

# Estatística para Cursos de Engenharia e Informática

Pedro Alberto Barbetta / Marcelo Menezes Reis / Antonio Cezar Bornia  
São Paulo: Atlas, 2004

## Cap. 11 – Correlação e Regressão

APOIO:

Fundação de Apoio à Pesquisa Científica e Tecnológica do Estado de Santa Catarina (FAPESC)

Departamento de Informática e Estatística – UFSC (INE/CTC/UFSC)

BARBETTA, REIS e BORNIA – Estatística para Cursos de Engenharia e Informática. Atlas, 2004

# Correlação

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

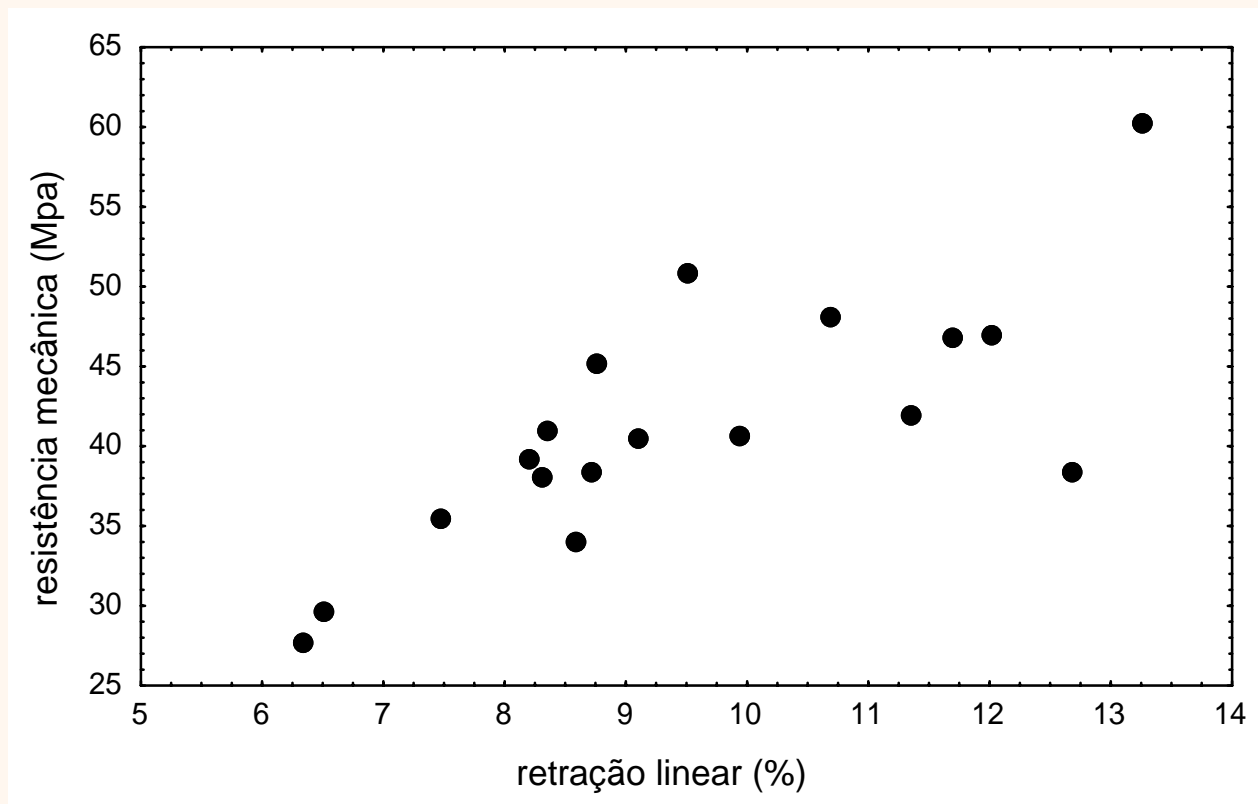
## Exemplo 11.1

- Processo de queima de massa cerâmica para pavimento
  - $X_1$  = retração linear (%),
  - $X_2$  = resistência mecânica (MPa) e
  - $X_3$  = absorção de água (%).

