Parte 10

Correlação e Regressão

Prof. Dr. Renato de Oliveira Moraes

Variância e Covariância

$$s_x^2 = s_{xx} = \frac{\sum (x_i - \bar{x})^2}{N-1} \longrightarrow s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}}$$

$$s_y^2 = s_{yy} = \frac{\sum (y_i - \bar{y})^2}{N-1} \longrightarrow s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{N-1}}$$

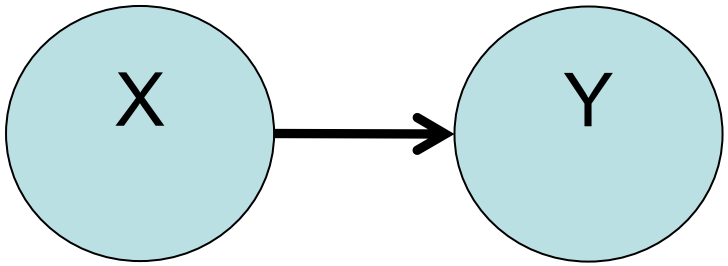
$$s_{xy} = \text{cov}(x.y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

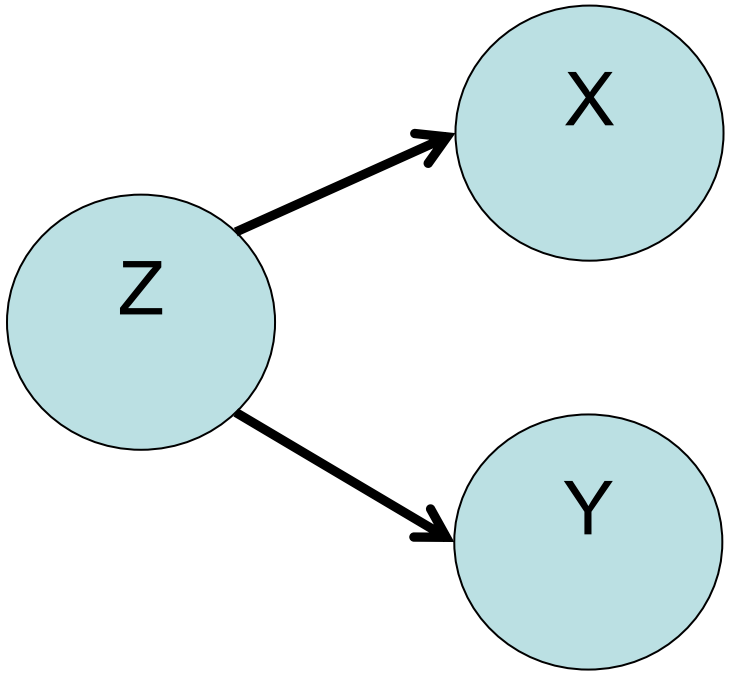
Correlação

Coeficiente de Correlação de Pearson

$$R = \frac{\text{cov}(x,y)}{s_x s_y} = \frac{s_{xy}}{s_x s_y} = \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}$$

Relação Causal e Correlação Linear entre “X” e “Y”





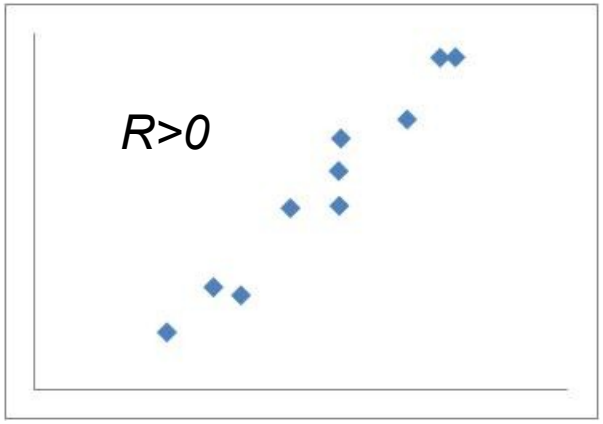
Indivíduo	Peso (kg)	Altura (m)
1	75	1,65
2	94	1,85
3	75	1,75
4	72	1,71
5	65	1,69
6	85	1,75

