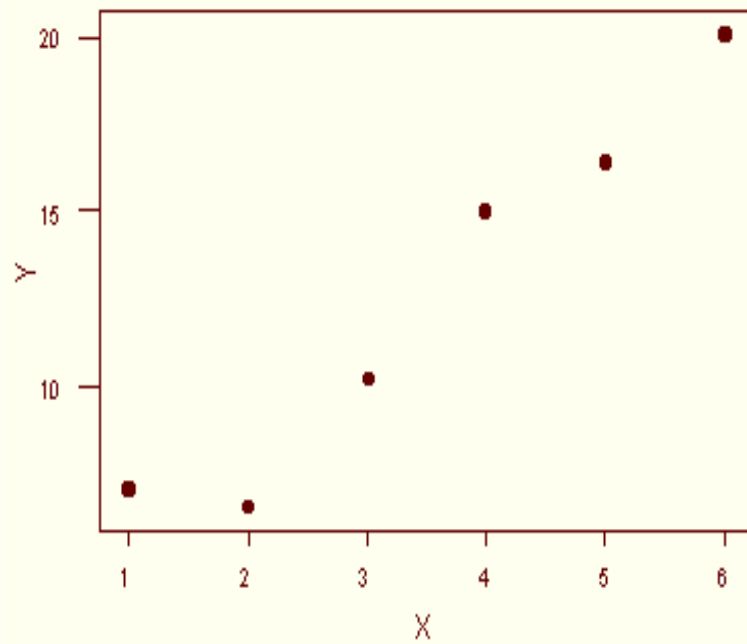
$r = -1$, correlação linear negativa e perfeita



