

Estatística Aplicada às Ciências Sociais

Sexta Edição

Pedro Alberto Barbetta

Florianópolis: Editora da UFSC, 2006

Cap. 13 – Correlação e Regressão

Correlação

- X e Y variáveis quantitativas



Correlação

- X e Y estão **positivamente correlacionadas** quando elas caminham num mesmo sentido;
- Estão **negativamente correlacionadas** quando elas caminham em sentidos opostos.

Correlação

- Correlação não implica relação de causa-e-efeito

Exemplo

- Amostra de municípios. Variáveis:
 - DistCap: distância à capital da respectiva Unidade da Federação.
 - EspVida: esperança de vida ao nascer
 - MortInf: mortalidade (número médio de mortes em 1.000) até um ano de idade.
 - Alfab: taxa de alfabetização (percentagem da população adulta alfabetizada).
 - Renda: renda *per capita* do município (R\$).

Tabela 13.1

Município	DistCap	EspVida	MortInf	Alfab	Renda
Araruna (PR)	365	67,99	23,19	86,23	188,29
Nova Redenção (BA)	278	61,19	56,56	63,00	74,79
Monção (MA)	150	59,58	63,32	63,64	66,96
Porto Rico do Maranhão (MA)	78	58,96	66,05	79,33	65,34
Campo Erê (SC)	468	68,10	31,71	83,38	173,38
Lagoa do Piauí (PI)	40	63,65	47,08	65,81	60,00
São José das Palmeiras (PR)	486	71,01	16,62	77,54	150,67
Paraíba do Sul (RJ)	83	71,36	15,69	89,28	264,55
Malhada dos Bois (SE)	65	64,46	44,18	69,95	80,69
Jandaíra (BA)	175	62,45	51,57	59,72	58,68
Vespasiano (MG)	14	68,68	32,81	90,43	196,51
Ipaba (MG)	167	67,42	37,04	81,82	125,75

Fonte: Atlas de Desenvolvimento Humano (www.pnud.org.br/atlas)

Diagrama de dispersão:

x	y
365	67,99
278	61,19
150	59,58

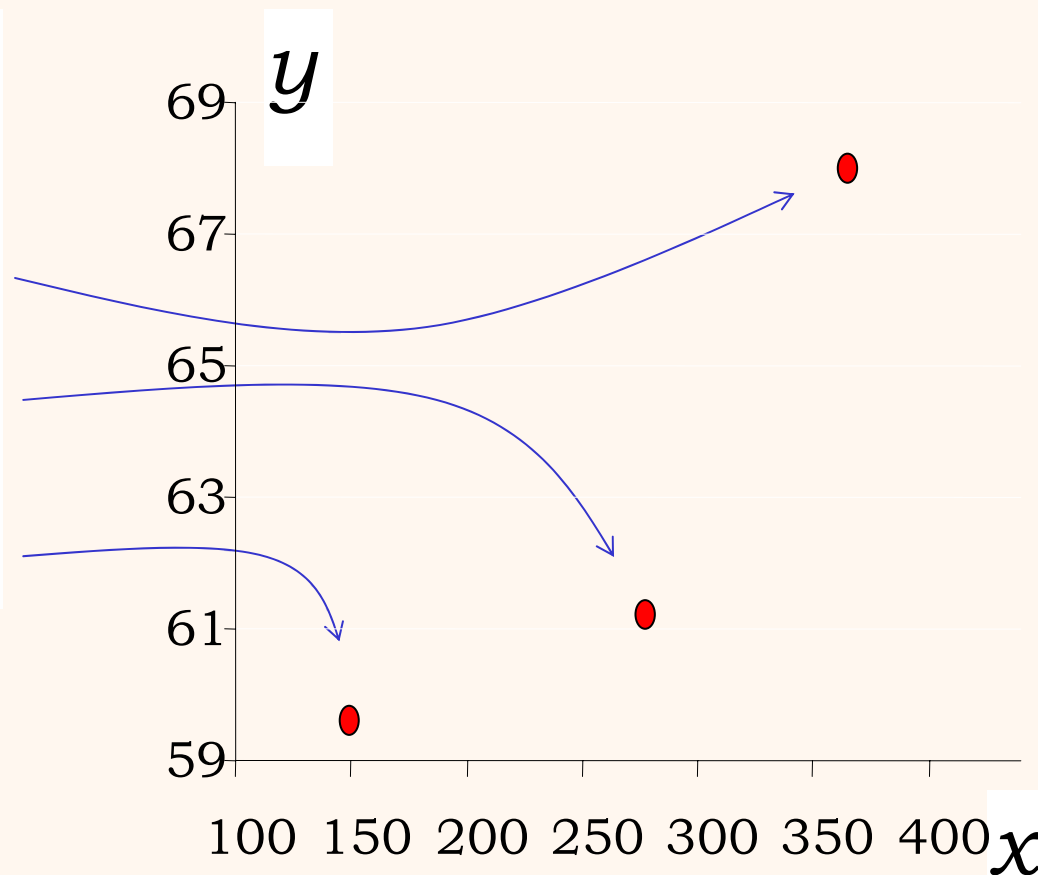
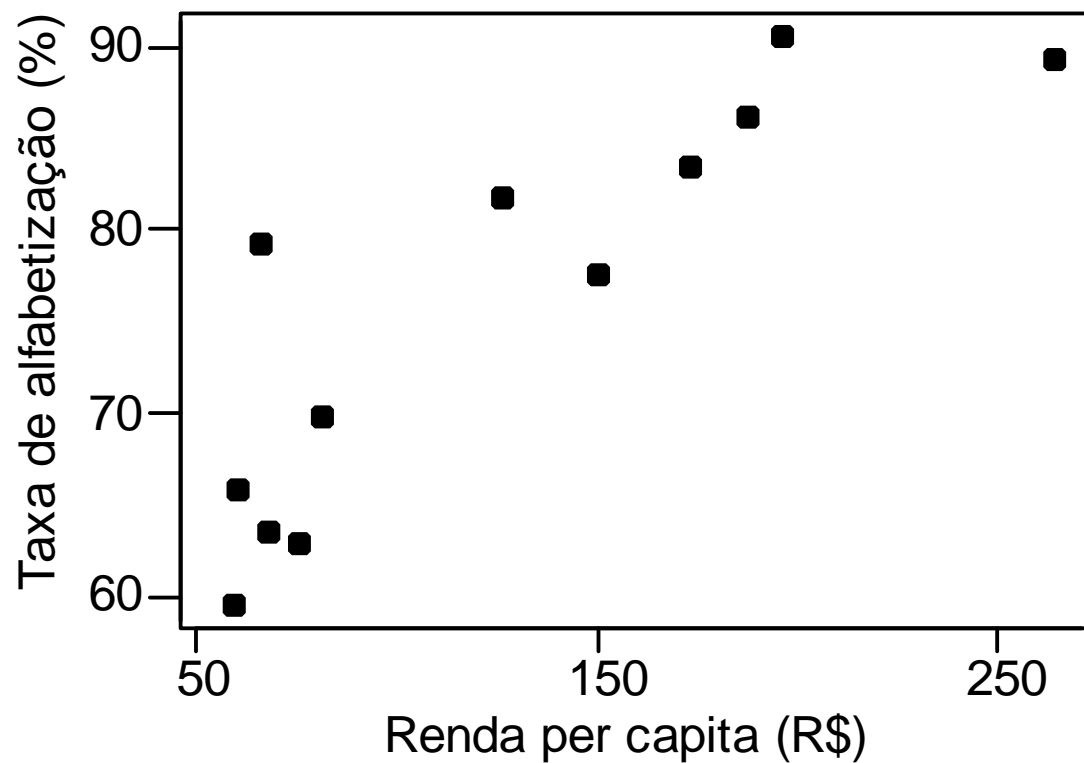
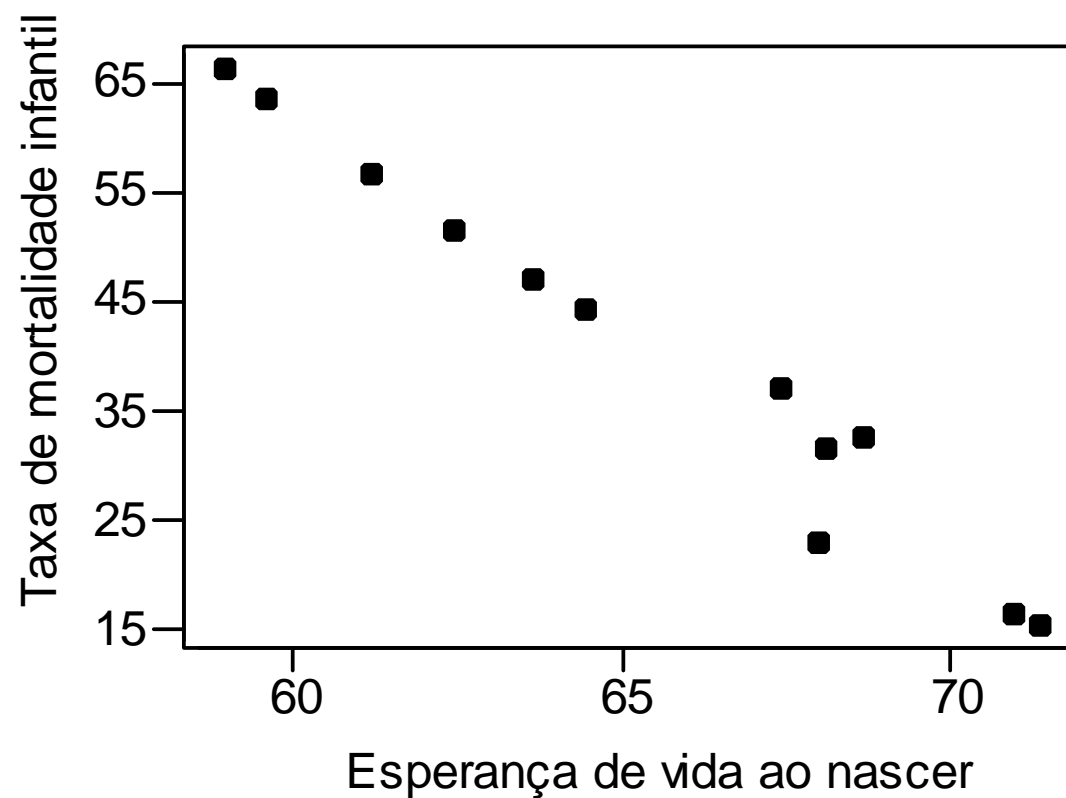