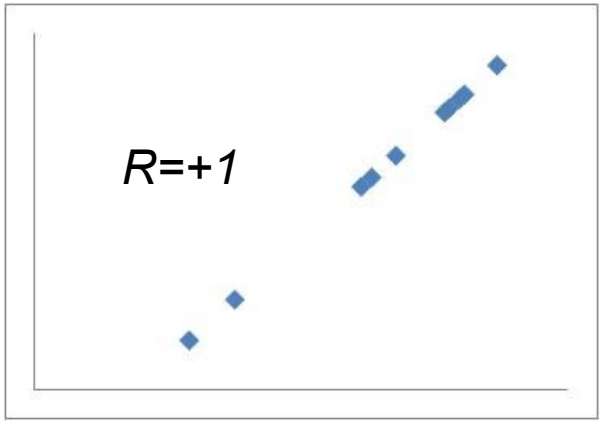
S_{Peso}	10,27
S_{altura}	0,07
Cov(Peso;Altura)	0,58
Coef de Correlação (R)	0,82

Indivíduo	Peso (kg)	Altura (cm)
1	75	165
2	94	185
3	75	175
4	72	171
5	65	169
6	85	175

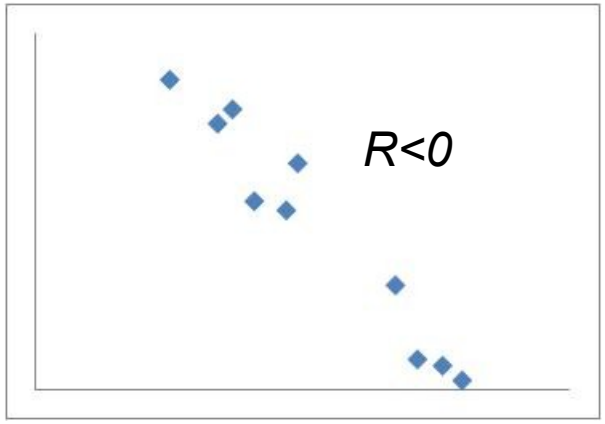
S_{Peso}	10,27
S_{altura}	6,86
Cov(Peso;Altura)	57,73
Coef de Correlação (R)	0,82

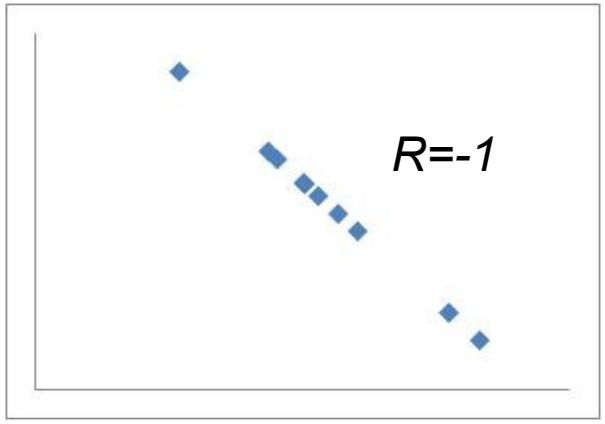
Correlação Positiva



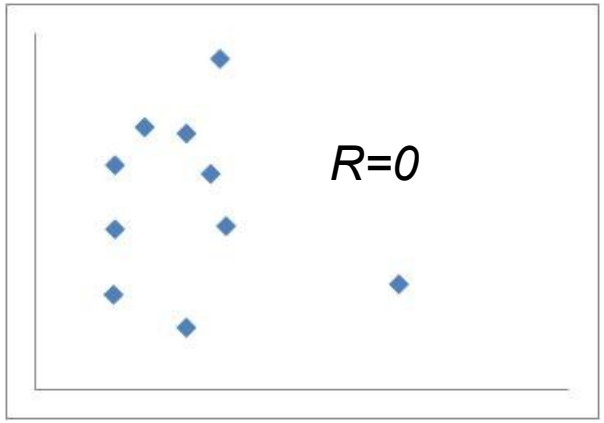


Correlação Negativa





Correlação Nula



Exemplo

Calcule a correlação entre as variáveis abaixo

X	Y
5	5
6	11
12	14
13	21
11	21

Exemplo

X_i	Y_i	$X_i - X_{médio}$	$Y_i - Y_{médio}$	$(X_i - X_{médio})^2$	$(Y_i - Y_{médio})^2$	$(X_i - X_{médio})(Y_i - Y_{médio})$
5	5	-4,4	-9,4	19,36	88,36	41,36
6	11	-3,4	-3,4	11,56	11,56	11,56
12	14	2,6	-0,4	6,76	0,16	-1,04
13	21	3,6	6,6	12,96	43,56	23,76
11	21	1,6	6,6	2,56	43,56	10,56
47	**72**			**53,2**	**187,2**	**86,2**
9,4	14,4					

