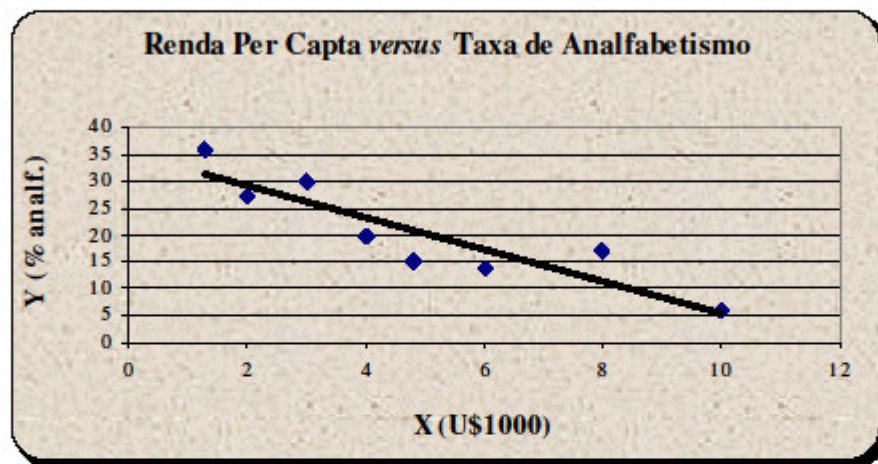

Estatística

Correlação

Correlação Linear - Introdução

Exemplo: amostra de 8 países

$\left\{ \begin{array}{l} X : \text{Renda Per Capita (U\$ 1000)} \\ Y : \text{Taxa de Analfabetismo (\%)} \end{array} \right.$



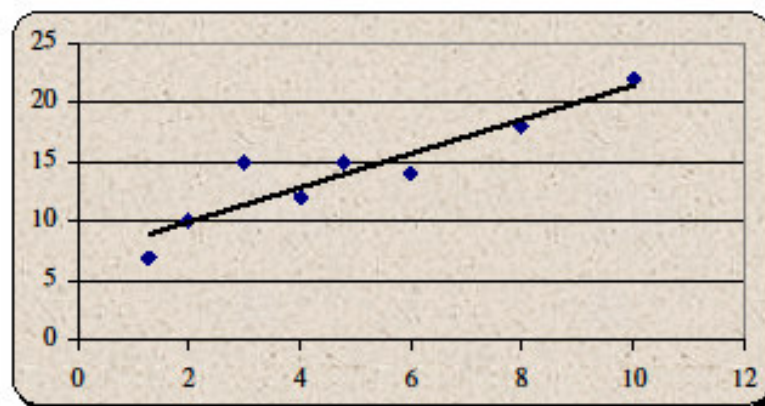
Verificação Visual:

Existe tendência dos maiores valores de X corresponderem aos menores valores de Y, ou seja:

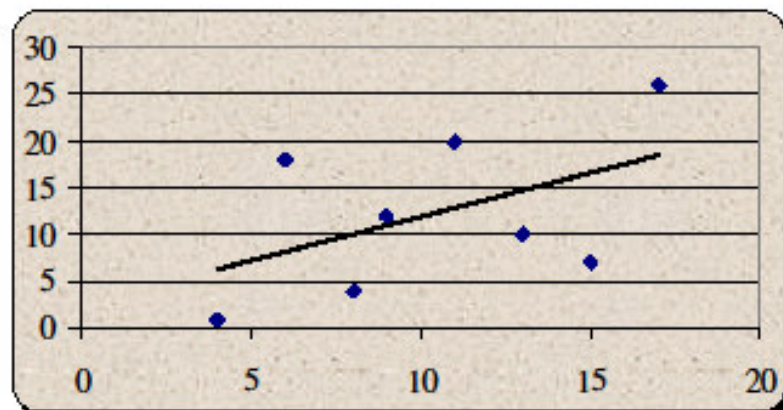
Existe **Correlação Linear Negativa** entre as variáveis

