Análise de Regressão

"método estatístico que utiliza relação entre duas ou mais variáveis de modo que uma variável pode ser estimada (ou predita) a partir da outra ou das outras"

Neter, J. et al. Applied Linear Statistical Models. McGraw Hill, 1996

A presença ou ausência de relação linear pode ser investigada sob dois pontos de vista:

a) Quantificando a força dessa relação: correlação.

b) Explicitando a forma dessa relação: regressão.

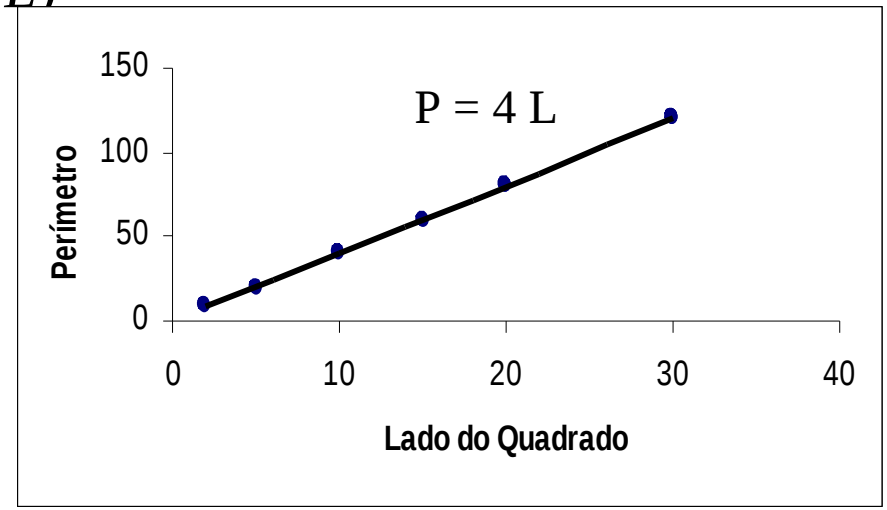
Representação gráfica de duas variáveis quantitativas: Diagrama de dispersão

Relação funcional x Correlação

As variáveis podem possuir dois tipos de relações:

Funcional: a relação é expressa por uma fórmula matemática: $Y = f(X)$

Ex: relação entre o perímetro (P) e o lado de um quadrado (L)

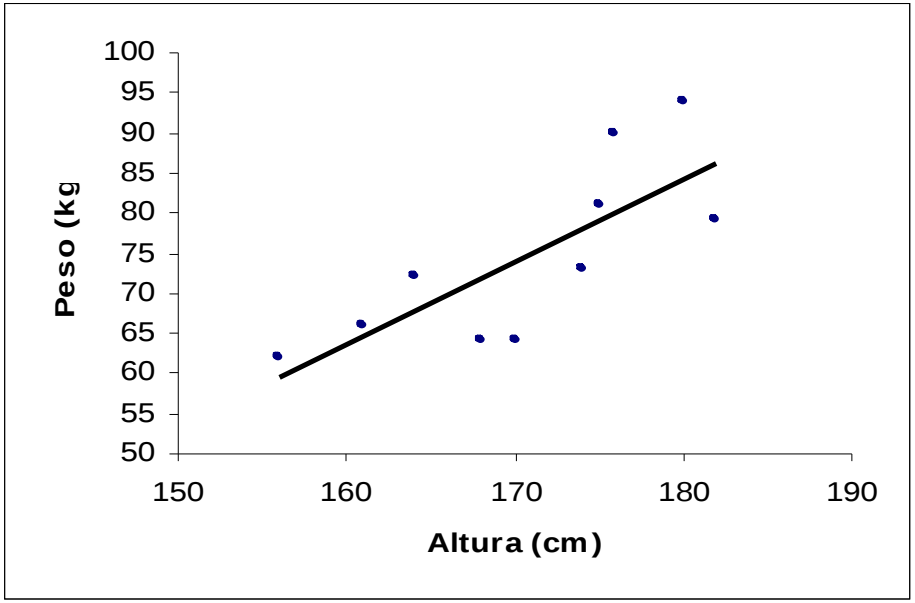


Todos os pontos caem na curva da relação funcional

Correlação: não há uma relação perfeita como no caso da relação funcional.

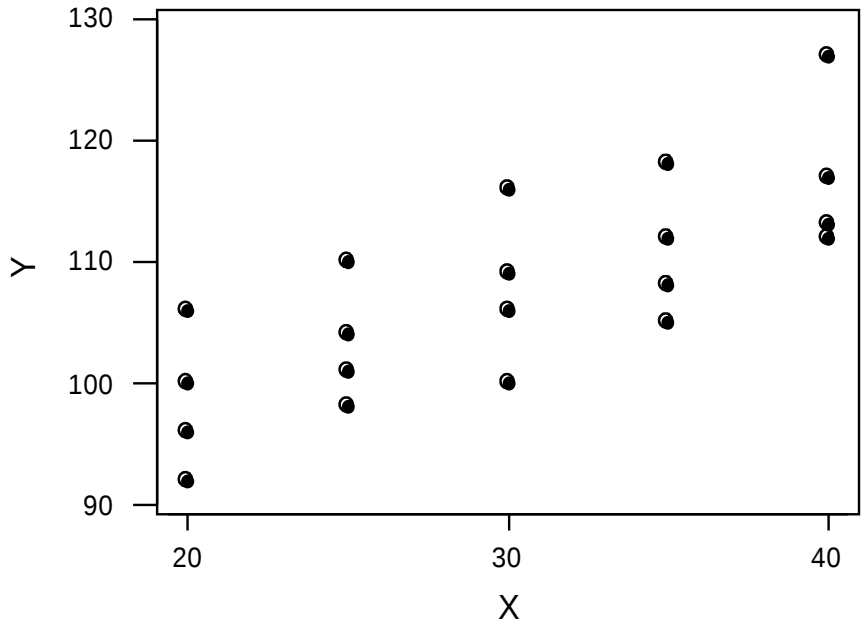
As observações em geral não caem exatamente na curva da relação.