$$s_x^2 = \frac{\sum (x_i - \bar{x})^2}{N-1} = \frac{53,2}{4} = 13,3 \longrightarrow s_x \cong 3,65$$

$$s_y^2 = s_{yy} = \frac{\sum (y_i - \bar{y})^2}{N-1} = \frac{187,2}{4} = 46,8 \longrightarrow s_y \cong 6,84$$

$$s_{xy} = \mathrm{cov}(x.y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1} = \frac{86,2}{4} = 21,55$$

$$R = \frac{\mathrm{cov}(x.y)}{s_x s_y} = \frac{21,55}{3,65 \times 6,84} \cong 0,864 \longrightarrow R^2 = 0,746 = 74,6\%$$

Inferência sobre o Coeficiente de Correlação de Pearson

H_0: $\rho = 0$

H_1: $\rho \neq 0$

$$t_{N-2;\alpha/2} = R\sqrt{\frac{N-2}{1-R^2}}$$

Exemplo

H_0: $\rho = 0$

H_1: $\rho \neq 0$

$$t_{N-2;\alpha/2} = R\sqrt{\frac{N-2}{1-R^2}}$$

$$t_{Calculado} = 0{,}864\sqrt{\frac{5-2}{1-0{,}746}} \cong 2{,}969$$

$$t_{Crítico} = t_{3;2{,}5\%} = 3{,}182$$

$$Sig.(\alpha) = 5{,}91\%$$

→ $t_{Crítico}$ =

,84

eria possível
a Hipótese Nula
!!

Exemplo

Se α *= 1%* → $t_{Crítico}$ *= 5,84*

Não seria possível rejeitar a Hipótese Nula !!

Cuidado com as correlações espúrias!!

Veja este link: http://www.tylervigen.com/spurious-correlations

Regressão Linear

Prof. Dr. Renato de Oliveira Moraes

Coeficientes da Regressão

$$f(x) = a + bx$$

$$b = \frac{n\sum(x_i y_i) - \left(\sum x_i\right)\left(\sum y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$a = \frac{\sum y_i - b\sum x_i}{n}$$