Grau de Correlação Linear

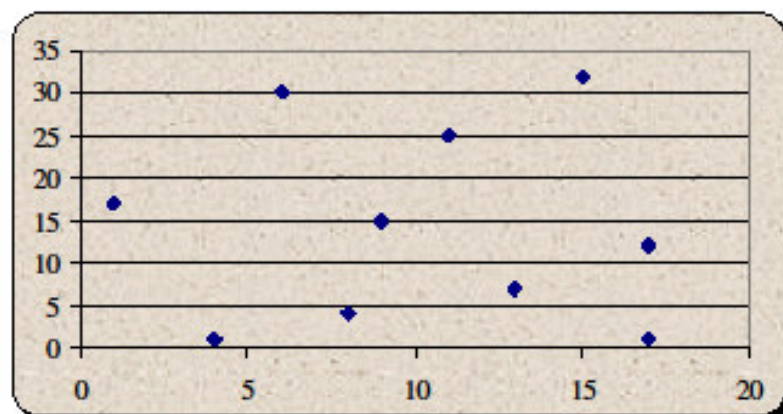
1) Grau Acentuado



2) Grau Pouco Acentuado



3) Grau Nulo



Medida do Grau de Correlação Linear

Covariância: Mede a *variabilidade* considerando duas variáveis

$$S_{xy} = \text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x}) * (y_i - \bar{y})}{n - 1}$$

$$S_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} = \text{Variância}$$

Coeficiente de Correlação Linear de *Pearson*

1. População:

$$\rho = \frac{\text{cov}(x, y)}{\sigma_x * \sigma_y}$$

2. Amostra:

$$r = \frac{S_{xy}}{\sqrt{S_x^2 * S_y^2}}$$

Coeficiente de Correlação Linear de *Pearson*

Sempre: $-1 \leq r \leq 1$

***r próximo de -1 : Correlação Linear
Negativa significativa***

***r próximo de 1 : Correlação Linear
Positiva significativa***

r = 0 : NÃO existe Correlação Linear

Coeficiente de Correlação Linear de *Pearson*

Exemplo: (p. 184, COSTA NETO)

Pessoas	x_i	y_i
	altura	massa
1	174	73
2	161	66
3	170	64
4	180	94
5	182	79
6	164	72
7	156	62
8	168	64
9	176	90
10	175	81
Total	1706	745

$$S_x^2 = 70,4889$$

$$S_y^2 = 126,722$$

$$S_{xy} = 72,888$$

2. Amostra:

$$r = \frac{S_{xy}}{\sqrt{S_y^2 S_x^2}}$$

Conclusão: como r se mostrou relativamente *próximo de 1*, significa que os pontos indicam uma razoavelmente alta *Correlação Linear Positiva*

Testes do coeficiente de Correlação

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho > 0 \end{cases}$$

Verificar se podemos, ao nível de 5% de significância, concluir pela existência de correlação linear positiva entre altura e peso.

CRITÉRIO \rightarrow **REJEITAR H_0** se $r > r_{\text{crítico}}$

onde

$$r_{\text{crítico}} = \sqrt{\frac{1}{1 + \frac{n-2}{(t_{n-2;\alpha})^2}}}$$

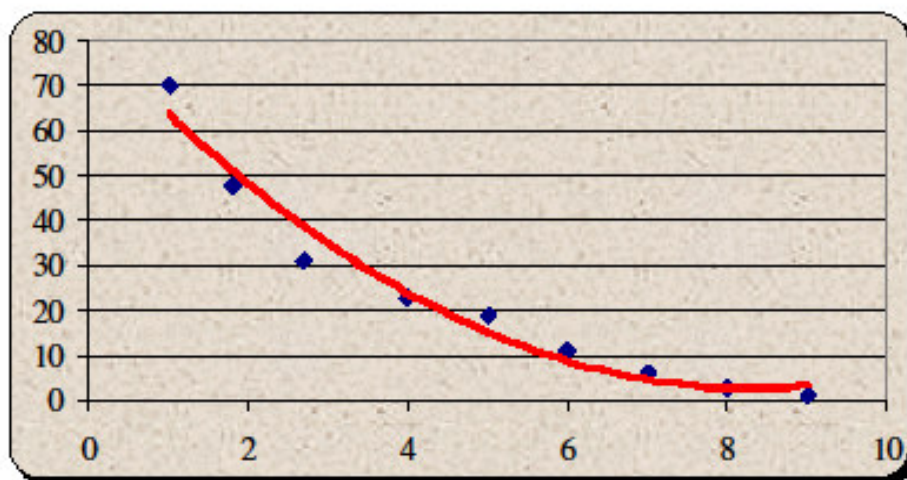
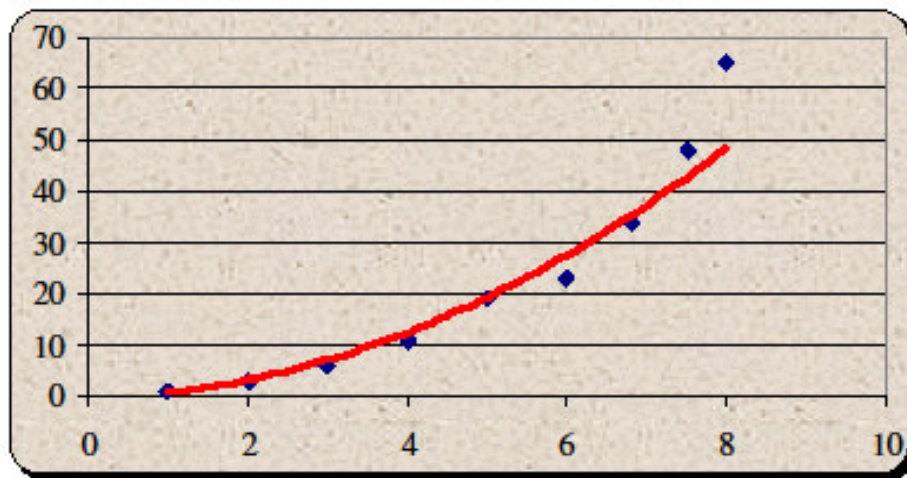
$$\text{para } \alpha = 0,05 \Rightarrow r_{\text{crítico}} = 0,55$$