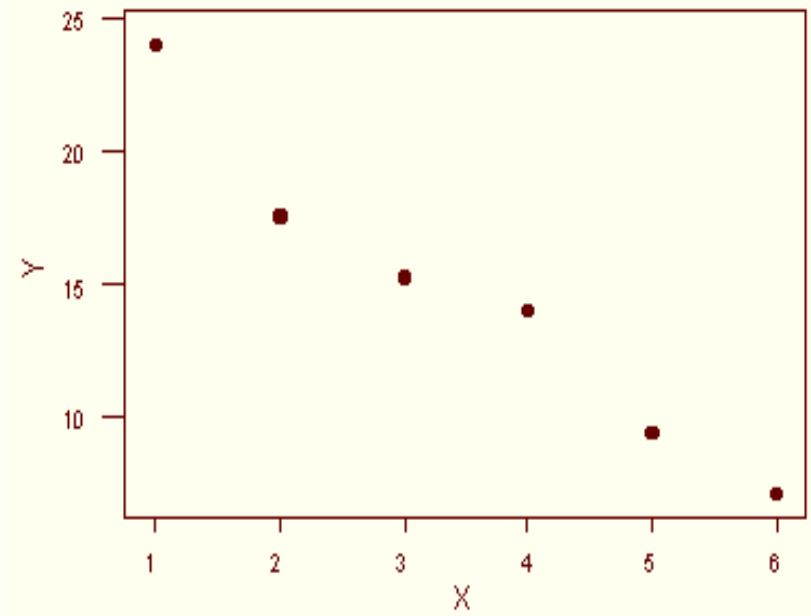
$$r \cong 0$$



$$r \cong 1$$



$$r \cong -1$$



Exemplo 3: criminalidade e analfabetismo

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

Y: taxa de criminalidade

X: taxa de analfabetismo

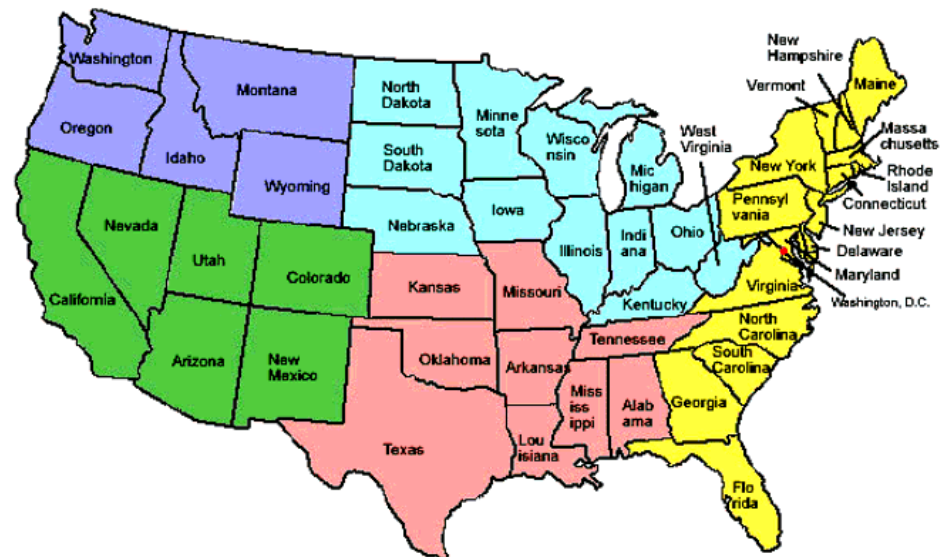
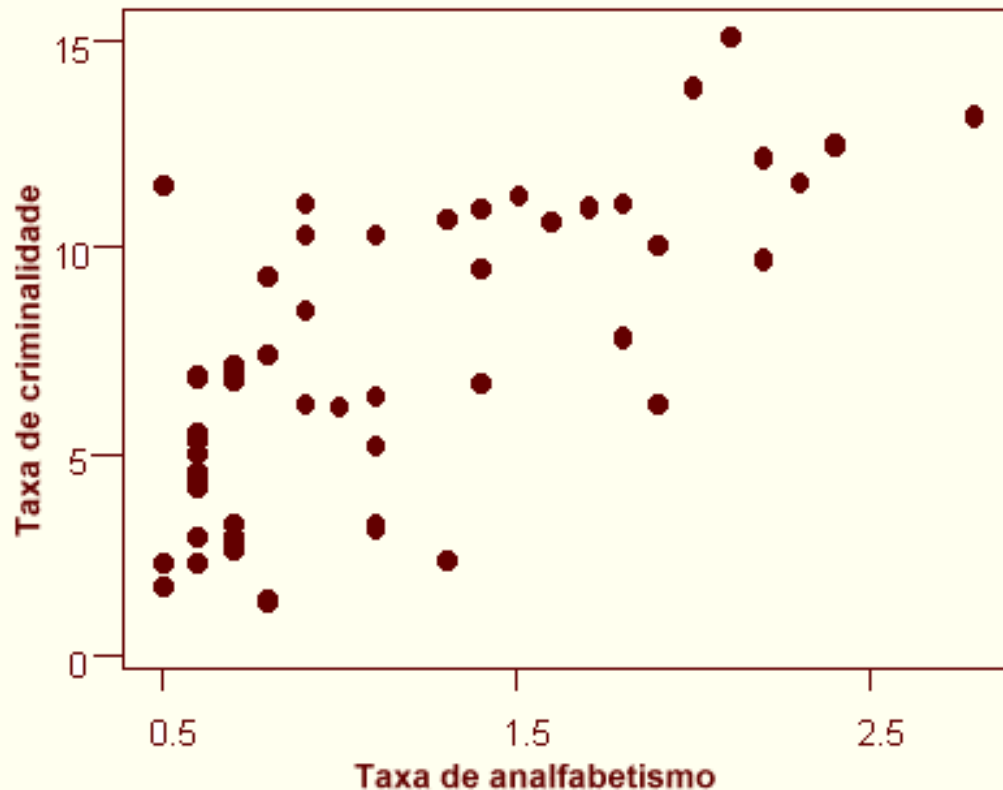


Diagrama de Dispersão



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:

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

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

Exemplo 4: expectativa de vida e analfabetismo

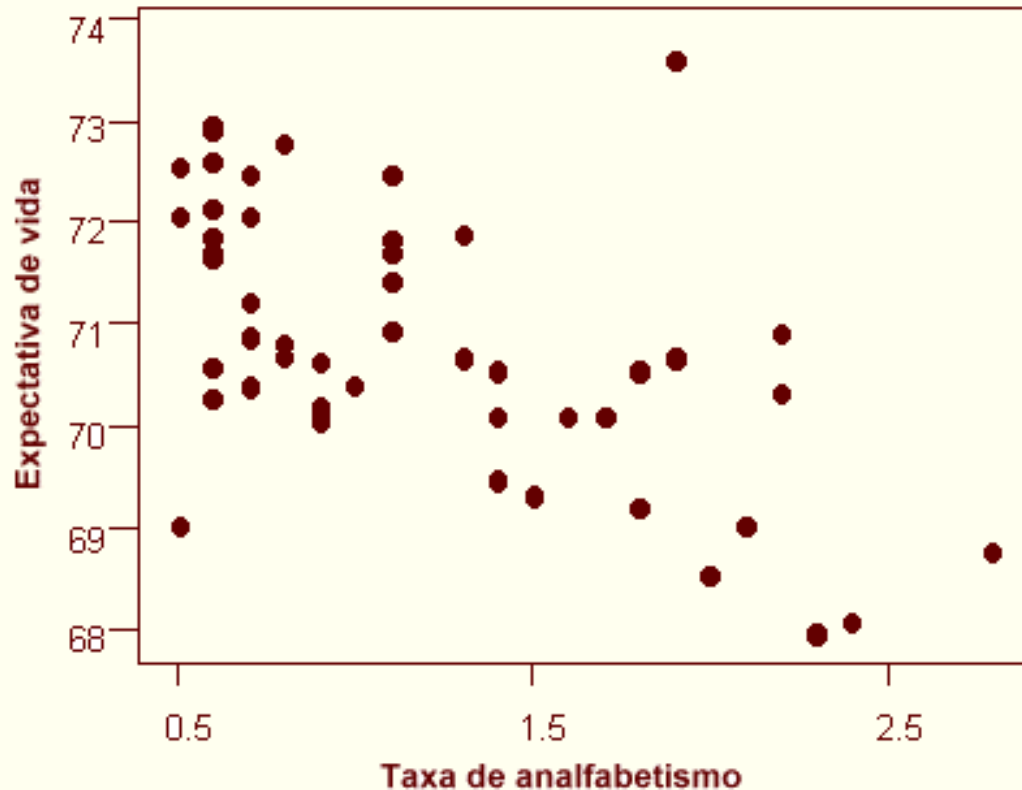
Considere as duas variáveis observadas em 50 estados norte-americanos.