Tabela 13.1 → Diagramas de dispersão:



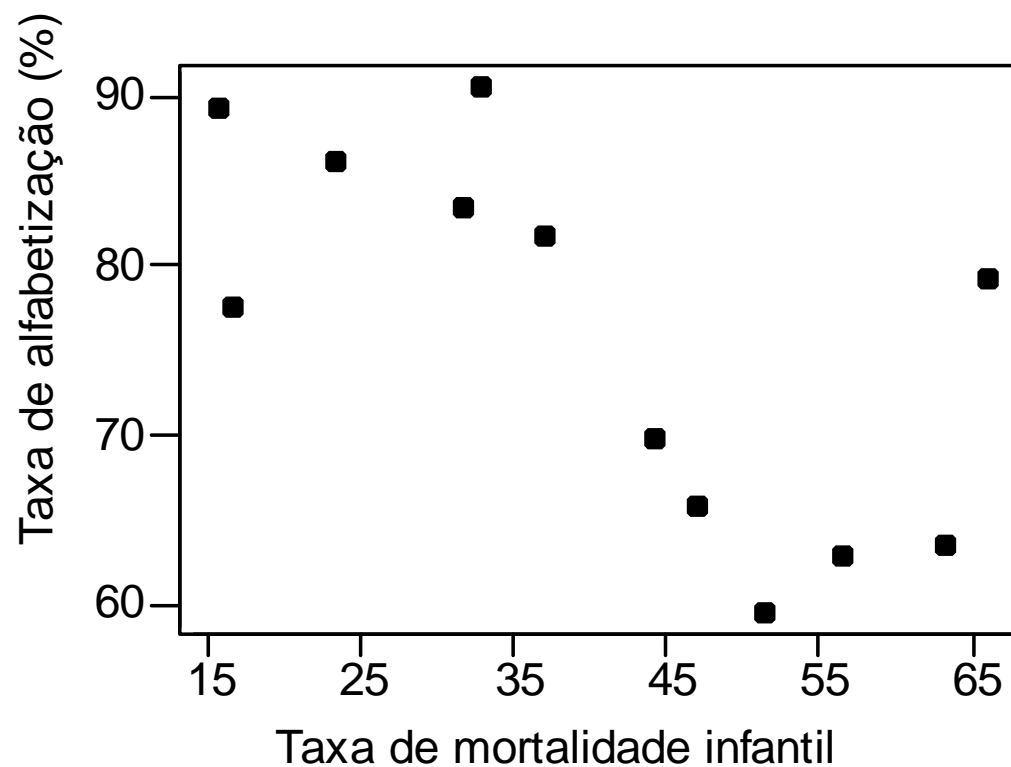
Interpretar a correlação entre as duas variáveis.

Tabela 13.1 → Diagramas de dispersão:



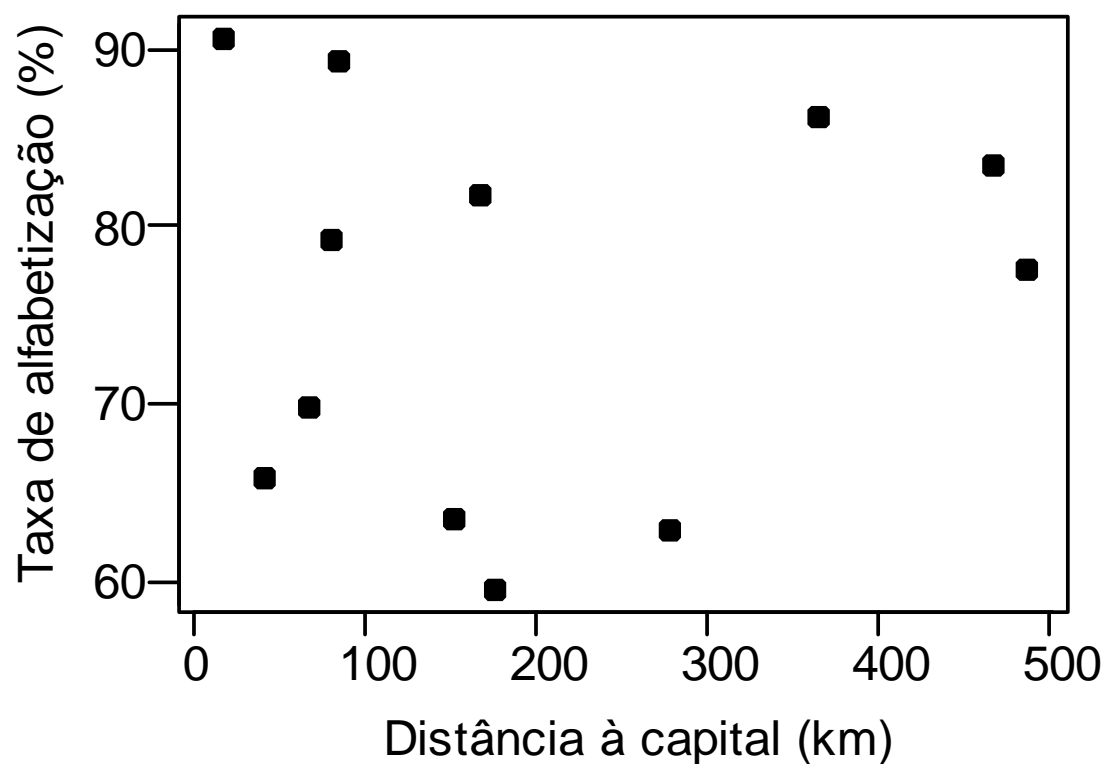
Interpretar a correlação entre as duas variáveis.