Logo: $r > r_{\text{crítico}}$ Portanto: REJEITA-SE H_0

Isto é, ao nível de significância de 5%, pode-se dizer que existe Correlação Linear entre altura e peso

Correlação Não - Linear

Variáveis apresentam correlação entre si, mas não explicada por uma função linear

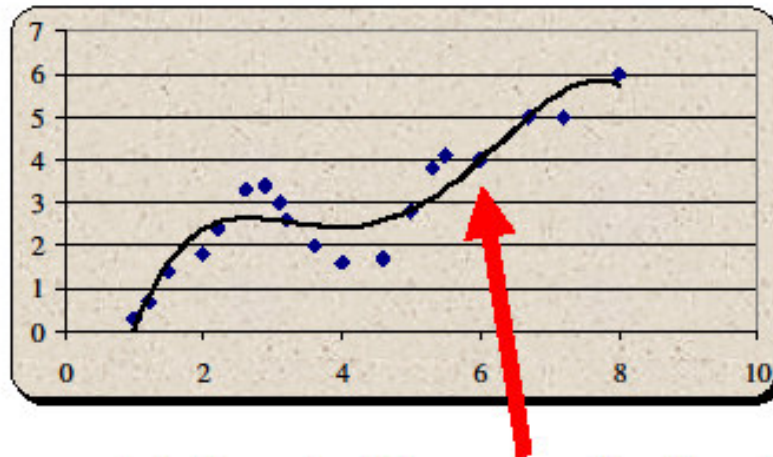


Estatística

Regressão

Regressão - Introdução

Função que exprime a relação funcional entre 2 ou mais variáveis



Linha de Regressão (polinomial)

Variação Residual : Variação em torno da linha de Regressão

Regressão Linear Simples : apenas duas variáveis: **Reta**

Regressão Polinomial : **Polinômio de grau superior a 1**

Regressão Linear Múltipla : mais de duas variáveis: **Reta**

Regressão - Introdução

Conceitos:

X : Variável Não-aleatória (sem erro devido ao acaso)

Y : Variável que tem uma parcela de variação aleatória

Exemplo:

Medir a temperatura de um forno, de 5 em 5 minutos

X : 0, 5, 10, 15, ...

Y_0 : Temperatura no instante inicial (forno é ligado)

Y_1 : Temperatura no instante 5 min

Y_2 : Temperatura no instante 10 min


e assim por diante ...