Ex: relação entre o peso (P) e a altura (A) de uma pessoa



A existência de uma relação estatística entre a variável dependente Y e a variável independente X não implica que Y dependa de X, ou que exista uma relação de causa-efeito entre X e Y.

Exemplo 1: Um psicólogo está investigando a relação entre o tempo que um indivíduo leva para reagir a um estímulo visual (Y) com o sexo (W), idade (X) e acuidade visual (Z, medida em porcentagem).



Correlação entre Y e X = 0,768

Análise de regressão: metodologia estatística que estuda (modela) a relação entre duas ou mais variáveis

1. Tempo de reação $\Rightarrow$ variável dependente ou resposta

   idade $\Rightarrow$ variável independente

   $\Downarrow$

   modelo de regressão linear simples

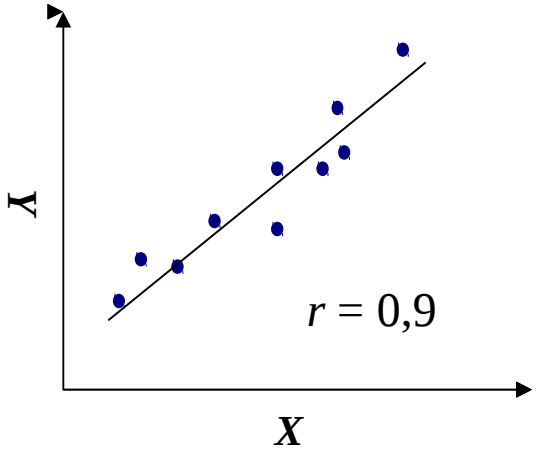
2. Tempo de reação $\Rightarrow$ variável dependente ou resposta

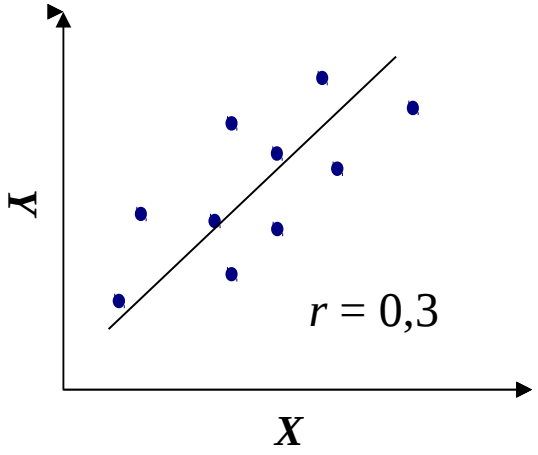
   sexo, idade, acuidade visual $\Rightarrow$ var. independentes