Tabela 13.1 → Diagramas de dispersão:



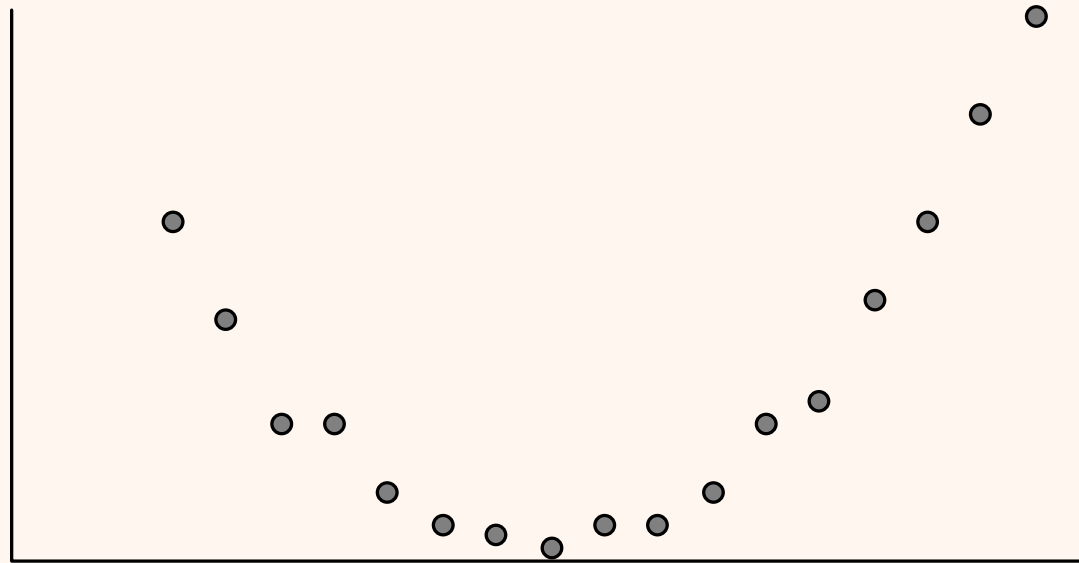
Interpretar a correlação entre as duas variáveis.

Tabela 13.1 → Diagramas de dispersão:



Interpretar a correlação entre as duas variáveis.

Correlação não linear



Coeficiente de Correlação

- Não deve depender da unidade de medida das variáveis
- Padronização $(x, y) \rightarrow (x', y')$ para cada par de valores:

$$x' = \frac{x - \bar{X}}{S_x}$$

Média dos valores de X

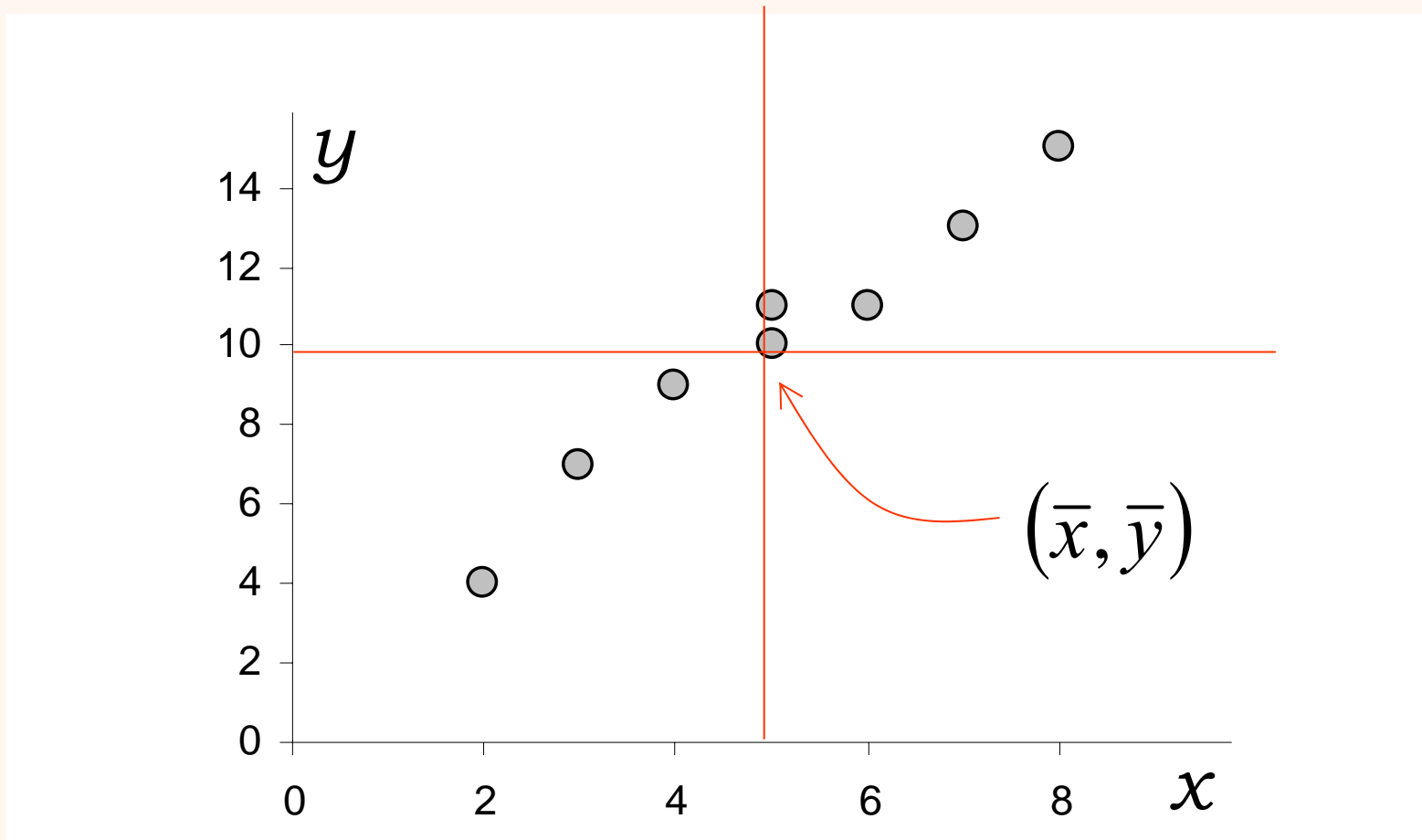
Desvio padrão dos valores de X

$$y' = \frac{y - \bar{Y}}{S_y}$$

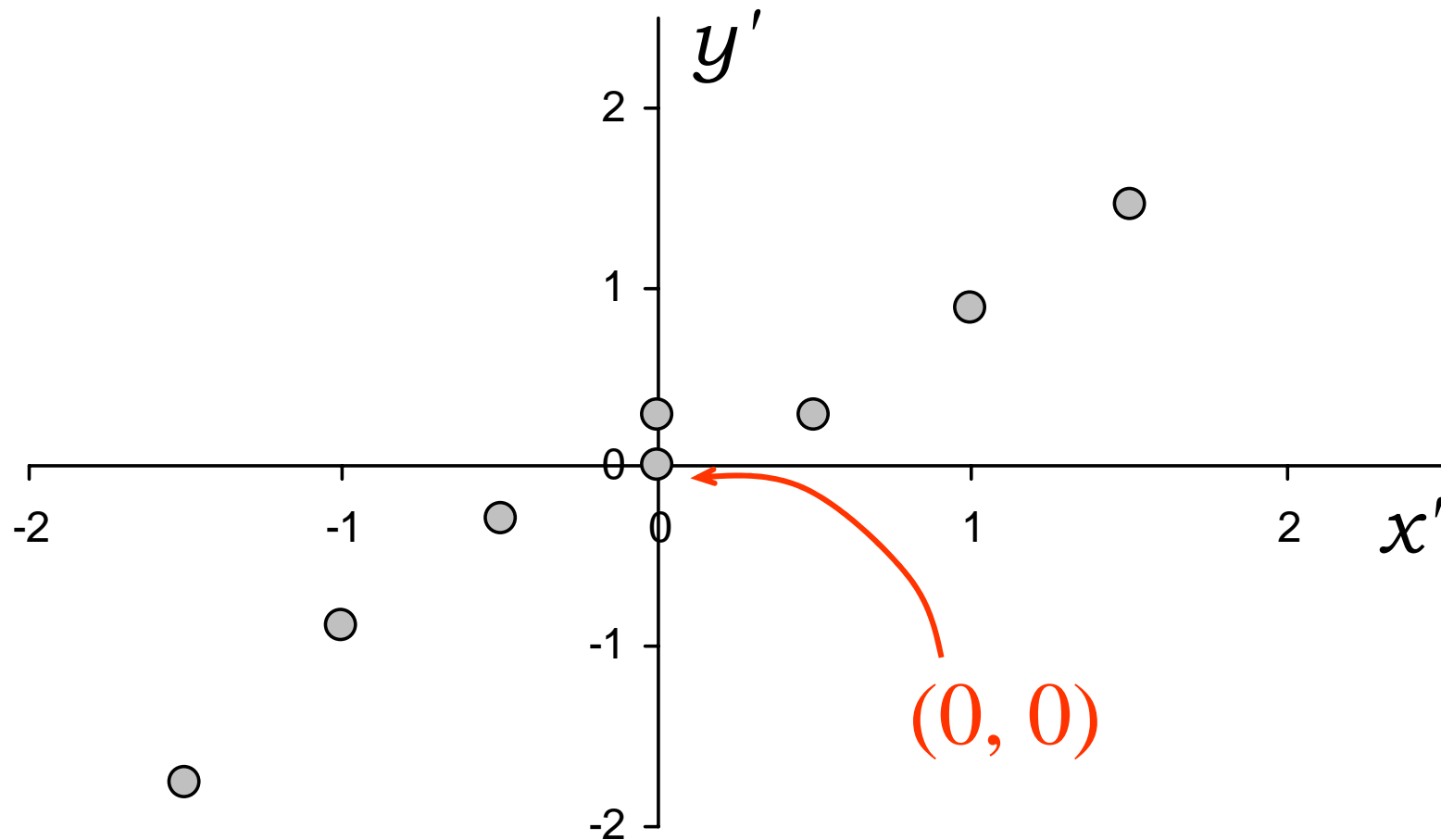
Média dos valores de Y

Desvio padrão dos valores de Y

Efeito da padronização



Padronização



Padronização - Exemplo 13.1

	Valores originais		Valores padronizados		Produtos
	X	Y	X'	Y'	$X' \cdot Y'$
	2	4	-1,50	-1,75	2,63
	3	7	-1,00	-0,88	0,88
	4	9	-0,50	-0,29	0,15
	5	10	0,00	0,00	0,00
	5	11	0,00	0,29	0,00
	6	11	0,50	0,29	0,15
	7	13	1,00	0,88	0,88
	8	15	1,50	1,46	2,19
Soma:	40	80	0,00	0,00	6,87
Média:	5,00	10,00	0,00	0,00	
Desvio padrão:	2,00	3,42	1,00	1,00	

Idéia de construção do Coef. de Correlação de Pearson

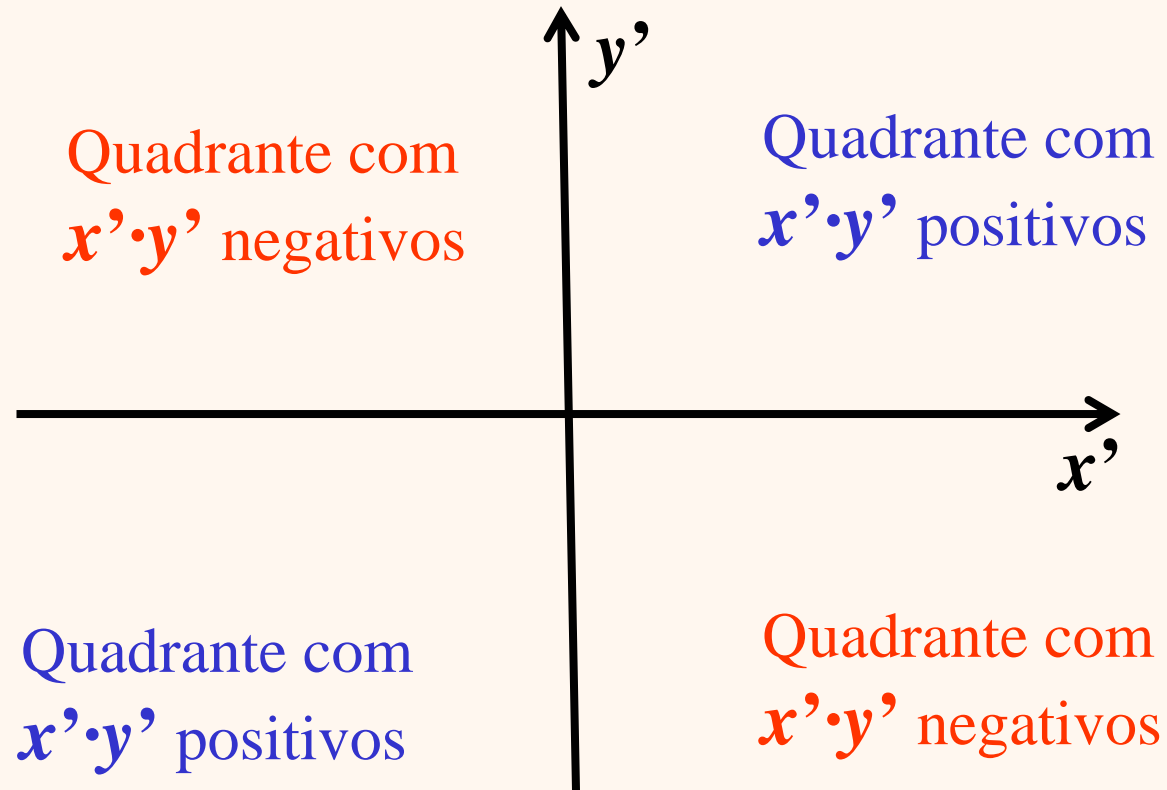
$$x' = \frac{x - \bar{X}}{S_x}$$

$$y' = \frac{y - \bar{Y}}{S_y}$$

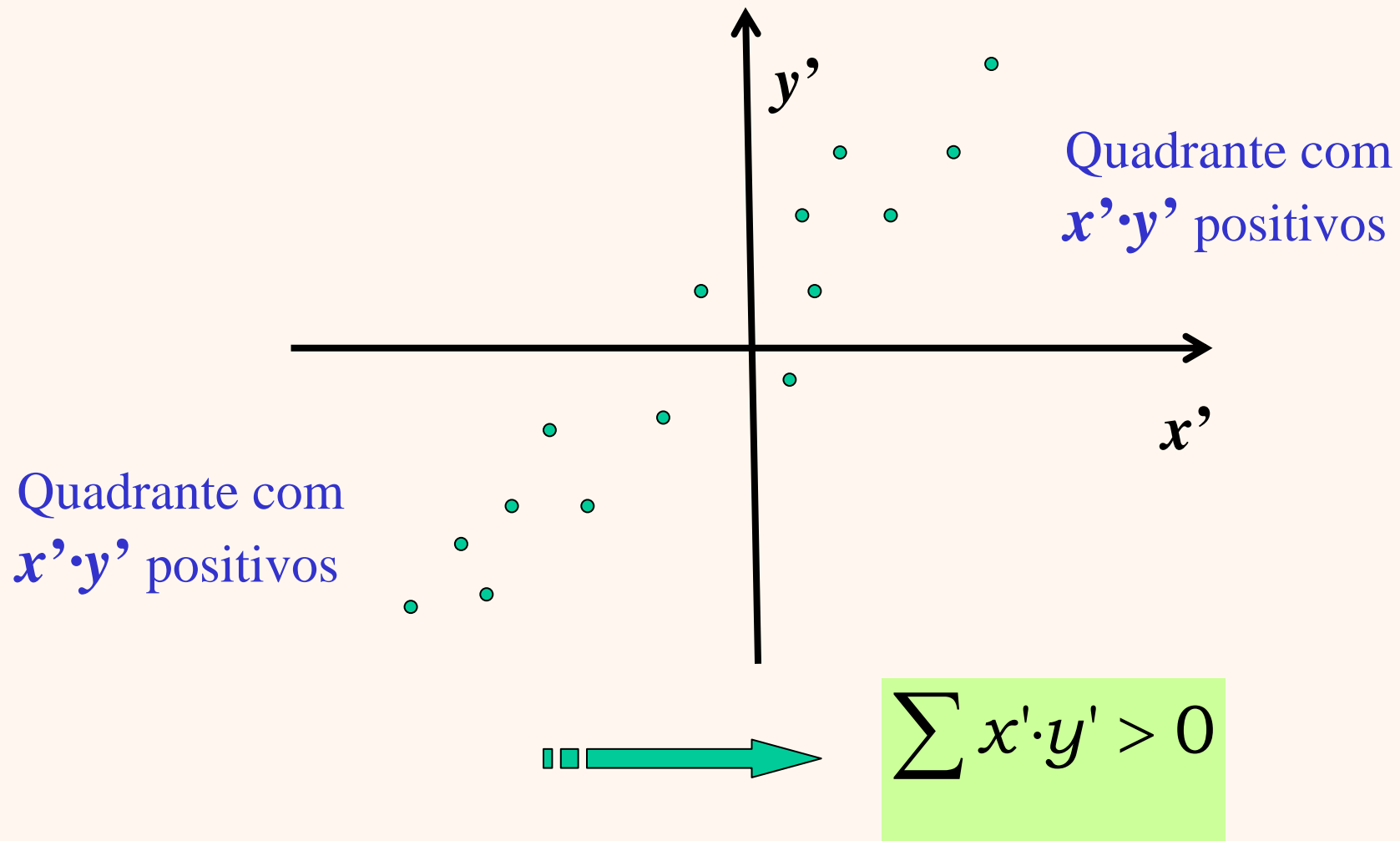
Considere os produtos dos valores padronizados:

$$x' \cdot y'$$

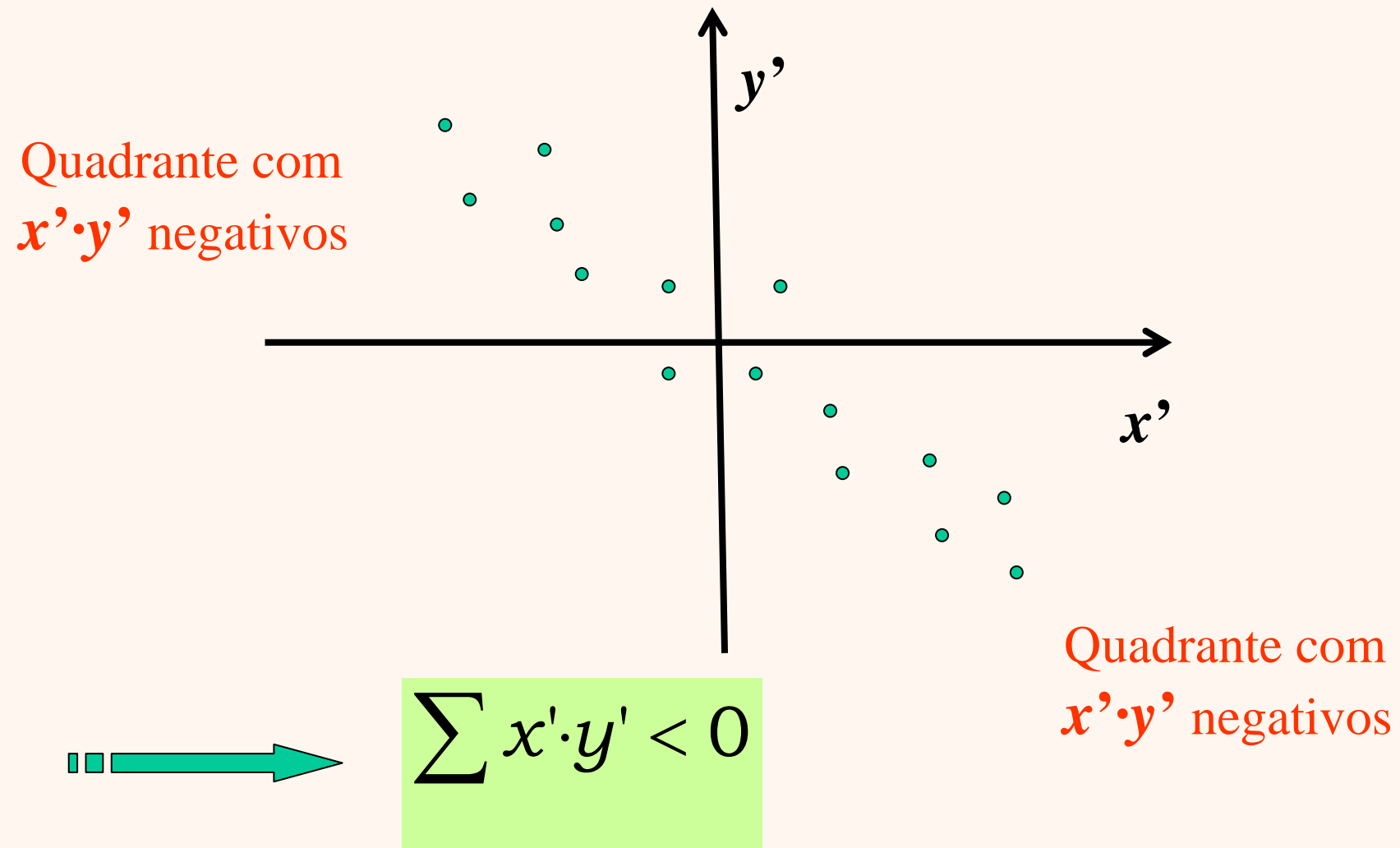
Sinais dos produtos dos valores padronizados:



Sinais dos produtos dos valores padronizados:



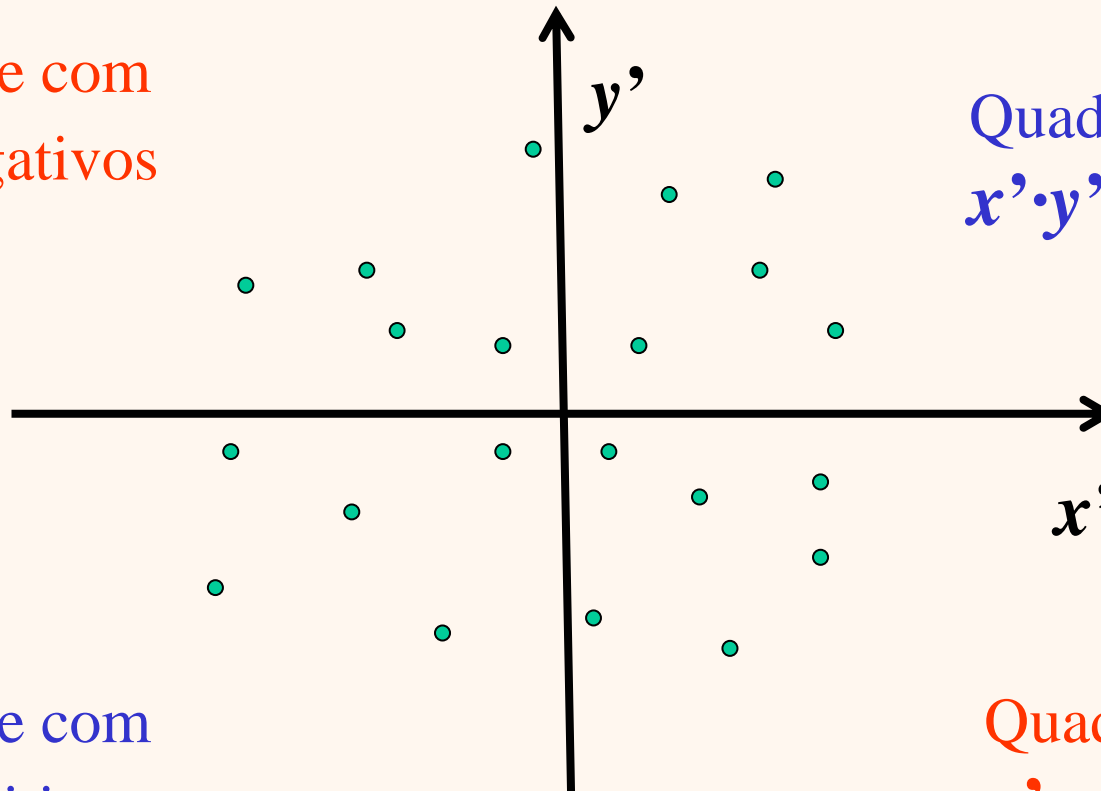
Sinais dos produtos dos valores padronizados:



Sinais dos produtos dos valores padronizados:

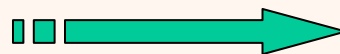
Quadrante com
 $x' \cdot y'$ negativos

Quadrante com
 $x' \cdot y'$ positivos



Quadrante com
 $x' \cdot y'$ positivos

Quadrante com
 $x' \cdot y'$ negativos



$$\sum x' \cdot y' \approx 0$$

Idéia de construção do Coef. de Correlação de Pearson

- Padronização $(x, y) \rightarrow (x', y')$:

$$x' = \frac{x - \bar{X}}{S_x}$$

$$y' = \frac{y - \bar{Y}}{S_y}$$

Coef. de Correlação de Pearson:

$$r = \frac{\sum (x' \cdot y')}{n - 1}$$

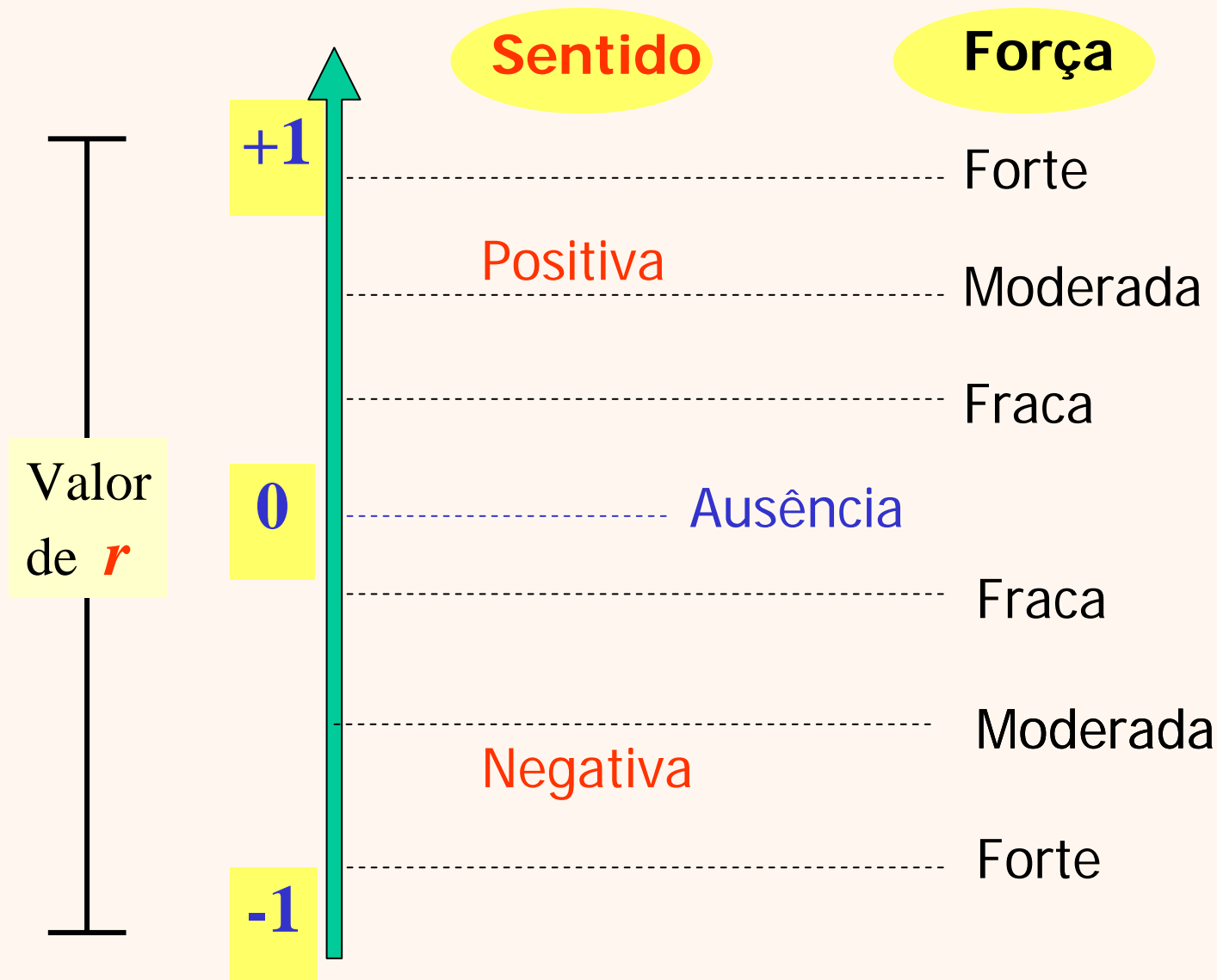
Mede a correlação linear entre X e Y.

Exemplo 13.1

	Valores originais		Valores padronizados		Produtos
	X	Y	X'	Y'	X'·Y'
	2	4	-1,50	-1,75	2,63
	3	7	-1,00	-0,88	0,88
	4	9	-0,50	-0,29	0,15
	5	10	0,00	0,00	0,00
	5	11	0,00	0,29	0,00
	6	11	0,50	0,29	0,15
	7	13	1,00	0,88	0,88
	8	15	1,50	1,46	2,19
Soma:	40	80	0,00	0,00	6,87
Média:	5,00	10,00	0,00	0,00	
Desvio padrão:	2,00	3,42	1,00	1,00	

$$r = \frac{\sum (x' \cdot y')}{n - 1} = \frac{6,87}{7} = 0,981$$

Valores possíveis de r e interpretação da correlação



Matriz de correlações. Dados da Tab. 13.1

	DISTCAP	ESPVIDA	MORTINF	ALF	RENDA
DISTCAP	1	0,337	-0,400	0,087	0,205
ESPVIDA	0,337	1	-0,983	0,718	0,865
MORTINF	-0,400	-0,983	1	-0,684	-0,860
ALF	0,087	0,718	-0,684	1	0,863
RENDA	0,205	0,865	-0,860	0,863	1

Interpretar.

Fórmula direta de calcular r

$$r = \frac{n \cdot \sum(X \cdot Y) - (\sum X) \cdot (\sum Y)}{\sqrt{n \cdot \sum X^2 - (\sum X)^2} \sqrt{n \cdot \sum Y^2 - (\sum Y)^2}}$$

Ver exemplo desses cálculos no livro (Tabela 13.3)

Ver, também, no livro:

- Teste de hipóteses sobre a correlação
- Correlação com variáveis indicadoras
- Correlação por postos

Regressão linear simples

- A análise de regressão é geralmente feita sob um referencial teórico que justifique a adoção de alguma relação matemática de causalidade.



Variável independente ou
Variável explicativa

Variável dependente ou
Variável resposta

Regressão linear simples

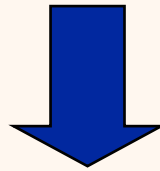
- Predizer valores de uma variável dependente (Y) em função de uma variável independente (X).
- Conhecer o quanto variações de X podem afetar Y.

Exemplos de regressão:

Variável independente (X)	→	Variável dependente (Y)
Renda	→	Consumo (R\$)
Gasto com o controle da qualidade (R\$)	→	Número de defeitos nos produtos
Memória RAM do computador (Gb)	→	Tempo de resposta do sistema (segundos)
Área construída do imóvel (m ²)	→	Preço do imóvel (R\$)

Regressão

Amostra de observações
de (X, Y)



Conhecer o relacionamento
entre X e Y

Regressão - Modelo

$$Y = \left[\begin{array}{c} \text{Predito por } X, \text{ se-} \\ \text{gundo uma função} \end{array} \right] + \left[\begin{array}{c} \text{Efeito aleatório} \end{array} \right]$$