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)

$\sum X_i Y_i = 4122,8$

Correlação entre X e Y:

$$r = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{(n-1) S_X S_Y}$$
$$r = \frac{4122,8 - 50 \cdot 70,88 \cdot 1,17}{49 \cdot 1,342 \cdot 0,609} = \frac{-23,68}{40,047} = -0,59$$

Comentário:

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

Suponha o seguinte exemplo: 6 variáveis são medidas em 11 indivíduos

Id	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 de Pearson entre X e Y1 = 0,816

Correlação de Pearson entre X e Y2 = 0,816

Correlação de Pearson entre X e Y3 = 0,816

Correlação de Pearson entre X e Y4 = 0,817

⇒ Mesmos valores de correlação.

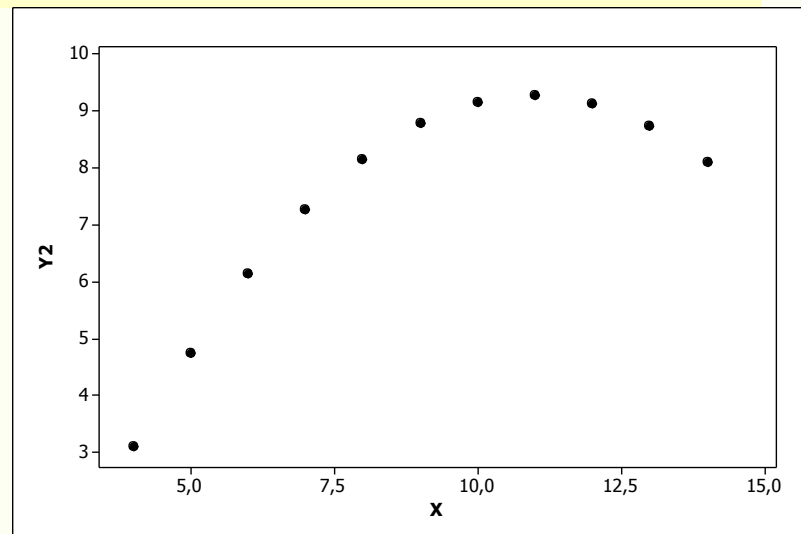
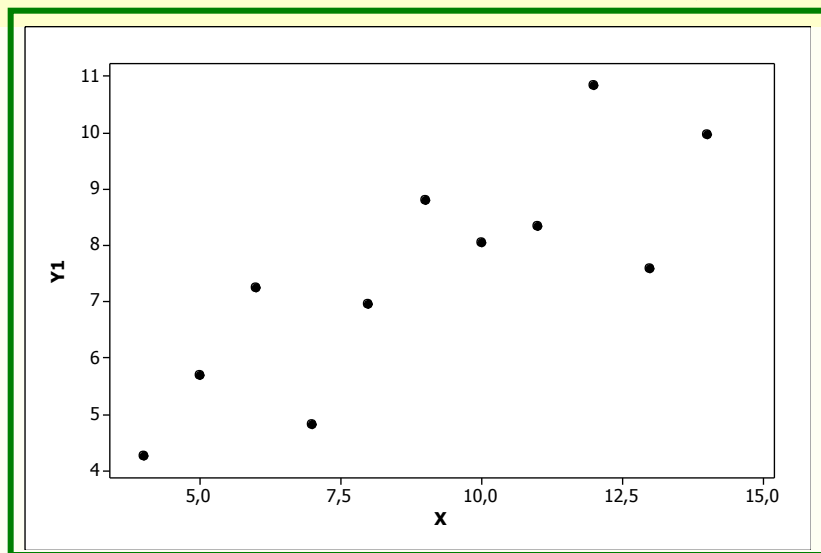
⇒ Qual a forma esperada da dispersão conjunta destas variáveis?

ARQUIVO FA.MTW

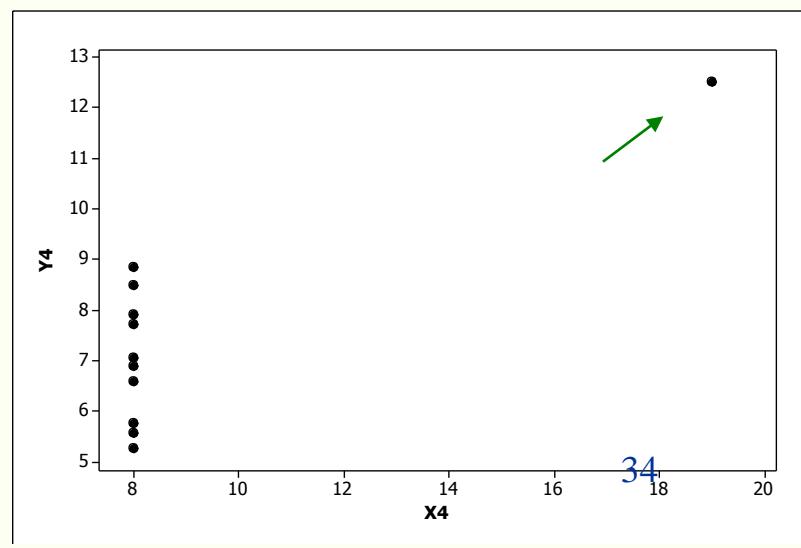
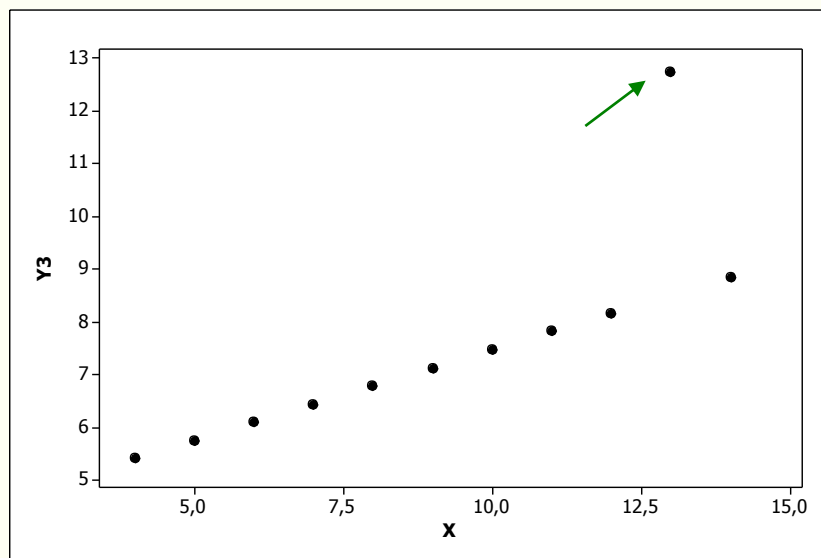
Diagramas de Dispersão e Coeficientes de Correlação