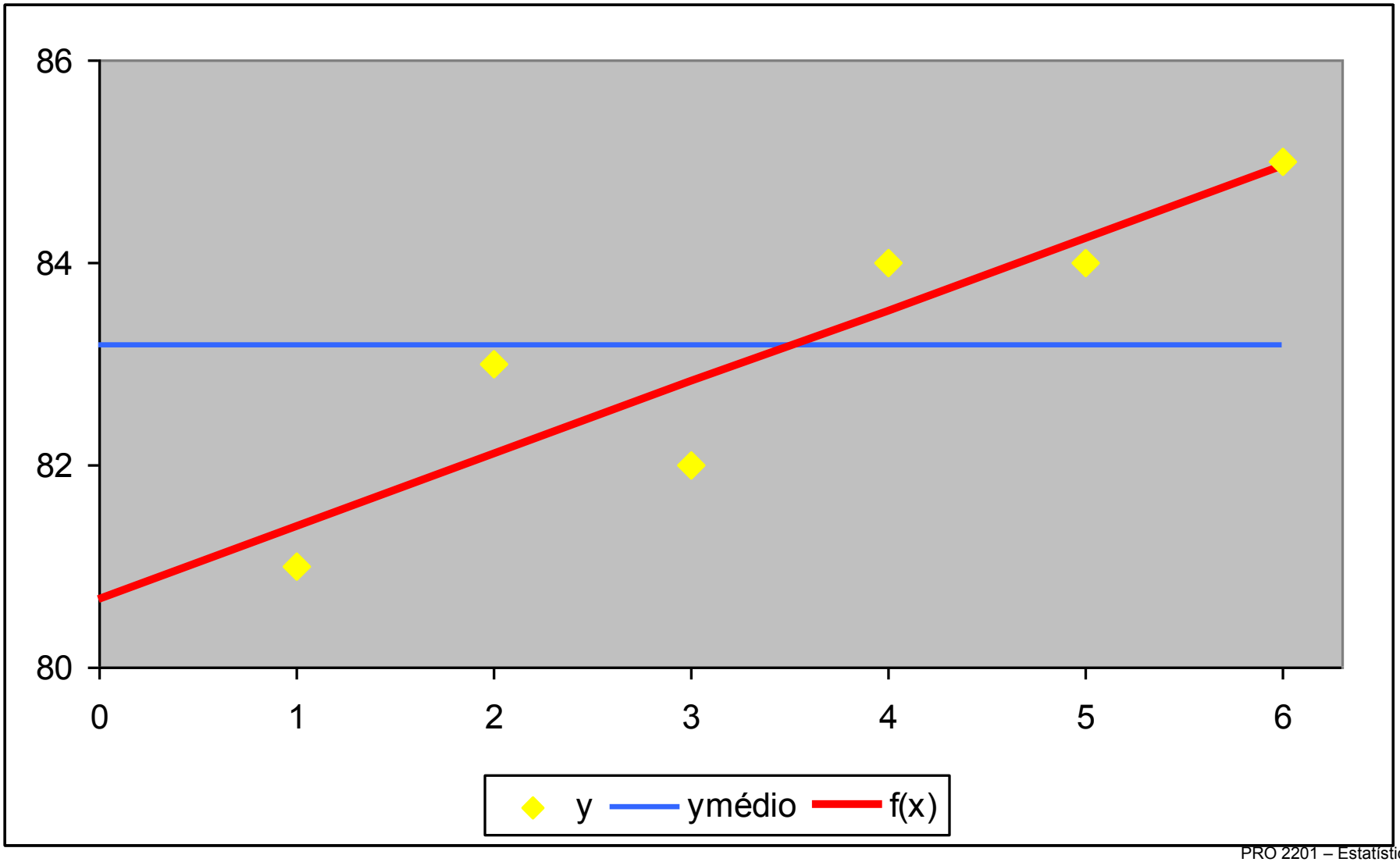
Exemplo

x	y	x^2	xy
1	81	1	81
2	83	4	166
3	82	9	246
4	84	16	336
5	84	25	420
6	85	36	510
21	**499**	**91**	**1759**

$$b = \frac{n\sum(x_i y_i) - \left(\sum x_i\right)\left(\sum y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{6\times 1759 - 21\times 499}{6\times 91 - (21)^2} \cong 0{,}714$$

$$a = \frac{\sum y_i - b\sum x_i}{n} = \frac{499 - 0{,}714\times 21}{6} \cong 80{,}667$$

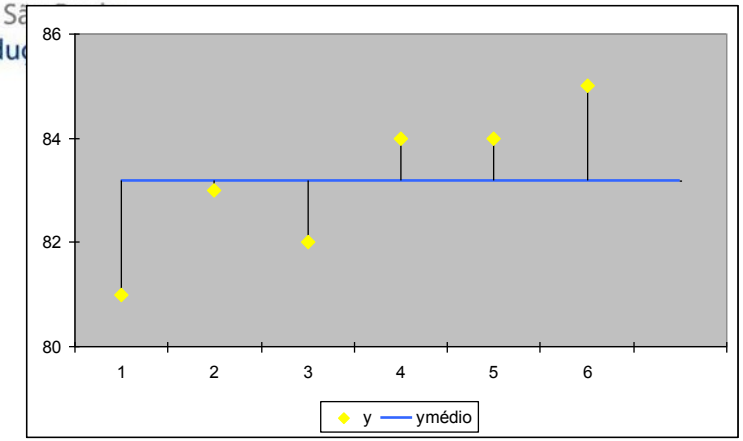
$$f(x) = 80{,}667 + 0{,}714x$$


86
84
82
80
0
1
2
3
4
5
6
y
ymédio
f(x)

$y_i - \bar{y}$

Variação total:

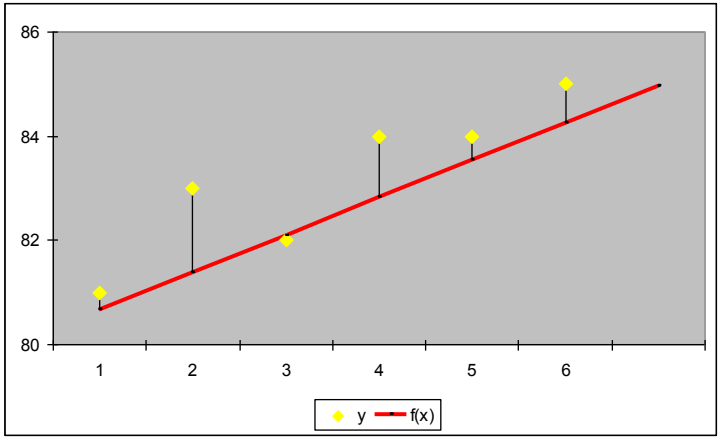
$$\sum (y_i - \bar{y})^2$$



$$y_i - \hat{y} = y_i - f(x)$$

Variação não explicada :

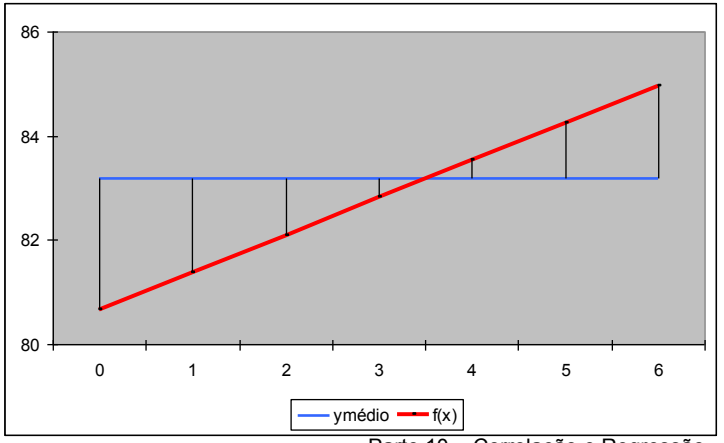
$$\sum (y_i - f(x))^2$$



$$\hat{y} - \bar{y} = f(x) - \bar{y}$$

Variação explicada pela regressão :

$$\sum (f(x) - \bar{y})^2$$



$$\text{Variação total} \quad \sum (y_i - \bar{y})^2 = \text{Variação não explicada} \quad \sum (y_i - f(x))^2 + \text{Variação explicada pela regressão} \quad \sum (f(x) - \bar{y})^2$$

Coeficiente de Determinação

$$R^2 = \frac{\sum (f(x) - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{\text{Variação explicada}}{\text{Variação total}}$$

PRO

Exemplo

x_i	y_i	x^2	xy	$y_{médio}$	$f(x)$	$(y_i-y_{médio})^2$	$(y_i-f(x))^2$	$(f(x)-y_{médio})^2$
0				83,17	80,67			
1	81	1	81	83,17	81,38	4,69	0,15	3,19
2	83	4	166	83,17	82,10	0,03	0,82	1,15
3	82	9	246	83,17	82,81	1,36	0,66	0,13
4	84	16	336	83,17	83,52	0,69	0,23	0,13
5	84	25	420	83,17	84,24	0,69	0,06	1,15
6	85	36	510	83,17	84,95	3,36	0,00	3,19
21	**499**	**91**	**1759**	**499**	**499,00**	**10,83**	**1,90**	**8,93**

$$R^2 = \frac{\sum\left(f(x)-\bar{y}\right)^2}{\sum\left(y_i-\bar{y}\right)^2} = \frac{8,93}{10,83}$$

$$R^2 \cong 0,824$$

$$R \cong 0,908$$

ANOVA

O poder de explicação do modelo de regressão tem um poder de explicação superior ao da média?