## Exemplo 11.1 - Dados:

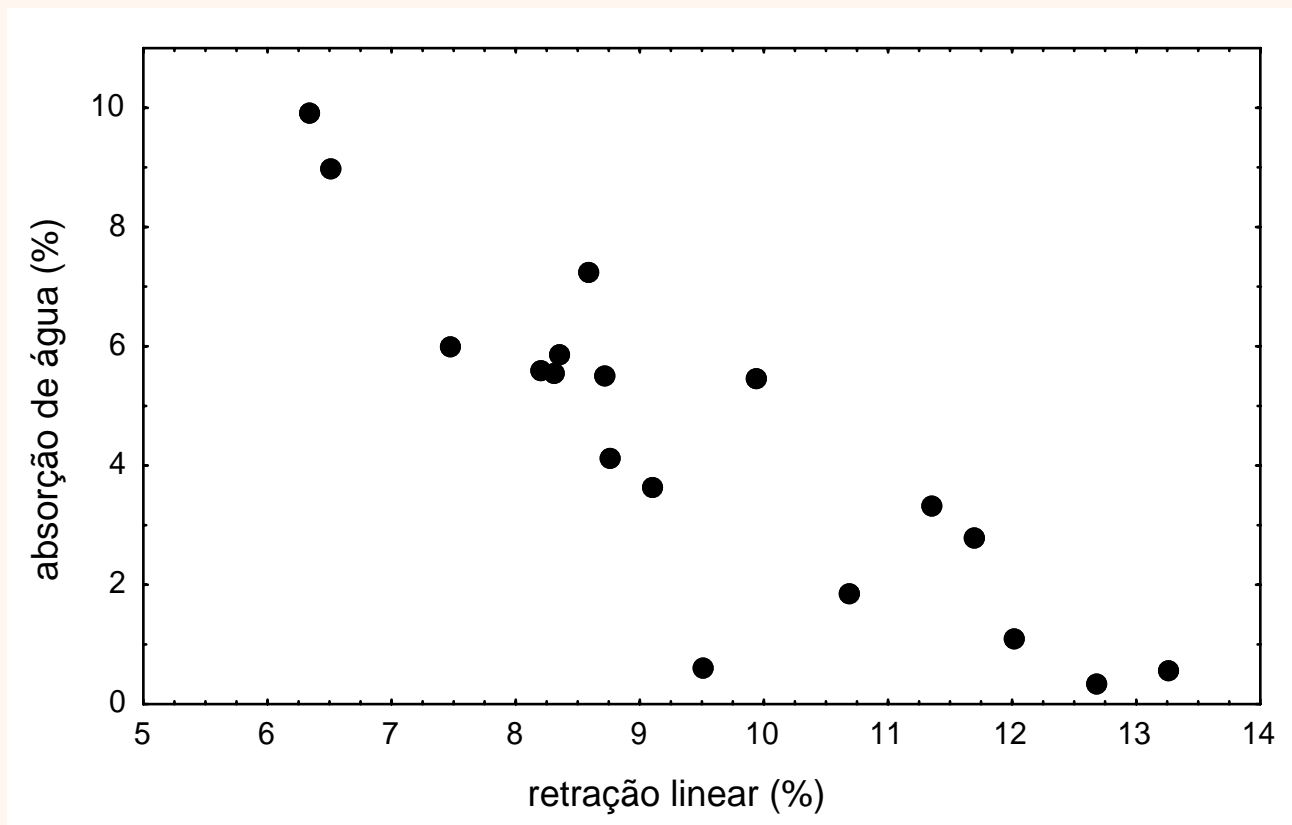
ensaio	$X_1$	$X_2$	$X_3$	ensaio	$X_1$	$X_2$	$X_3$
1	8,70	38,42	5,54	10	13,24	60,24	0,58
2	11,68	46,93	2,83	11	9,10	40,58	3,64
3	8,30	38,05	5,58	12	8,33	41,07	5,87
4	12,00	47,04	1,10	13	11,34	41,94	3,32
5	9,50	50,90	0,64	14	7,48	35,53	6,00
6	8,58	34,10	7,25	15	12,68	38,42	0,36
7	10,68	48,23	1,88	16	8,76	45,26	4,14
8	6,32	27,74	9,92	17	9,93	40,70	5,48
9	8,20	39,20	5,63	18	6,50	29,66	8,98