$$r = 0,816$$

Dispersão
esperada!

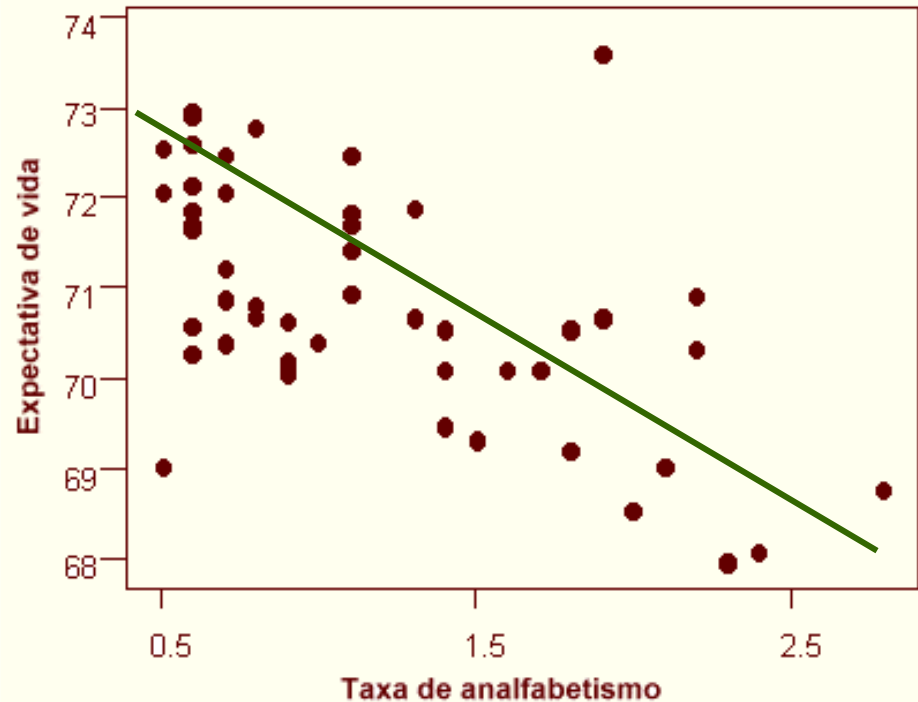
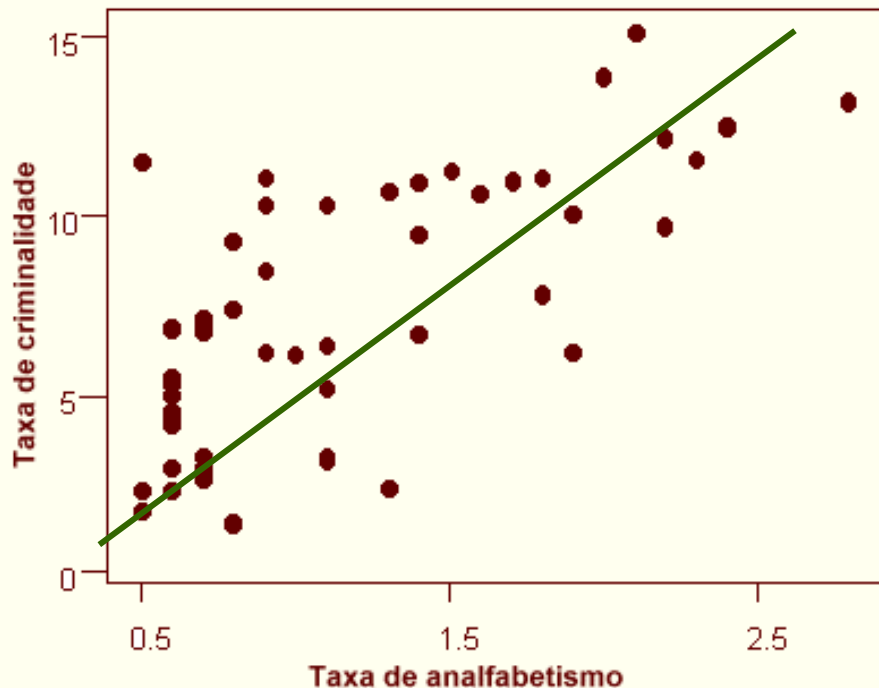


Pontos
influentes!