Conclusões:

Vemos que os valores de X independem de Y , pois foram arbitrados

No entanto, os valores de Y dependem dos de X
(por exemplo, supor que Y aumenta com o aumento de X)

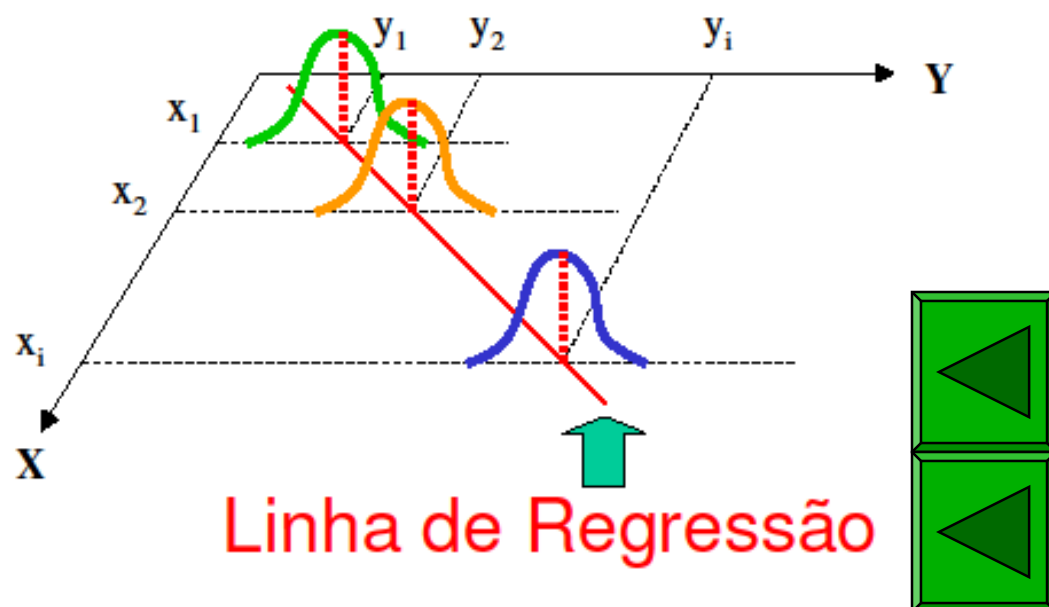
Assim:

X  Variável independente

Y  Variável dependente

Regressão Linear Simples

Hipótese : Variação de Y em torno da linha de regressão tem Distribuição Normal com $\mu = 0$ e $\sigma = \text{constante}$



Função da População : $y = \alpha + \beta \cdot x$

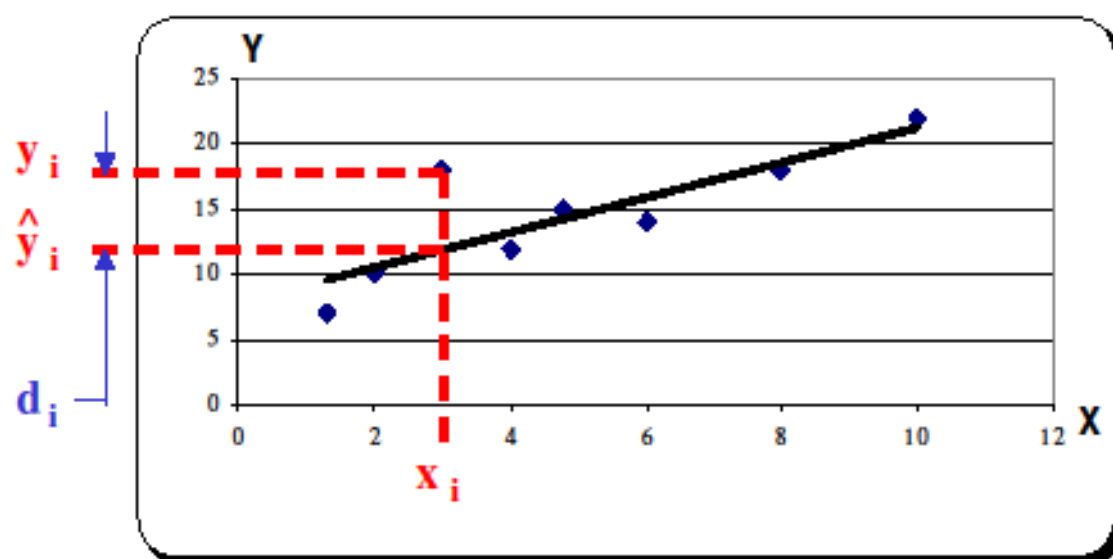
Função da Amostra : $\hat{y} = a + b \cdot x$

Parâmetros **a** e **b** obtidos experimentalmente:

a : estimativa de α

b : estimativa de β (Coeficiente de Regressão Linear)

Regressão Linear Simples- Método dos Mínimos Quadrados



$$d_i = \hat{y}_i - y_i \quad \text{Função da Amostra : } \hat{y} = a + b \cdot x$$

OBJETIVO:

Procurar a reta: $\hat{y} = a + b \cdot x$, para a qual $\sum_{i=1}^n d_i^2$ é mínima !!!