## Exemplo 11.1 - Diagramas de dispersão:



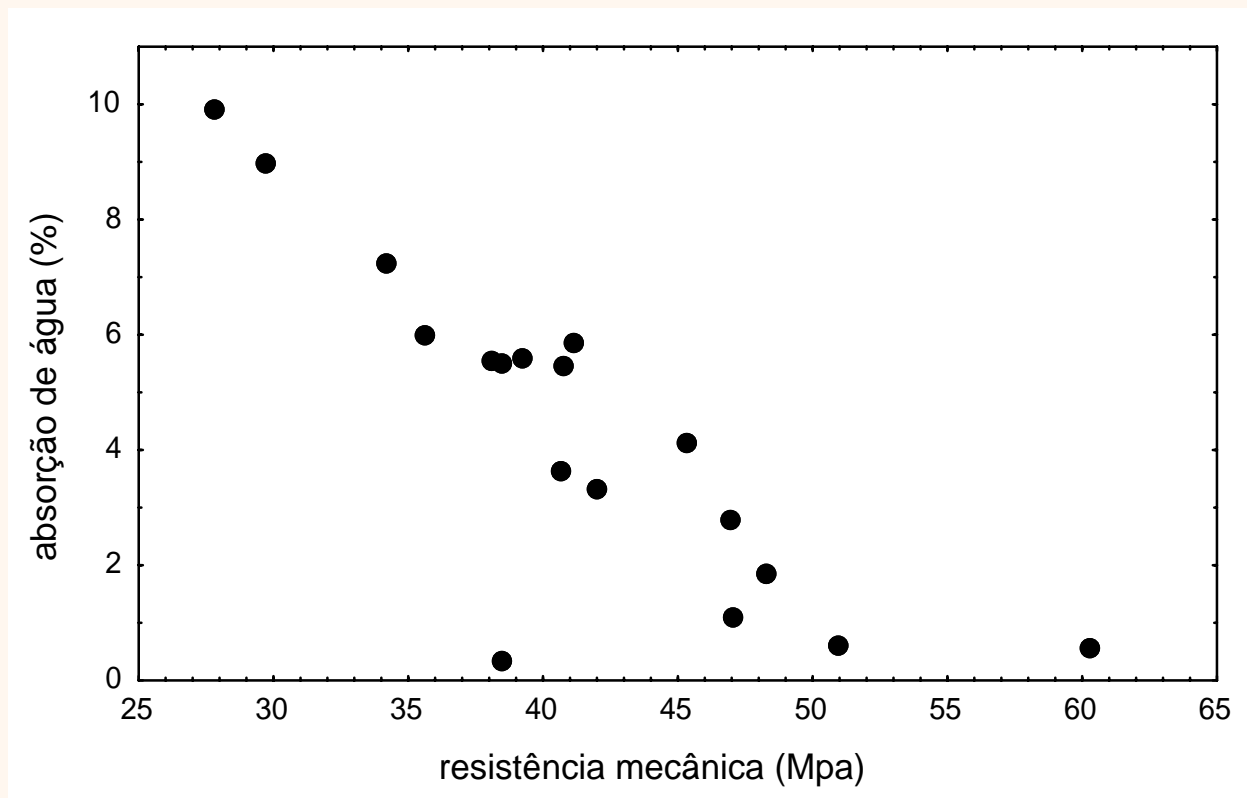
Interpretar a correlação entre as duas variáveis.

## Exemplo 11.1 - Diagramas de dispersão:



Interpretar a correlação entre as duas variáveis.

## Exemplo 11.1 - Diagramas de dispersão:



Interpretar a correlação entre as duas variáveis.