Análise de Regressão

Diagramas de Dispersão



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

Análise de Regressão

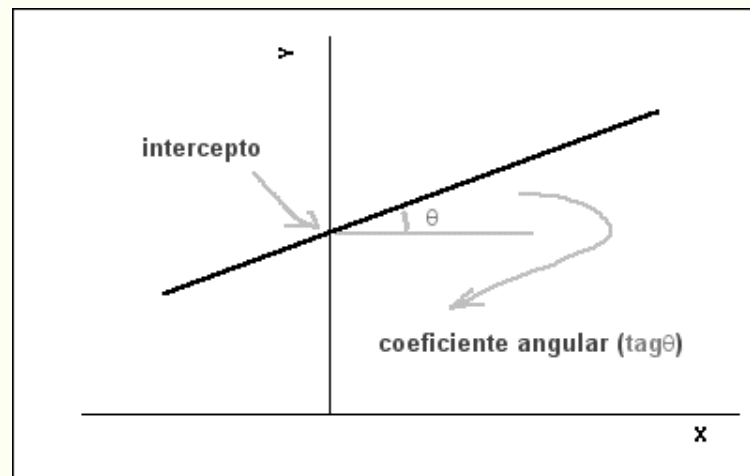
Reta ajustada:

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

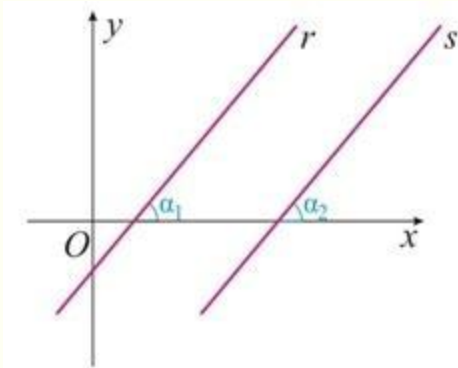
O que são a e b ?

a : intercepto

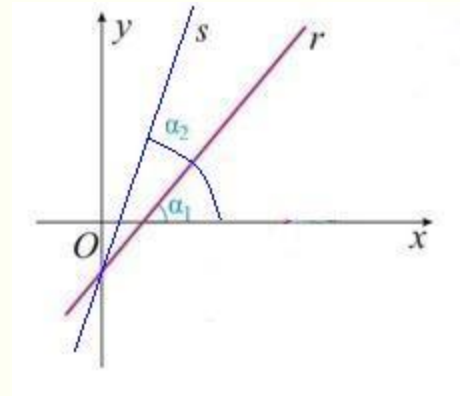
b : inclinação ou coeficiente angular



Análise de Regressão



- Iguais coeficientes angulares
- Diferentes interceptos



- Diferentes coeficientes angulares
- Iguais interceptos

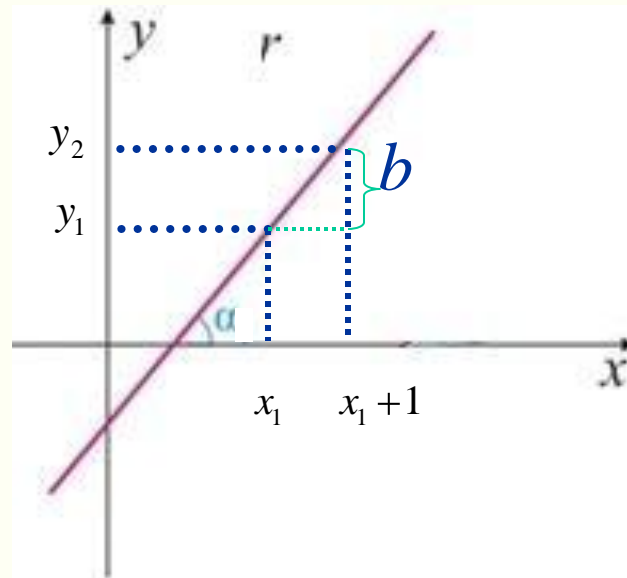
Reta ajustada:

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

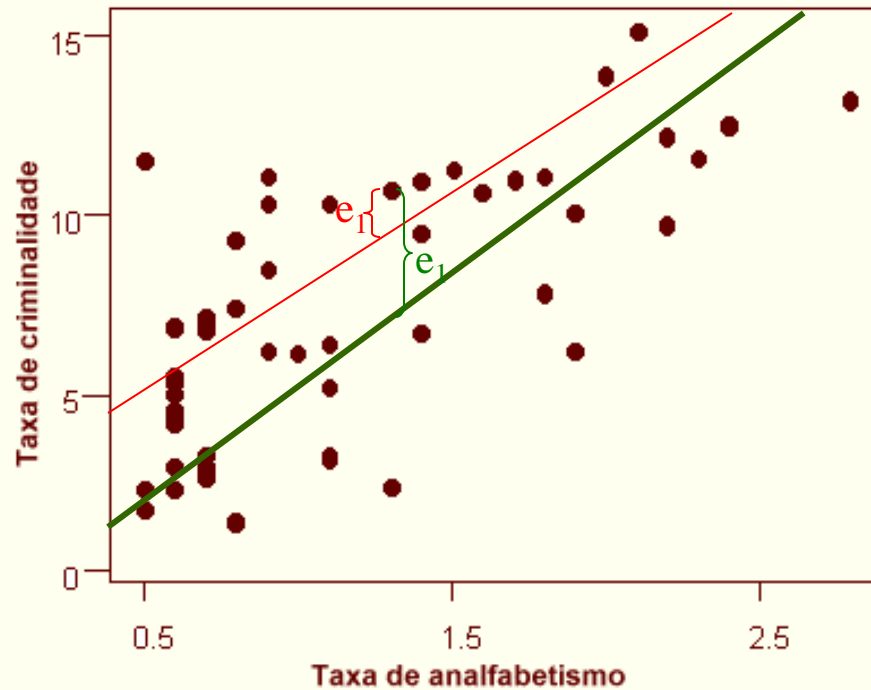
Interpretação de b :