$y_i \approx f(x_i) = a + bx$

$y_i \approx y_{Médio}$

Fonte de variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	$F_{Calculado}$
Regressão	1	$SQR = \sum (f(x) - \bar{y})^2$	$QMR = \frac{SQR}{1}$	$F_{Calc} = \frac{QMR}{QME}$
Erro	N-2	$SQE = \sum (y - f(x))^2$	$QME = \frac{SQE}{N-2}$	
Total	N-1	$SQT = \sum (y - \bar{y})^2$	$QMT = \frac{SQT}{N-1}$	

Exemplo

Fonte de variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	$F_{Calculado}$
Regressão	1	$SQR = \sum (f(x) - \bar{y})^2$	$QMR = \frac{SQR}{1}$	$F_{Calc} = \frac{QMR}{QME}$
Erro	N-2	$SQE = \sum (y - f(x))^2$	$QME = \frac{SQE}{N-2}$	
Total	N-1	$SQT = \sum (y - \bar{y})^2$	$QMT = \frac{SQT}{N-1}$	

Fonte de variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	$F_{Calculado}$
Regressão	1	SQR = 8,93	QMR = 8,93	18,8
Erro	4	SQE = 1,90	QME = 0,475	
Total	5	SQT = 10,83	QMT = 2,166	

$F_{Crítico} = F_{1;4;5\%} = 7,709$ ou $\alpha = 1,23\%$

Inferência sobre o coeficiente angular

H_0: $\beta = \beta_0$

H_1: $\beta \neq \beta_0$

$$s_e = \sqrt{QME} = \sqrt{\frac{SQE}{N-2}}$$

$$s_e = \sqrt{\frac{\sum y_i^2 - a\sum y - b\sum x_i y_i}{N-2}}$$

$$s_b = s_e \sqrt{\frac{N}{N\sum x_i^2 - \left(\sum x_i\right)^2}}$$

$$t = \frac{b - \beta_0}{s_b}$$

Intervalo de confiança para o coeficiente angular

$$\beta = b \pm t_{N-2;\alpha}\, s_b$$

Inferência sobre o coeficiente angular

$QME = 0,476$

$S_e = 0,690$

$S_b = 0,165$

$t_{Calc} = 4,330$

$t_{Tab} = t_{N-2;\alpha} = 2,776$

$t\ S_b = 0,458$

$b_{Mín} = 0,256$

$b_{Mán} = 1,172$ $\qquad$ $0,256 \leq b \leq 1,172$

Inferência sobre o coeficiente linear

H_0: $\alpha = \alpha_0$

H_1: $\alpha \neq \alpha_0$

$$s_e = \sqrt{QME} = \sqrt{\frac{SQE}{N-2}}$$

$$s_e = \sqrt{\frac{\sum y_i^2 - a\sum y - b\sum x_i y_i}{N-2}}$$

$$s_a = s_e \sqrt{\frac{1}{N} + \frac{\left(\frac{\sum x_i}{N}\right)^2}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{N}}}$$