$$y = \alpha + \beta \cdot x + e$$

Parâmetros

Regressão
Linear
Simples

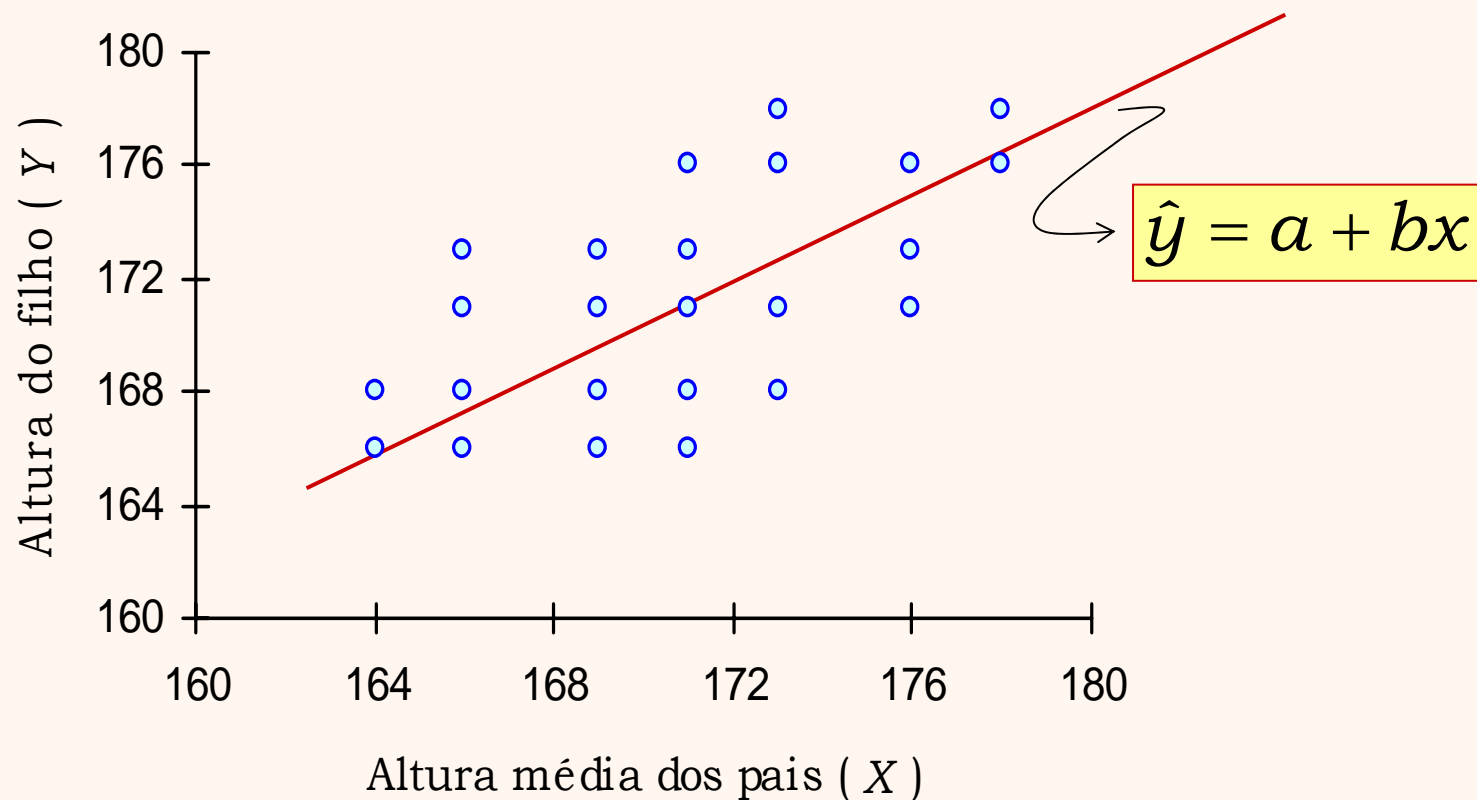
Pressupostos do modelo de regressão

$$y = \alpha + \beta.x + e$$

- Os erros (e 's) são independentes e variam aleatoriamente segundo uma distribuição (normal) com média zero e variância constante.

Estimativas dos parâmetros α e β

- Construção da equação de regressão com base nos dados:



Estimativas dos parâmetros α e β

- Construção da equação de regressão com base nos dados:

$$\hat{y} = a + bx$$

$$b = \frac{n \cdot \sum(X \cdot Y) - (\sum X) \cdot (\sum Y)}{n \cdot \sum X^2 - (\sum X)^2} \quad (\text{estimativa do beta})$$

$$a = \frac{\sum Y - b \cdot \sum X}{n} \quad (\text{estimativa do alfa})$$

Estimativas dos parâmetros α e β

- Construção da equação de regressão com base nos dados.

Exemplo 13.5:

$$b = \frac{n \cdot \sum(X \cdot Y) - (\sum X) \cdot (\sum Y)}{n \cdot \sum X^2 - (\sum X)^2}$$

$$a = \frac{\sum Y - b \cdot \sum X}{n}$$

$$b = \frac{9 \cdot (263.483) - (1.539) \cdot (1.540)}{9 \cdot (263.333) - (1.539)^2} = \frac{1.287}{1.476} = 0,872$$

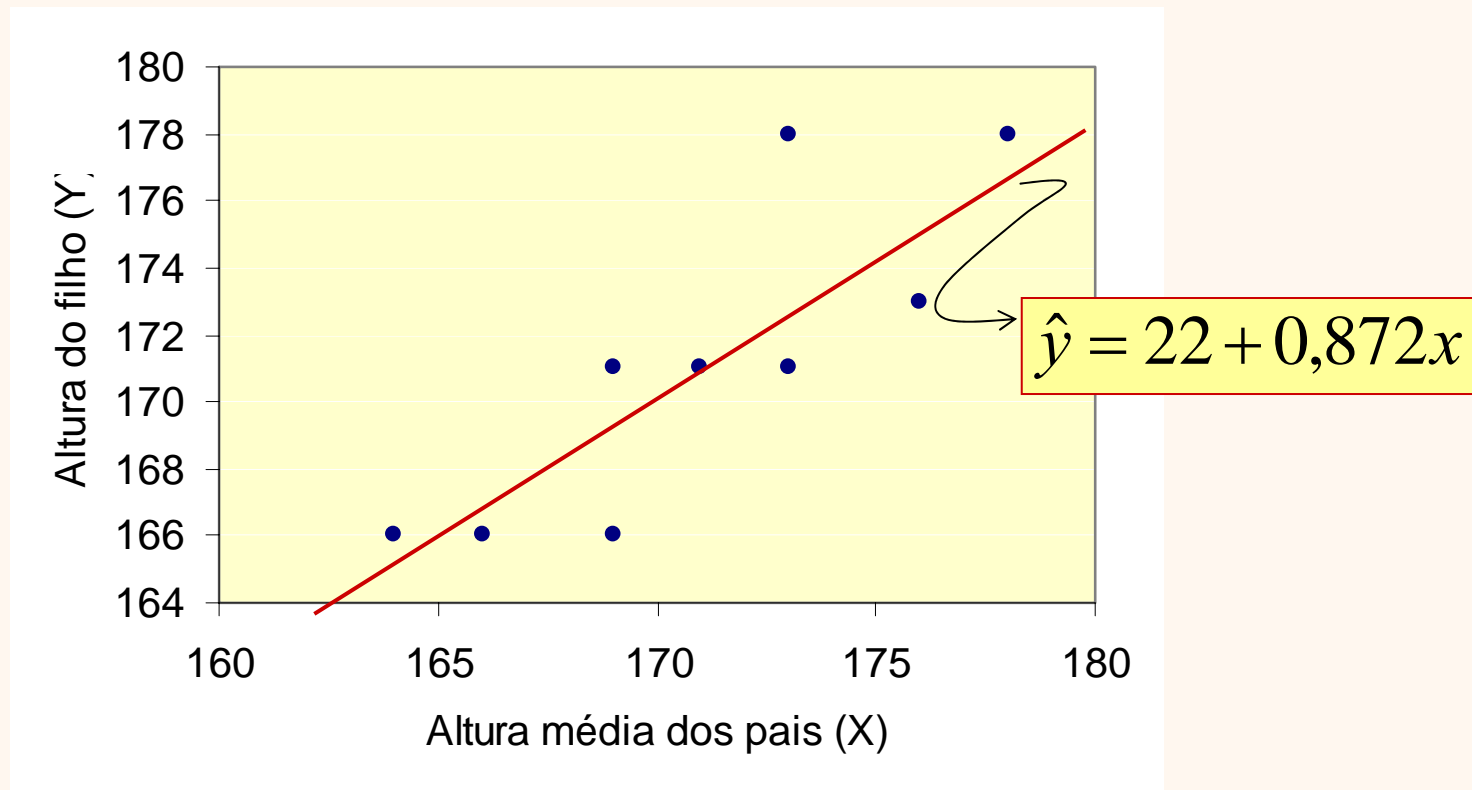
$$a = \frac{1.540 - (0,872) \cdot (1.539)}{9} = 22,00$$

Dados		Cálculos	
X	Y	X ²	X · Y
164	166	26.896	27.224
166	166	27.556	27.556
169	171	28.561	28.899
169	166	28.561	28.054
171	171	29.241	29.241
173	171	29.929	29.583
173	178	29.929	30.794
176	173	30.976	30.448
178	178	31.684	31.684
1.539	1.540	263.333	263.483

Estimativas dos parâmetros α e β

- Construção da equação de regressão com base nos dados.