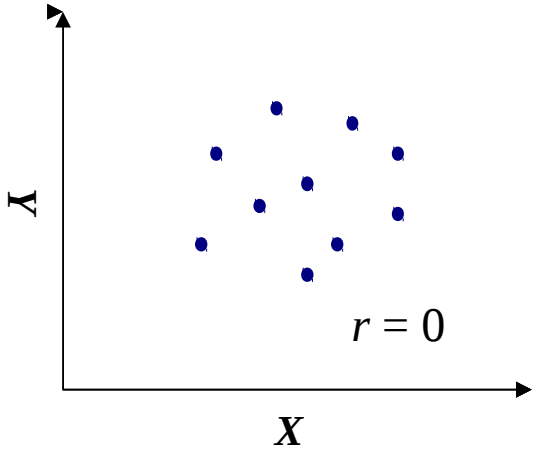
   $\Downarrow$

   modelo de regressão linear múltipla

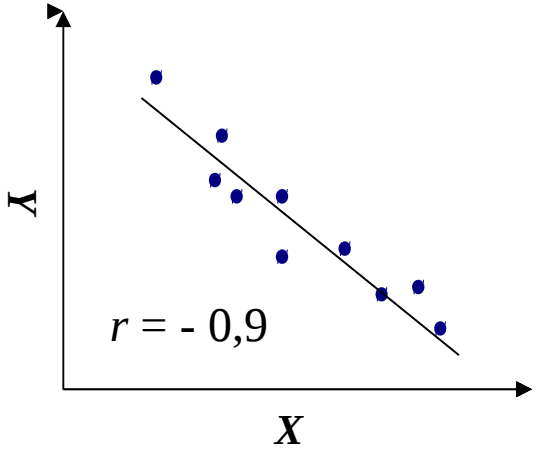
Medida de Associação







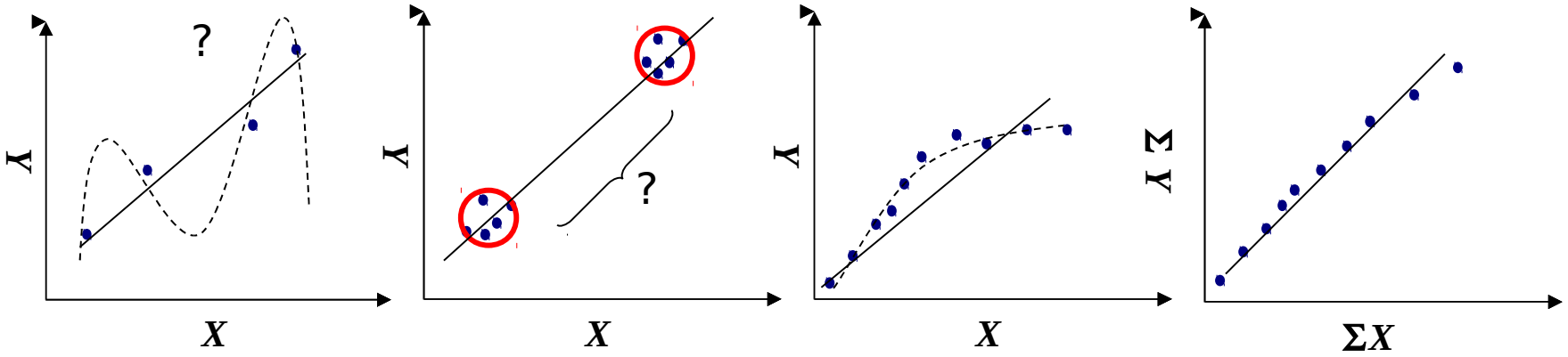
Coeficiente de Correlação (de Pearson)
mede o grau de relação linear entre X e Y



Coeficiente de Correlação

Interpretações errôneas dos coeficientes de correlação

1. Um alto coeficiente de correlação nem sempre indica que a equação de regressão estimada está bem ajustada aos dados.

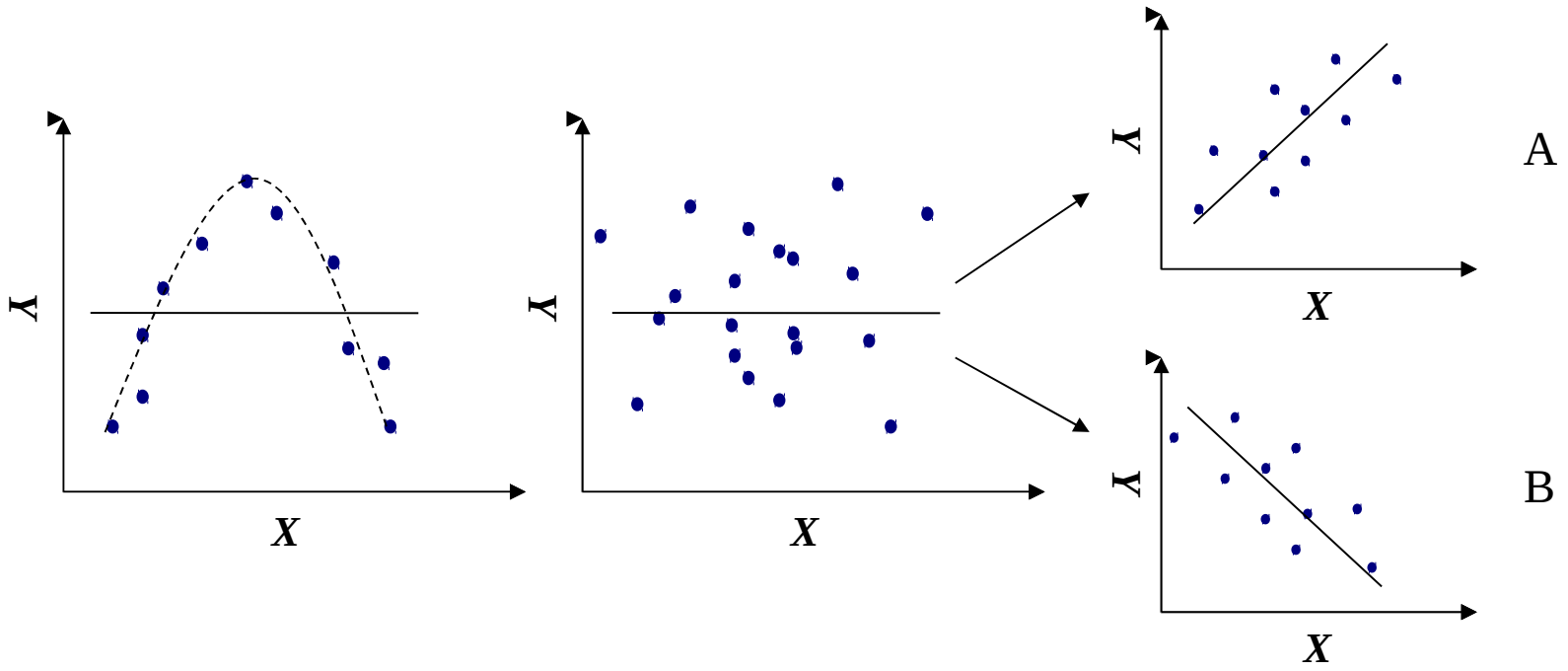


$$Y_i = Y_{i-1} + \Delta y_i \qquad \Delta y_i \neq 0$$
$$X_i = X_{i-1} + \Delta x_i \qquad \Delta x_i \neq 0$$

Coeficiente de Correlação

Interpretações errôneas dos coeficientes de correlação

1. Um coeficiente de correlação próximo de zero nem sempre indica que X e Y não são relacionadas.



Análise de Regressão

1. Determinar como duas ou mais variáveis se relacionam.
2. Estimar a função que determina a relação entre as variáveis.
3. Usar a equação ajustada para prever valores da variável dependente.