$$t = \frac{a - \alpha_0}{s_a}$$

Intervalo de confiança para o coeficiente linear

$$\alpha = a \pm t_{N-2;\alpha}\, s_a$$

Inferência sobre o coeficiente linear

$QME = 0,476$

$S_e = 0,690$

$S_a = 0,642$

$t_{Calc} = 125,568$

$t_{Tab} = t_{N-2;\alpha} = 2,776$

$t\ S_a = 1,784$

$b_{Mín} = 78,883$

$b_{Mán} = 82,450$ $\qquad 78,883 \leq b \leq 82,450$

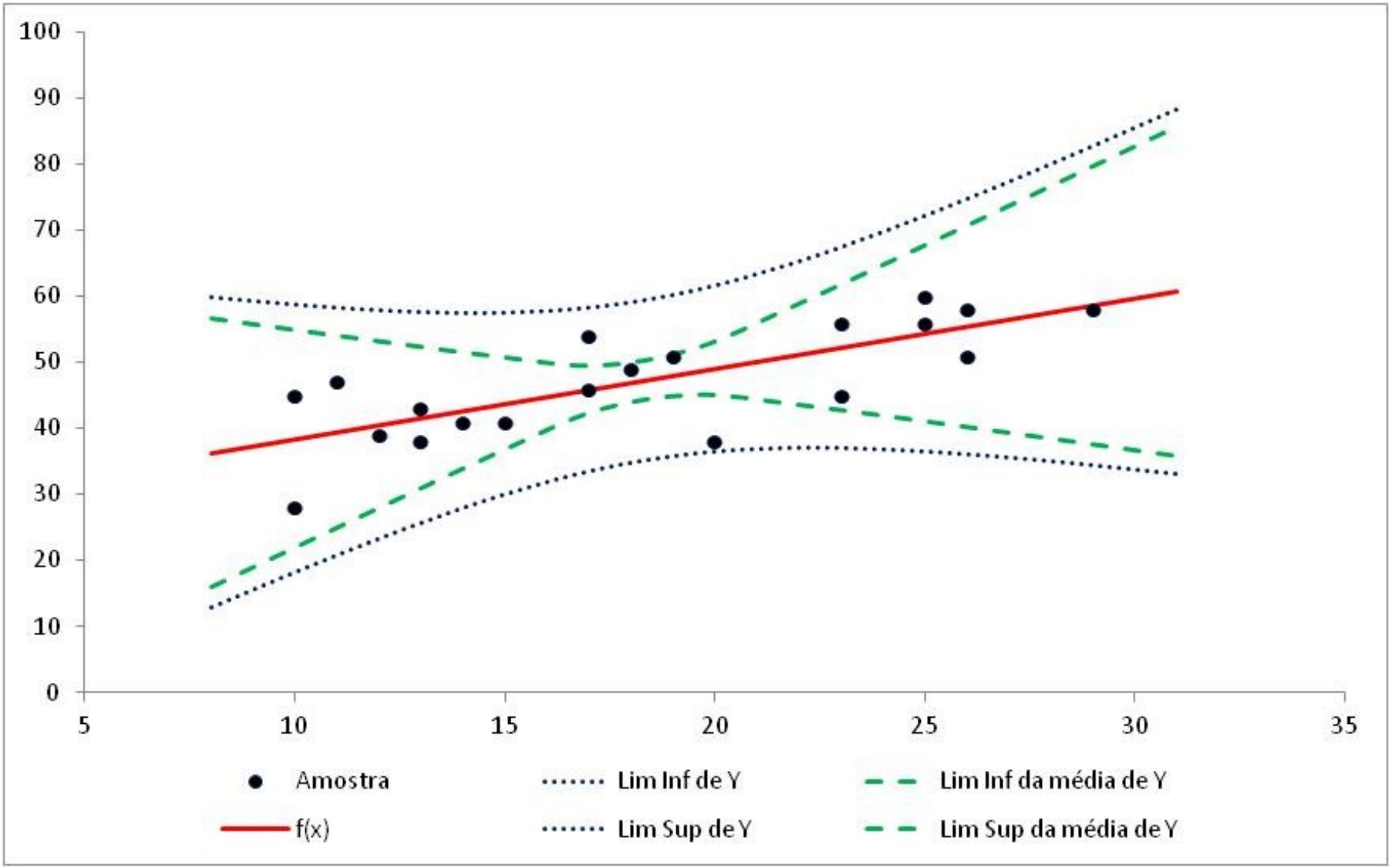
IC para o valor esperado de y
IC para o valor de y (IC de predição)

$$\mu(y) = f(x) \pm t_{N-2;\alpha/2} s_e \sqrt{\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{N s_x^2}}$$

$$\mu(y) = f(x) \pm t_{N-2;\alpha/2} s_e \sqrt{\frac{1}{N} + \frac{N(x_0 - \bar{x})^2}{N(\sum x^2) - (\sum x)^2}}$$

$$y = f(x) \pm t_{N-2;\alpha/2} s_e \sqrt{1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{N s_x^2}}$$

$$\mu(y) = f(x) \pm t_{N-2;\alpha/2} s_e \sqrt{1 + \frac{1}{N} + \frac{N(x_0 - \bar{x})^2}{N(\sum x^2) - (\sum x)^2}}$$


100
90
80
70
60
50
40
30
20
10
0
5
10
15
20
25
30
35
Amostra
f(x)
Lim Inf de Y
Lim Sup de Y
Lim Inf da média de Y
Lim Sup da média de Y

Exercício

Com os dados da tabela ao lado

1. Ache a reta de regressão
2. Construa IC (95%) para os coeficientes da reta
3. Calcule o coeficiente de determinação
4. Faça uma previsão para y e E[y] quando
   a) x=5
   b) x=9

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	2,0