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

$$\begin{aligned} \text{tag}(\alpha) &= \frac{y_2 - y_1}{x_2 - x_1} = \frac{y_2 - y_1}{x_1 + 1 - x_1} \\ &= y_2 - y_1 = b \end{aligned}$$



Reta ajustada (método de mínimos quadrados)



Reta ajustada (método de mínimos quadrados)

Os coeficientes a e b são calculados da seguinte maneira:

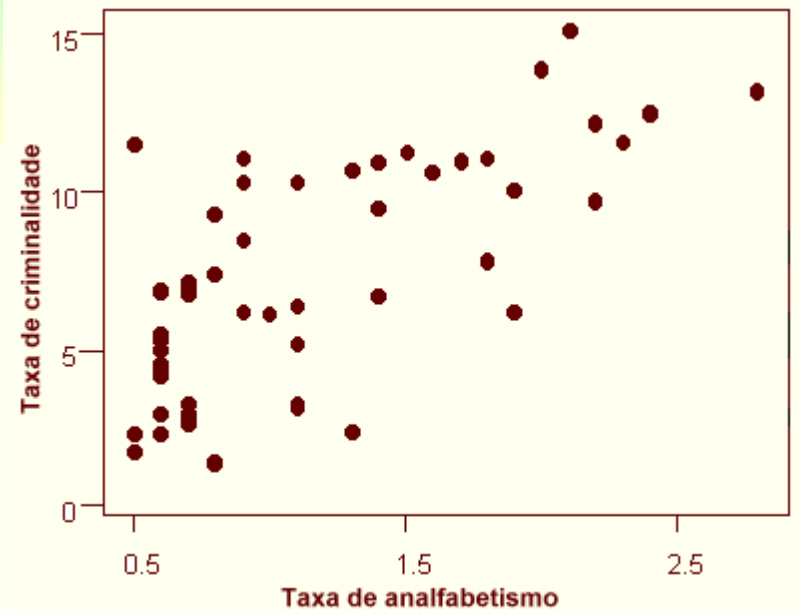
$$b = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{n - 1 \bar{S}_X^2}$$

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

No Exemplo 3,

A reta ajustada é:

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



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

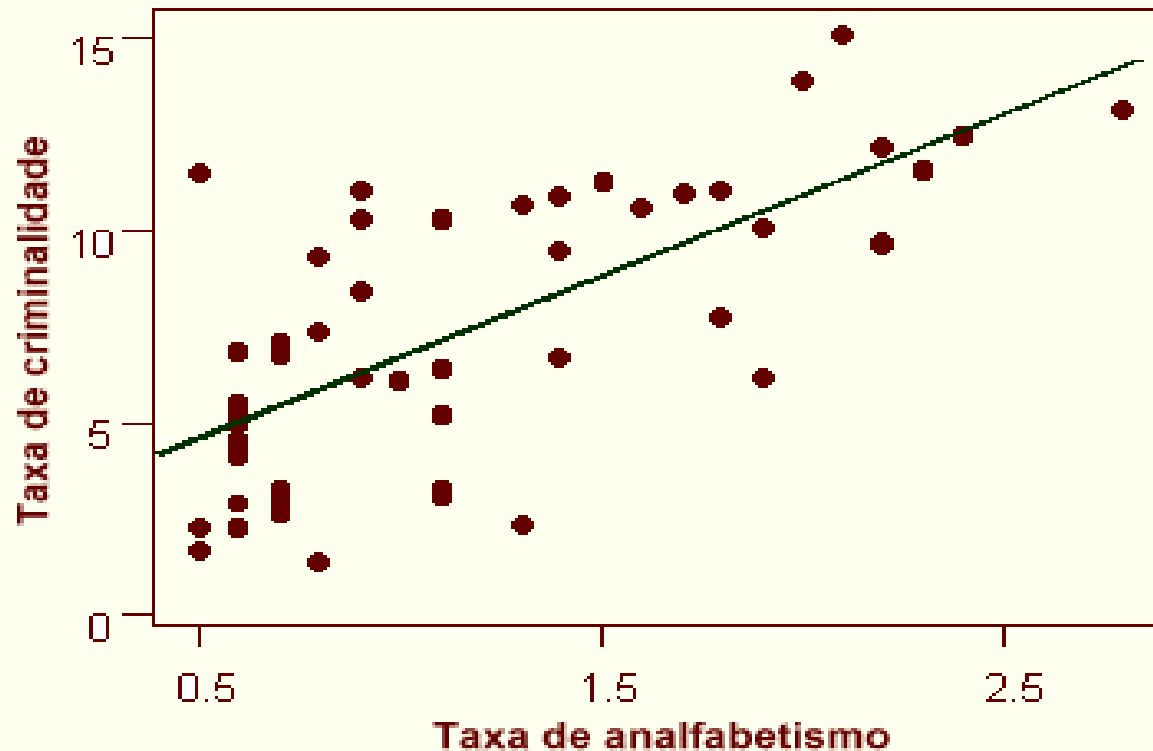
X : taxa de analfabetismo

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

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

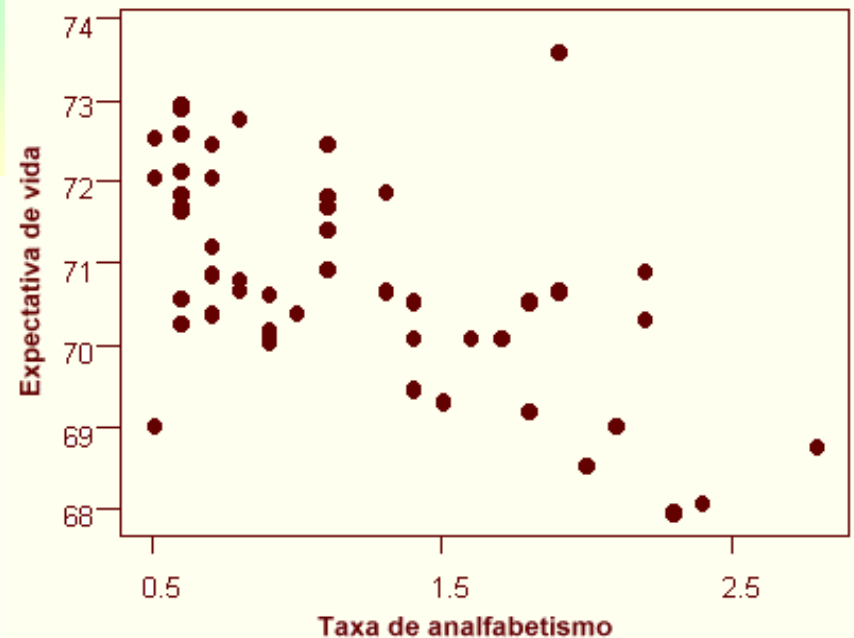


Como desenhar a reta no gráfico?

No exemplo 4,

A reta ajustada é:

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



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

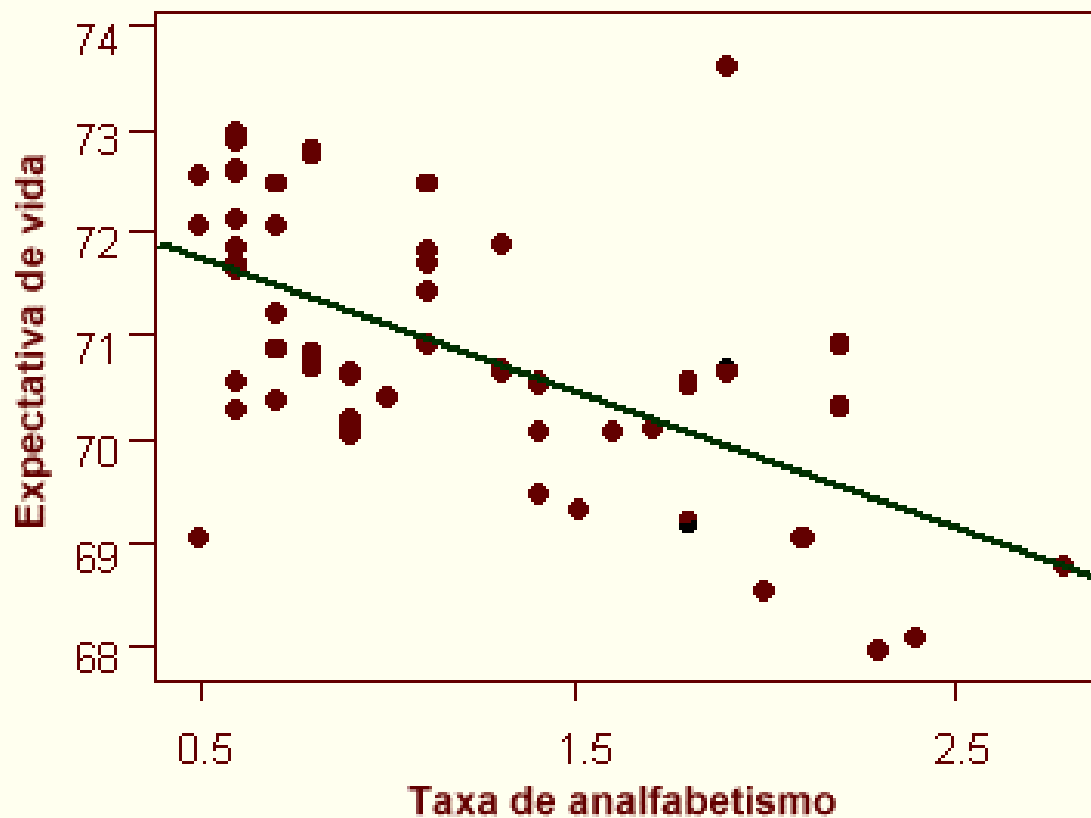
X : taxa de analfabetismo

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

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



Continuando com o arquivo CEA05P11

The image shows the R Commander interface with the 'Estatísticas' menu open, highlighting 'Ajuste de Modelos'. The 'Regressão Linear' dialog box is also open, showing the model name 'RegModel.3', the response variable 'Vit_a', and the explanatory variables 'EGS_a', 'EGS_d', 'Idade', and 'Individuo'.