- determinação de d_i^2 mín

$$\min \sum_{i=1}^n d_i^2 = \min \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^n (y_i - a - b \cdot x_i)^2$$

- determinação dos valores de a e b que minimizam

$$\frac{\partial}{\partial a} \sum_{i=1}^n (y_i - a - b \cdot x_i)^2 = 0 \Rightarrow -2 \cdot \sum_{i=1}^n (y_i - a - b \cdot x_i) = 0$$

$$\frac{\partial}{\partial b} \sum_{i=1}^n (y_i - a - b \cdot x_i)^2 = 0 \Rightarrow -2 \cdot \sum_{i=1}^n (y_i - a - b \cdot x_i) \cdot x_i = 0$$

Regressão Linear Simples- Método dos Mínimos Quadrados

- Resolvendo o sistema de equações, temos:

$$b = \frac{S_{xy}}{S_x^2}$$

$$a = \bar{y} - b * \bar{x}$$

Coeficiente de Correlação Linear de *Pearson*

Exemplo: (p. 184, COSTA NETO)

Pessoas	x_i	y_i
	altura	massa
1	174	73
2	161	66
3	170	64
4	180	94
5	182	79
6	164	72
7	156	62
8	168	64
9	176	90
10	175	81
Total	1706	745

$$S_x^2 = 70,4889$$

$$S_y^2 = 126,722$$

$$S_{xy} = 72,888$$

- Determinar a reta: $y = a + b \cdot x$

$$b = \frac{S_{xy}}{S_x^2} = \frac{72,888}{70,4889} = 1,03$$

$$a = \bar{y} - b \cdot \bar{x} = 74,5 - 1,03 \cdot 170,6 = -101,22$$

$$y = -101,22 + 1,03 x$$

Estimação e Testes de Hipóteses sobre Parâmetros da Reta

Reta Teórica: $y = \alpha + \beta \cdot x$

Reta Estimativa: $\hat{y} = a + b \cdot x$

Parâmetros **a** e **b** obtidos experimentalmente

Teste de Hipóteses:

$$\begin{cases} H_0 : \beta = 0 \text{ (NÃO existe Regressão)} \\ H_1 : \beta \neq 0 \text{ (Existe Regressão)} \end{cases}$$

Seja: s_R^2 = Variância Residual, ou Variância em torno da reta dos Mínimos Quadrados

$$s_R^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}$$

onde: $\hat{y}_i = a + b \cdot x_i$

$n - 2$ = graus de liberdade (para estimar \hat{y}_i , é necessário estimar os dois parâmetros: a e b)

$$S_R^2 = \left(\frac{n-1}{n-2} \right) \frac{S_x^2 S_y^2 - S_{xy} S_{xy}}{S_x^2}$$

• Teste de Hipótese:

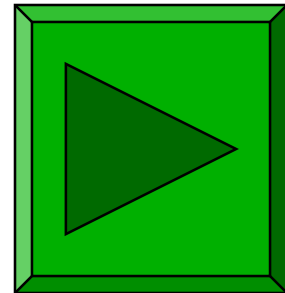
$$\begin{cases} H_0 : \beta = \beta_0 \\ H_1 : \beta \neq \beta_0 \end{cases}$$

$$\text{Reta T: } \mathbf{y = \alpha + \beta.x}$$

$$\text{Reta E: } \mathbf{\hat{y} = a + b.x}$$

$$\boxed{\mathbf{b}} \xrightarrow{\text{Distrib. Normal}} \begin{cases} \mu(b) = \beta \\ \sigma^2(b) \text{ desconhecido} \end{cases}$$

$$\boxed{S^2(b) = \frac{S_R^2}{(n-1)S_x^2}}$$



Critério: Rejeitar H_0 se $|b_{\text{calculado}}| > b_{\text{crítico}}$

$$b_{\text{crítico}} = t_{n-2, \alpha/2} \cdot S(b) + \beta_0$$

Estimação e Testes de Hipóteses sobre Parâmetros da Reta

$$\begin{cases} H_0 : \beta = 0 & (\text{NÃO existe Reta de Regressão}) \\ H_1 : \beta \neq 0 \end{cases}$$