Regressão Linear Simples

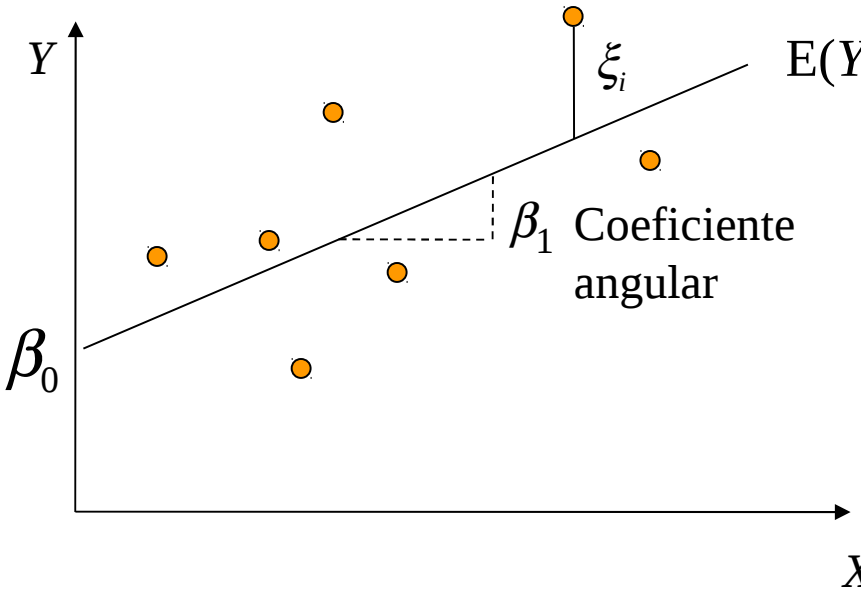
$$Y_i = \beta_0 + \beta_1 X_i + \xi_i$$

$$\mathrm{E}(\xi_i) = 0$$

$$\mathrm{Var}(\xi_i) = \sigma^2$$

$$COV(\xi_i, \xi_j) = 0 \quad \forall i \neq j$$

Modelo de Regressão Linear Simples



$$E(Y) = \beta_0 + \beta_1 X$$

Erro Aleatório

$$Y_i = \beta_0 + \beta_1 X_i + \xi_i$$

Variável Dependente

Variável Independente

Estimação dos parâmetros

Em geral não se conhece os valores de β_0, β_1 e σ^2

Eles podem ser estimados através de dados obtidos por amostras.

O método utilizado na estimação dos parâmetros é o método dos mínimos quadrados, o qual considera os desvios dos Y_i de seu valor esperado:

$$\xi_i = Y_i - (\beta_0 + \beta_1 X_i)$$

Em particular, o método dos mínimos quadrados requer que consideremos a soma dos n desvios quadrados, denotado por Q:

$$Q = \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i]^2$$

Estimação dos parâmetros

De acordo com o método dos mínimos quadrados, os estimadores de β_0 e β_1 são aqueles, denotados por b_0 e b_1, que tornam mínimo o valor de Q.

Derivando

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}[Y_i - \beta_0 - \beta_1 X_i]$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}[Y_i - \beta_0 - \beta_1 X_i]X_i$$

Igualando-se essas equações a zero obtém-se os valores b_0 e b_1 que minimizam Q:

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$b_0 = \overline{Y} - b_1\overline{X}$$

$$E(Y) = \beta_0 + \beta_1 X$$

$$\hat{Y} = b_0 + b_1 X$$

$$e_i = Y_i - \hat{Y}_i \quad \text{(resíduo)}$$

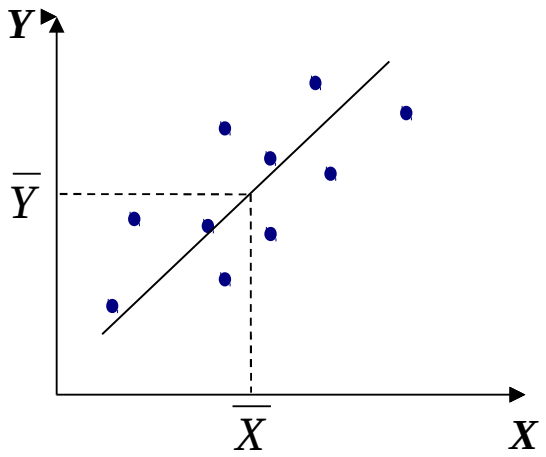
Propriedades da equação de regressão

1) $\sum_{i=1}^{n} e_i = 0$

2) $\sum_{i=1}^{n} e_i^2$ é mínima

3) $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$

4) A reta de regressão passa sempre pelo ponto $(\overline{X}, \overline{Y})$



No exemplo:

n=20, Σy_i= 2150, Σx_i=600, $\Sigma x_i y_i$=65400, Σx_i^2=19000

$$\hat{\beta} = \frac{65400 - 20.30.107,5}{11000 - 20.30^2} = 0,90$$

$$\hat{\alpha} = 107,50 - 0,90.30 = 80,50$$

$$\hat{y}_i = 80,50 + 0,90x_i$$

Interpretação: Para um aumento de 1 ano na idade, o tempo médio de reação aumenta 0,90.