R Commander

Arquivo Editar Dados Estatísticas Gráficos Modelos Distribuições Ferramentas Ajuda

Conjunto de Dados: [dados] Ver conjunto de dados Modelo: <sem modelo>

Janela do Script

```
CEA05P11 <- sqlQuery(conexao, "select * from [Dados_CEA05P11 (1)$] where 1=1")
names(CEA05P11) <- c("EGS_a", "EGS_d", "Idade", "Individuo", "Vit_a", "Vit_d")
```

Estatísticas

- Resumos
- Tabelas de Contingência
- Médias
- Frequências/Proporções
- Variâncias
- Testes Não-Paramétricos
- Análise Dimensional
- Ajuste de Modelos**

Ajuste de Modelos

- Regressão Linear...**
- Modelo Linear...
- Modelo Linear Generalizado (GLM)...
- Modelo Logito Multinomial...
- Modelo de regressão ordinal...

Regressão Linear

Defina um nome p/ o modelo: RegModel.3

Variável resposta (escolha 1): Vit_a

Variáveis Explicativas (escolha 1 ou mais): EGS_a, EGS_d, Idade, Individuo

Expressão (subset expression): <todos casos válidos>

OK Cancelar Ajuda

select * from [Dados_CEA05P11 (1)\$] where 1=1

CEA05P11))

Saída

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	27.56474	7.56471	3.644	0.000377	***
EGS_a	0.41886	0.09806	4.272	3.55e-05	***

a



b



$$\hat{Y} = 27,56474 + 0,41886\text{EGS_a}$$

Exemplo 5: consumo de cerveja e temperatura

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

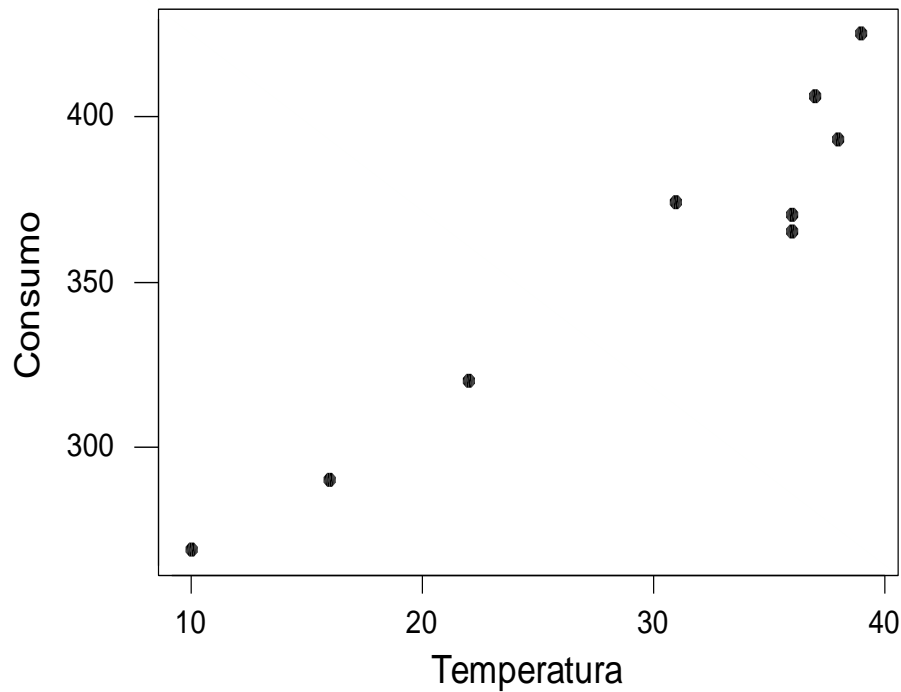
X : temperatura máxima (em °C).

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

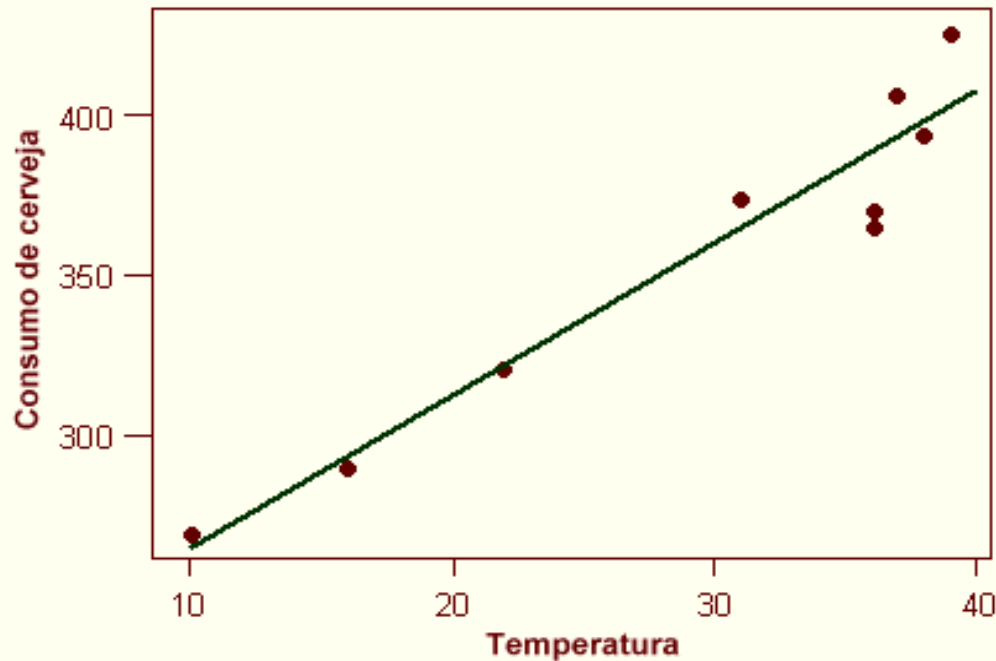
Diagrama de Dispersão



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 25°C?

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