Critério: Rejeitar H_0 se $|b_{\text{calculado}}| > b_{\text{crítico}}$

$$b_{\text{crítico}} = t_{n-2, \alpha/2} \cdot S(b)$$

Intervalo de Confiança para β

		<u>Estimador</u> →
Distribuição. Normal	{	$\mu(b) = \beta \xrightarrow{\text{red arrow}} b$
		$\sigma^2(b) \xrightarrow{\text{red arrow}} S^2(b) = \frac{S_R^2}{(n-1)S_x^2}$

$$\Pr\left[b - t_{n-2, \alpha/2} * S(b) \leq \beta \leq b + t_{n-2, \alpha/2} * S(b)\right] = 1 - \alpha$$

Estimação e Testes de Hipóteses sobre Parâmetros da Reta

Exemplo: Altura X Massa (n = 10 pessoas)

Resultados já obtidos: $a = -101,22$ $b = 1,03$

$$S_x^2 = 70,4889$$

$$S_y^2 = 126,722$$

$$S_{xy} = 72,888$$

$$S_R^2 = \left(\frac{n-1}{n-2} \right) \frac{S_x^2 S_y^2 - S_{xy}^2}{S_x^2}$$

$$S^2(b) = \frac{S_R^2}{(n-1)S_x^2}$$

$$S_R^2 = \left(\frac{9}{8} \right) \frac{70,4889 \cdot 126,722 - (72,888)^2}{70,4889} = 57,8$$

$$S(b) = \sqrt{\frac{57,8}{(9) \cdot 70,4889}} = 0,302$$

• Teste de Hipótese:

$$\begin{cases} H_0 : \beta = 0 & (\text{NÃO existe Reta de Regressão}) \\ H_1 : \beta \neq 0 \end{cases} \quad \alpha = 5\%$$

Critério: Rejeitar H_0 se $|b_{\text{calculado}}| > b_{\text{crítico}}$

$$b_{\text{crítico}} = t_{n-2, \alpha/2} \cdot S(b) + \beta_0 = 2,306 \cdot 0,302 = 0,696$$

Logo: $|b| > b_{\text{crítico}}$ ($1,03 > 0,696$) **REJEITA-SE H_0**

Isto é, ao nível de significância de 5%, pode-se dizer que existe Reta de Regressão

Estimação e Testes de Hipóteses sobre Parâmetros da Reta

Exemplo: Altura X Massa ($n = 10$ pessoas)

Resultados já obtidos: $a = -101,22$ $b = 1,03$

$$S_x^2 = 70,4889$$

$$S_y^2 = 126,722$$

$$S_{xy} = 72,888$$

$$S_R^2 = \left(\frac{n-1}{n-2} \right) \frac{S_x^2 S_y^2 - S_{xy}^2}{S_x^2}$$

$$S^2(b) = \frac{S_R^2}{(n-1)S_x^2}$$

$$S_R^2 = \left(\frac{9}{8} \right) \frac{70,4889 \cdot 126,722 - (72,888)^2}{70,4889} = 57,8$$

$$S(b) = \sqrt{\frac{57,8}{(9) \cdot 70,4889}} = 0,302$$

• Intervalo de Confiança para β

$$\Pr \left[b - t_{n-2, \alpha/2} * S(b) \leq \beta \leq b + t_{n-2, \alpha/2} * S(b) \right] = 1 - \alpha$$

$$P[1,03 - 2,306 * 0,302 \leq \beta \leq 1,03 + 2,306 * 0,302] = 95\%$$

$$P(0,33 \leq \beta \leq 1,73) = 95\%$$

Conclusão: Isto é, podemos afirmar, com uma probabilidade de 95%, que β se encontra no intervalo entre 0,33 e 1,73