Podemos prever, por exemplo, o tempo médio de reação para pessoas de 20 anos $\Rightarrow$ $\hat{y}(20) = 80,50 + 0,90.20 = 98,50$

$$\hat{y}(25) = 103 \quad \hat{y}(30) = 107{,}50 \quad \hat{y}(35) = 112 \quad \hat{y}(40) = 116{,}5$$

⇓

Vantagem: permite estimar o tempo médio de reação para idades não observadas

⇓

$$\hat{y}(33) = 80{,}50 + 0{,}90.33 = 110{,}20$$

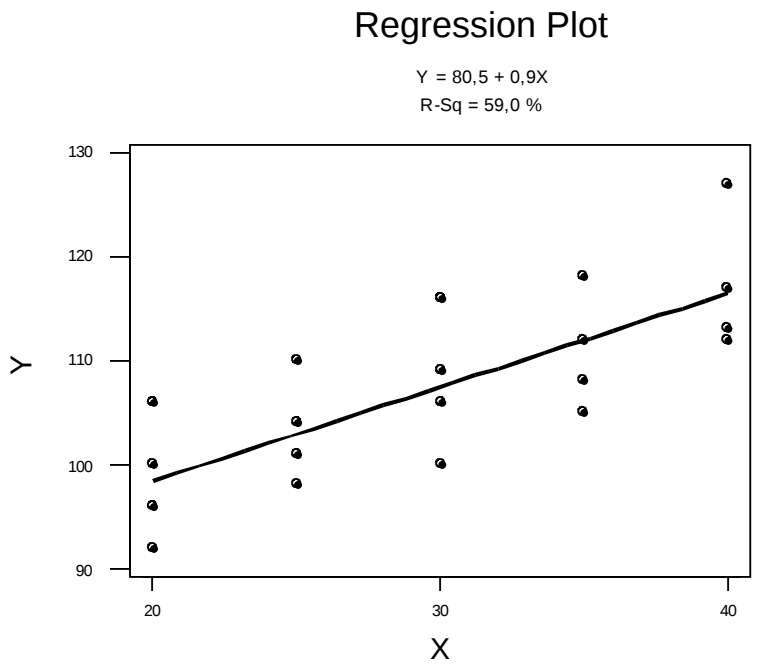
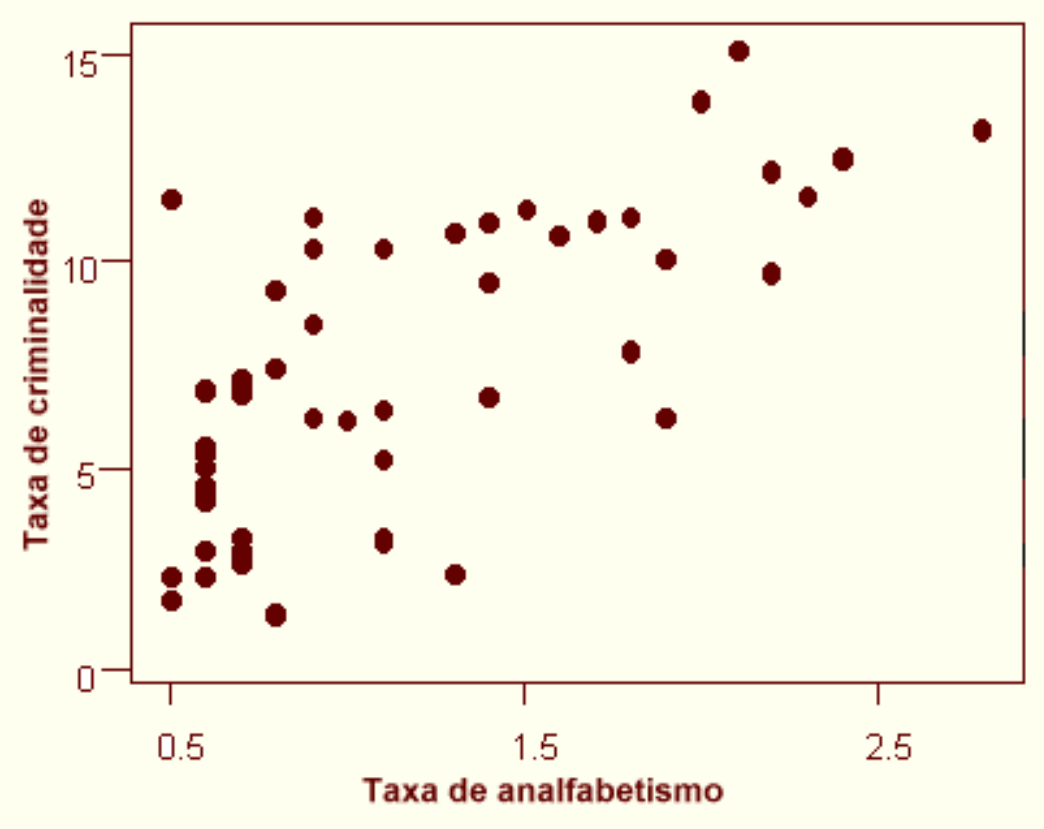


Diagrama de dispersão



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

Correlação entre X e Y: 0,702

a reta ajustada é:

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

^

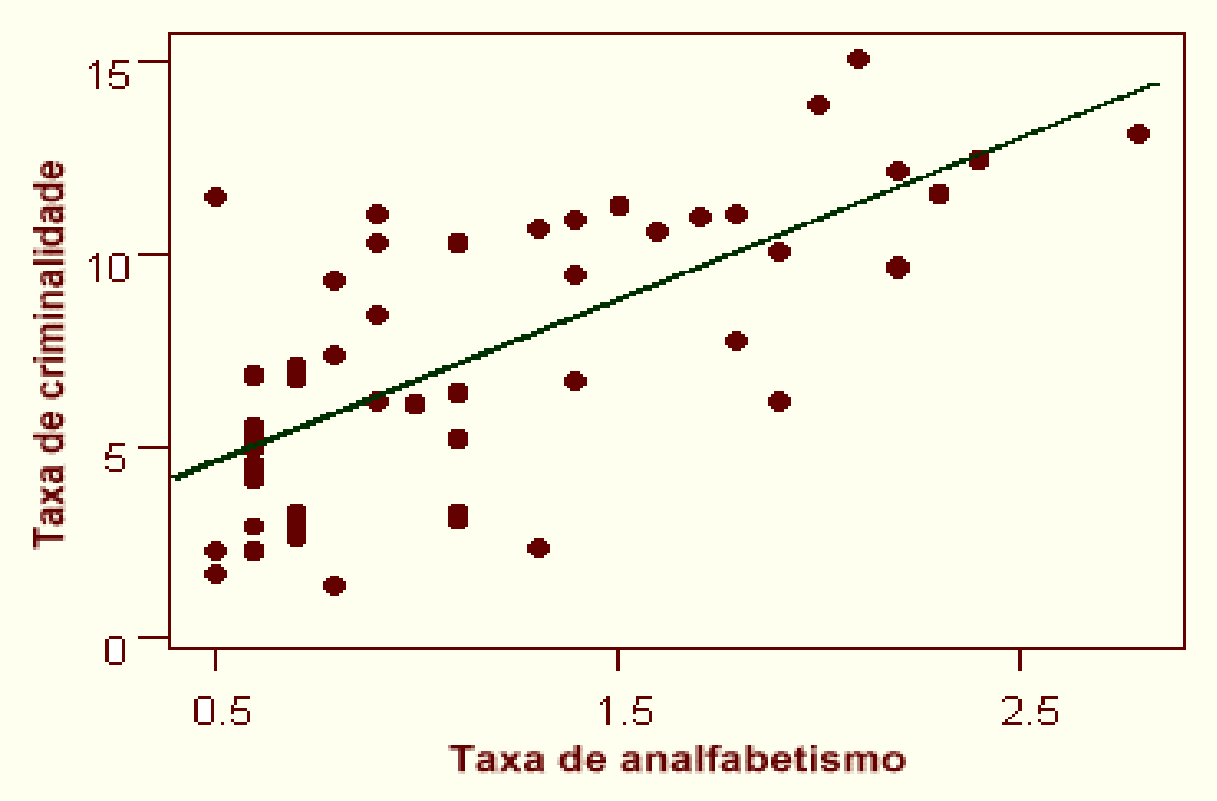
Y : valor predito para a taxa de criminalidade

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, 4,257 unidades.

Graficamente, temos



Como desenhar a reta no gráfico?

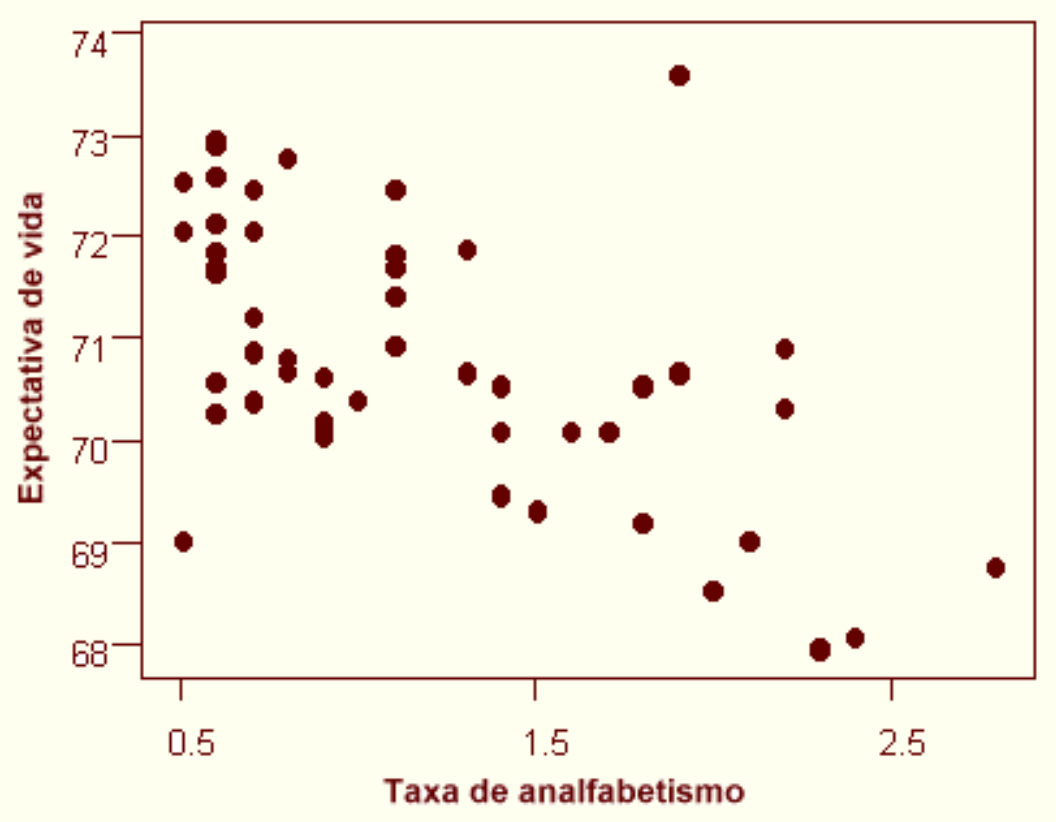
Exemplo 3: expectativa de vida e analfabetismo

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.