Exemplo:



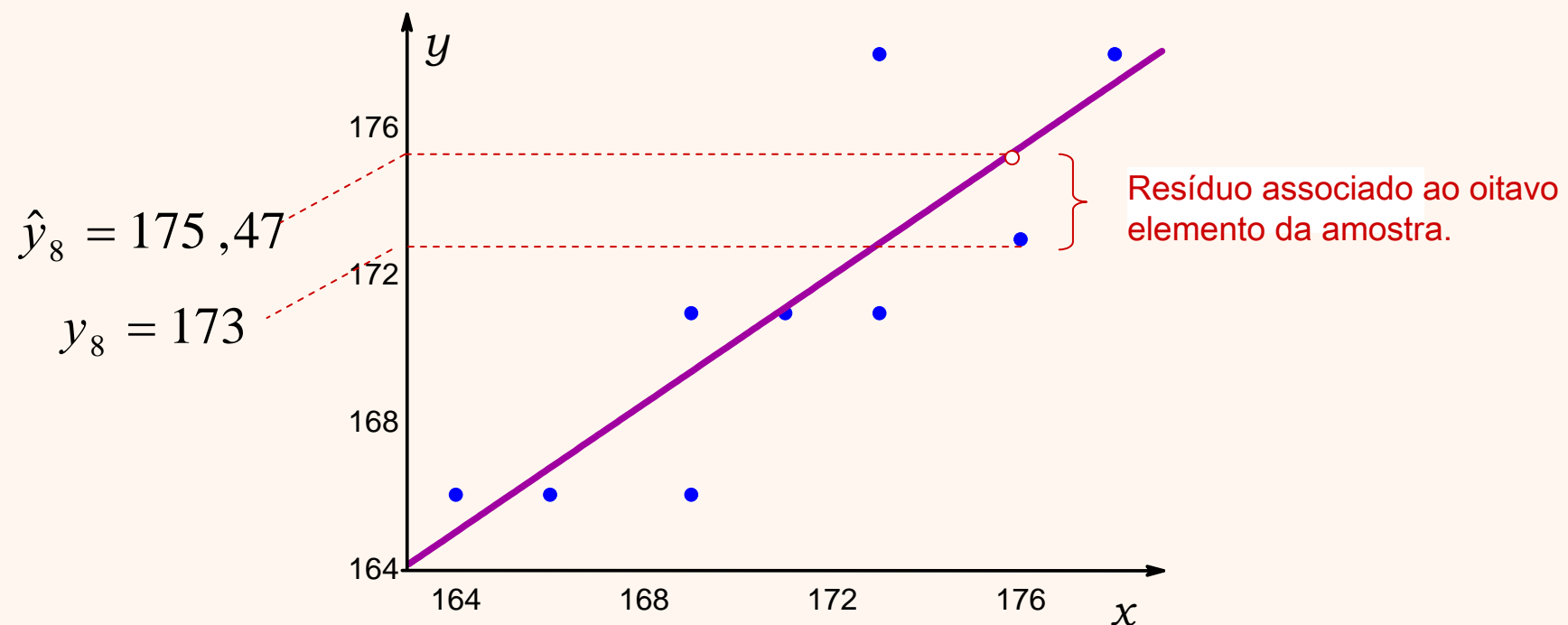
Valores preditos e resíduos

x	y	Predito	Resíduo
164	166	165,01	0,992
166	166	166,75	-0,752
169	171	169,37	1,632
169	166	169,37	-3,368
171	171	171,11	-0,112
173	171	172,86	-1,856
173	178	172,86	5,144
176	173	175,47	-2,472
178	178	177,22	0,784

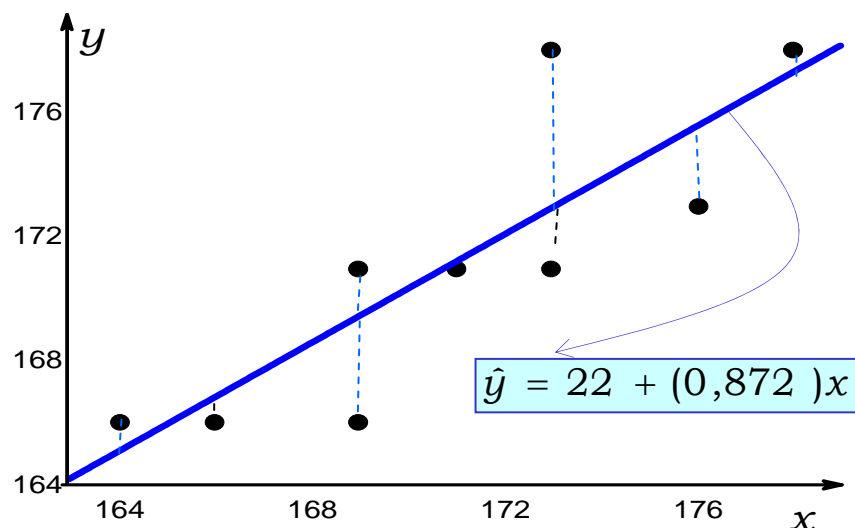
$\hat{y} = 22 + 0,872x$

$\hat{e} = y - \hat{y}$

Valores preditos e resíduos



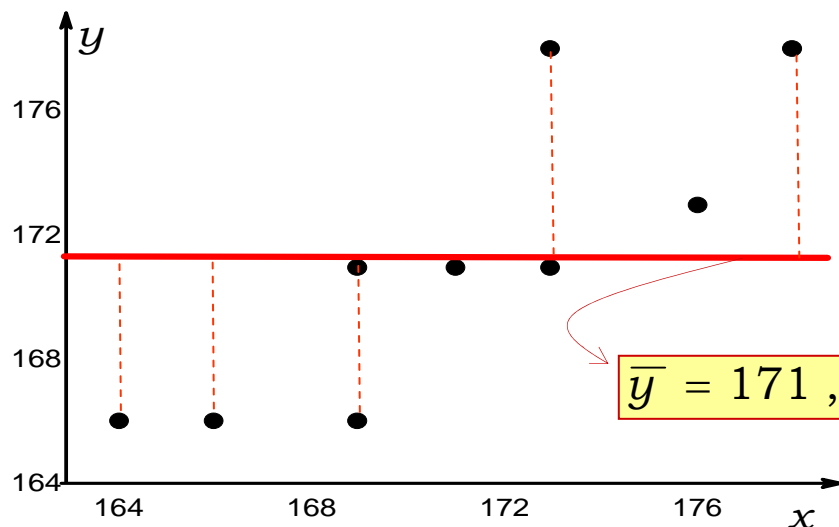
Variação explicada e não-explicada



Variação explicada
pelo modelo de regressão

Soma de quadrados devida ao
erro aleatório:

$$SQE = \sum (y - \hat{y})^2$$



Variação em relação à média
aritmética (variação total)

Soma de quadrado total:

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

Variação explicada e não-explicada

Soma de quadrado total:

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

Soma de quadrados do erro:

$$SQE = \sum (y - \hat{y})^2$$

Soma de quadrados da regressão:

$$SQR = SQT - SQE$$

Coeficiente de determinação:

$$R^2 = \frac{SQR}{SQT} = \frac{\text{variação explicada}}{\text{variação total}}$$

$$0 \leq R^2 \leq 1$$

Variação explicada e não-explicada. Exemplo 13.5

x	y	Média \bar{y}	$y - \bar{y}$	$(y - \bar{y})^2$
164	166	171,11	-5,11	26,11
166	166		-5,11	26,11
169	171		-0,11	0,01
169	166		-5,11	26,11
171	171		-0,11	0,01
173	171		-0,11	0,01
173	178	176,89	6,89	47,47
176	173		1,89	3,57
178	178		6,89	47,47
			0	177

X = altura média dos pais

Y = altura do filho

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

Variação explicada e não-explicada. Exemplo 13.5

x	y	Preditos \hat{y}	$y - \hat{y}$	$(y - \hat{y})^2$
164	166	165,01	0,992	0,98
166	166	166,75	-0,752	0,56
169	171	169,37	1,632	2,66
169	166	169,37	-3,368	11,36
171	171	171,11	-0,112	0,01
173	171	172,86	-1,856	3,46
173	178	172,86	5,144	26,42
176	173	175,47	-2,472	6,1
178	178	177,22	0,784	0,61
			0	52



$$SQE = \sum (y - \hat{y})^2$$

Variação explicada e não-explicada. Exemplo 13.5

Fonte de variação	Somas de quadrados
Explicada por X pelo modelo de regressão (variação <i>explicada</i>)	$SQR = 125$
Devida ao erro aleatório (variação <i>não-explicada</i>)	$SQE = 52$
Variação total	$SQT = 177$

$$R^2 = \frac{SQR}{SQT} = \frac{125}{177} = 0,706 \text{ ou } 70,6\%$$

Interpretar.