# Idéia de construção do Coef. de Correlação de Pearson

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

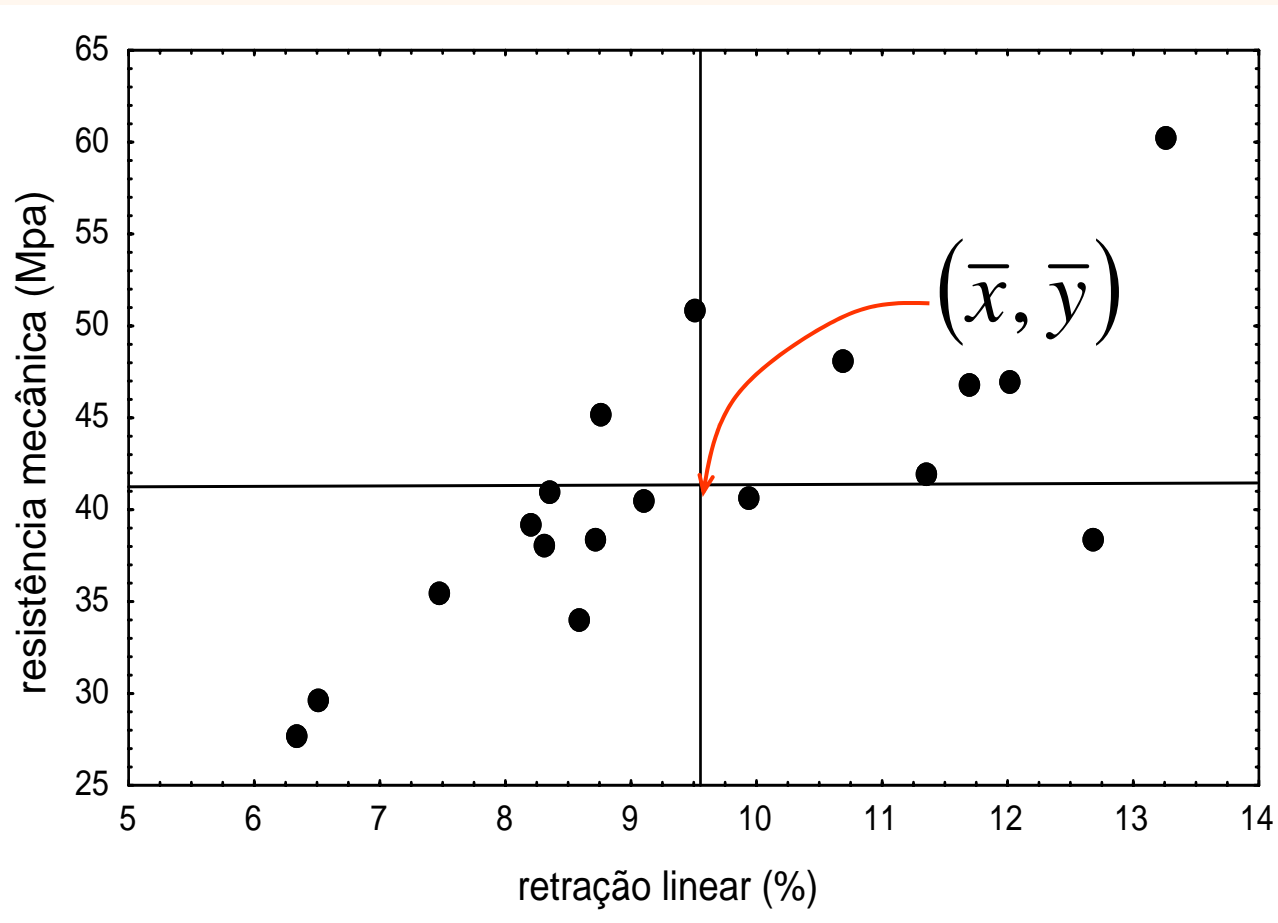
$$x'_i = \frac{x_i - \bar{x}}{s_x}$$