Correlação entre X e Y:- 0,59

a reta ajustada é:

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

^

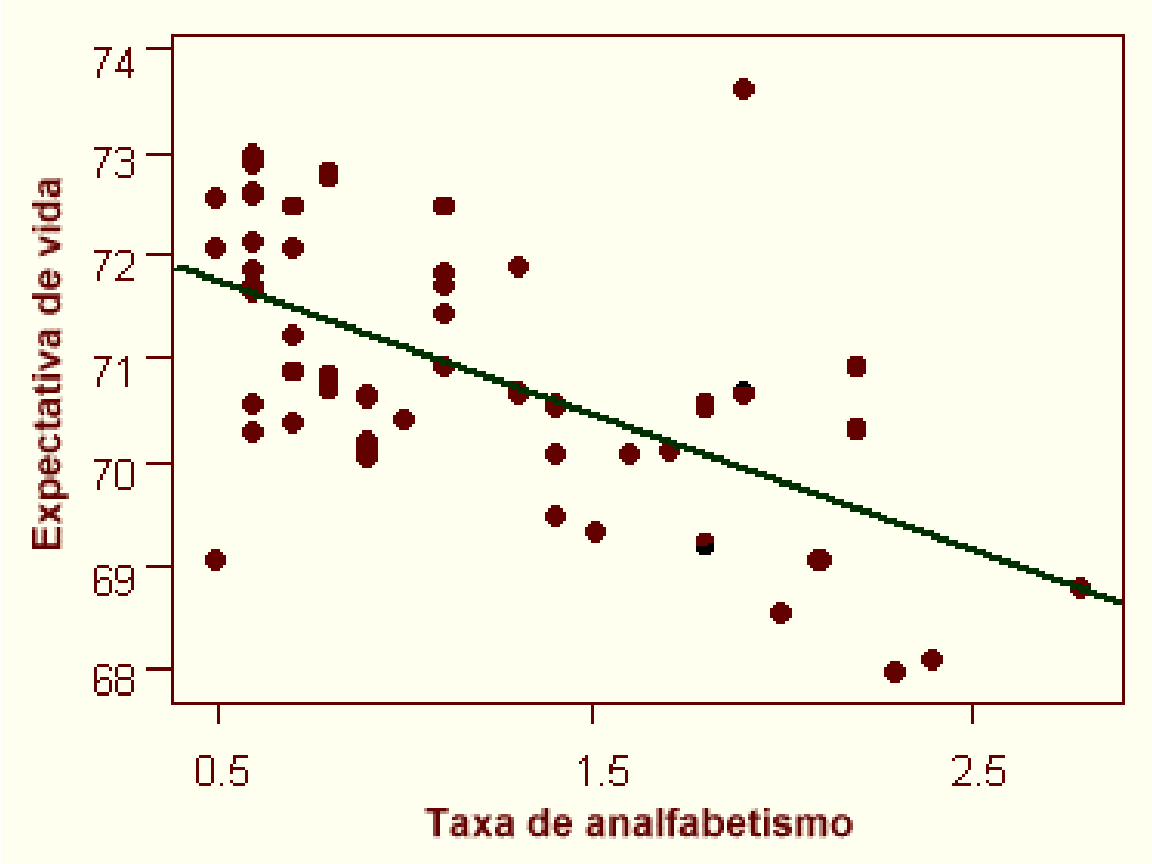
Y : valor predito para a expectativa de vida

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



Resíduos

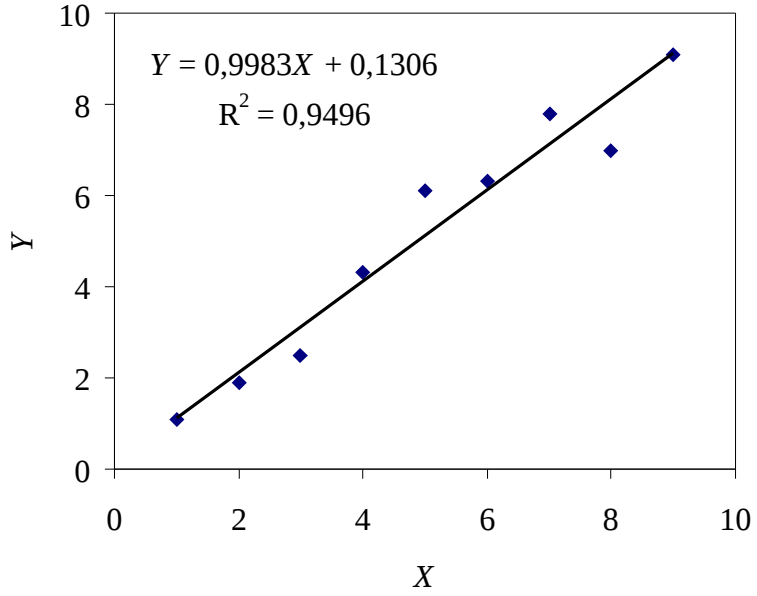
Resíduo é a diferença entre o valor observado e o valor ajustado pela reta, isto é, $Y - \hat{Y}$