Estimação e Testes de Hipóteses sobre Parâmetros da Reta

• Teste de Hipótese:

$$\left\{ \begin{array}{ll} H_0 : \alpha = \alpha_0 & \text{Reta T: } y = \alpha + \beta \cdot x \\ H_1 : \alpha \neq \alpha_0 & \text{Reta E: } \hat{y} = a + b \cdot x \end{array} \right.$$

$$\begin{array}{l} \text{Distr.} \\ \text{Normal} \end{array} \quad \left\{ \begin{array}{ll} \mu(a) = \alpha & \longrightarrow a \\ \sigma^2(a) ? & \longrightarrow S^2(a) \end{array} \right. \quad \begin{array}{c} \text{Estimador} \end{array}$$

$$S^2(a) = \frac{s_R^2 \cdot \sum_{i=1}^n x_i^2}{n(n-1)S_x^2}$$

Critério: Rejeitar H_0 se $|a_{\text{calculado}}| > a_{\text{crítico}}$

$$a_{\text{crítico}} = t_{n-2, \alpha/2} \cdot S(a) + \alpha_0$$

Intervalo de Confiança para α

$$P\left[a - t_{n-2, \alpha/2} \cdot s(a) \leq \alpha \leq a + t_{n-2, \alpha/2} \cdot s(a)\right] = 1 - \alpha$$

$$a = \bar{y} - b * \bar{x}$$

$$S^2(a) = \frac{s_R^2 \cdot \sum_{i=1}^n x_i^2}{n(n-1)S_x^2}$$

Coeficiente de Correlação Linear de *Pearson*

2. Amostra:

$$r = \frac{S_{xy}}{\sqrt{S_x^2 * S_y^2}}$$

Coeficiente de determinação

$$r^2 = \frac{S_{xy} S_{xy}}{S_x^2 * S_y^2}$$

Indica a porcentagem de variação de Y explicada pela reta de regressão.

Indica quanto a reta de regressão fica bem determinada em função da correlação entre os pontos experimentais.

Região de previsão para y'

Determinar um intervalo no qual, com $1 - \alpha$ de certeza, possamos prever que o valor experimental de Y , obtido para dado x' , venha a estar contido.

$$\hat{y}' \pm t_{n-2, \alpha/2} \cdot s_R \sqrt{1 + \frac{1}{n} + \frac{(x' - \bar{x})^2}{(n-1)S_x^2}}$$

onde

$$s_R^2 = \left(\frac{n-1}{n-2} \right) \frac{S_x^2 S_y^2 - S_{xy}^2}{S_x^2}$$

Região de previsão para y'

Exemplo: (p. 195, COSTA NETO)

x_i	y_i
1	0,5
2	0,6
3	0,9
4	0,8
5	1,2
6	1,5
7	1,7
8	2
36	9,2

$$\bar{x} = \frac{36}{8} = 4,5$$

$$\bar{y} = \frac{9,2}{8} = 1,15$$

$$S_{XY} = 1,3 \quad S_x^2 = 6 \quad S_y^2 = 0,294$$

$$b = \frac{S_{xy}}{S_x^2} = \frac{1,3}{6} = 0,217$$

$$a = \bar{y} - b \cdot \bar{x} = 1,15 - 0,217 \cdot 4,5 = 0,174$$

$$y = 0,174 + 0,217x$$

Região de previsão para y'

$$y = 0,174 + 0,217 x$$

		Região de Previsão	
x'	y'	Limite inferior	Limite superior
1	0,391	0,044	0,738
2	0,608	0,279	0,937
3	0,825	0,509	1,141
4	1,042	0,732	1,352
5	1,259	0,949	1,569
6	1,476	1,160	1,792
7	1,693	1,364	2,022
8	1,91	1,563 *	2,257 **

$$\hat{y}' = 0,174 + 0,217 x (8) = 1,91$$

$$S_R^2 = \left(\frac{n-1}{n-2} \right) \frac{S_x^2 S_y^2 - S_{xy} S_{xy}}{S_x^2}$$

$$S_R = \sqrt{\left(\frac{7}{6} \right) \frac{6 \cdot 0,294 - (1,3)^2}{6}} = 0,119$$

$$t_{n-2, \alpha / 2} = t_{6, 2,5\%} = 2,447$$

$$\hat{y}' \pm t_{n-2, \alpha / 2} \cdot S_R \sqrt{1 + \frac{1}{n} + \frac{(x' - \bar{x})^2}{(n-1) S_x^2}}$$

* Limite inferior: $1,91 - 2,447 \cdot 0,119 \sqrt{1 + \frac{1}{8} + \frac{(8 - 4,5)^2}{7 \cdot 6}} = 1,563$

** Limite superior: $1,91 + 2,447 \cdot 0,119 \sqrt{1 + \frac{1}{8} + \frac{(8 - 4,5)^2}{7 \cdot 6}} = 2,257$

Região de previsão para y'