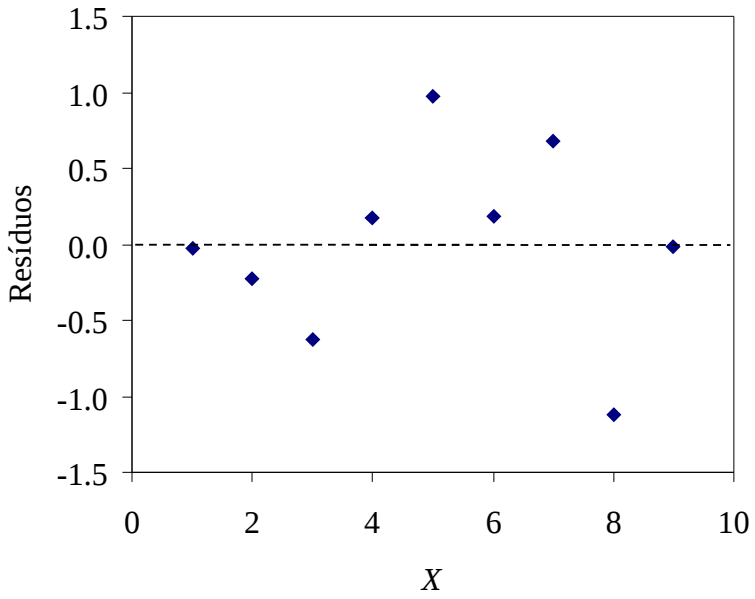
Para verificar a adequação do ajuste deve-se fazer uma *análise dos resíduos*.

Análise de Resíduos

$\hat{Y} = 0{,}9983X + 0{,}1306$

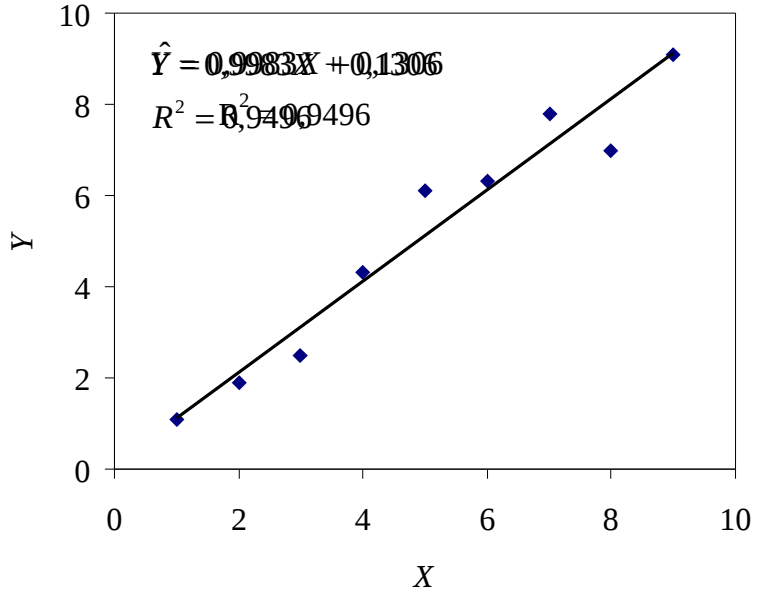
$R^2 = 0{,}9496$

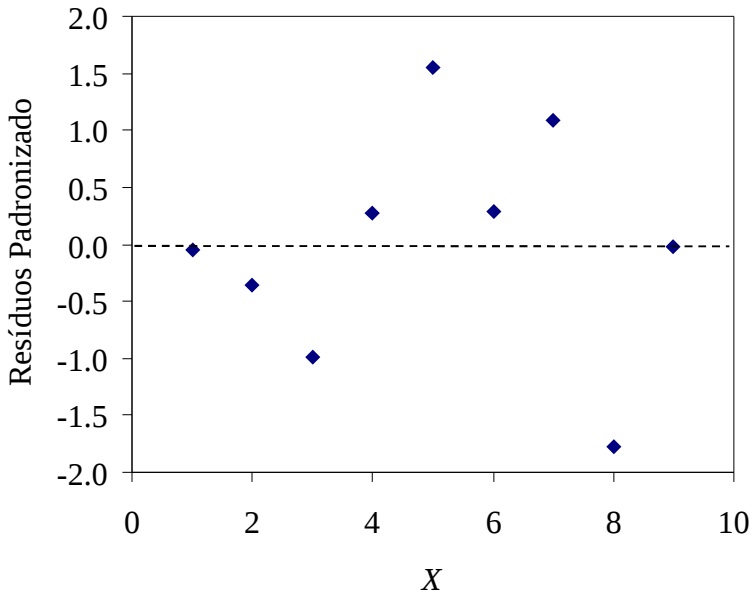




Resíduo = $e_i = Y_i - \hat{Y}_i$

Análise de Resíduos





Resíduo Padronizado = $e_i \Big/ \sqrt{MQRes}$

Análise de Resíduos