$$y'_i = \frac{y_i - \bar{y}}{s_y}$$

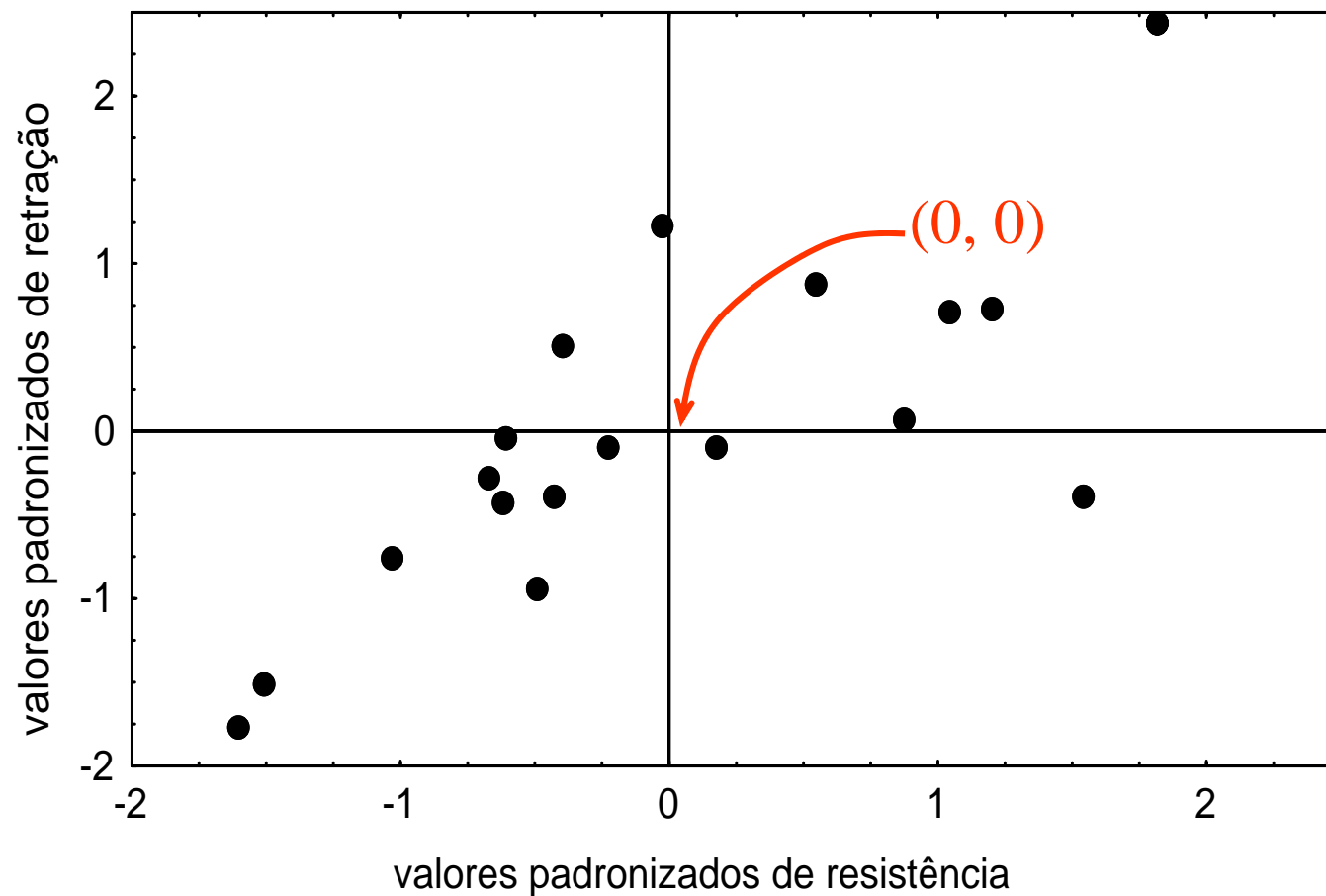
$$(i = 1, 2, \dots, n)$$



## Padronização (Ex. 11.1 a):



## Padronização (Ex. 11.1 a):



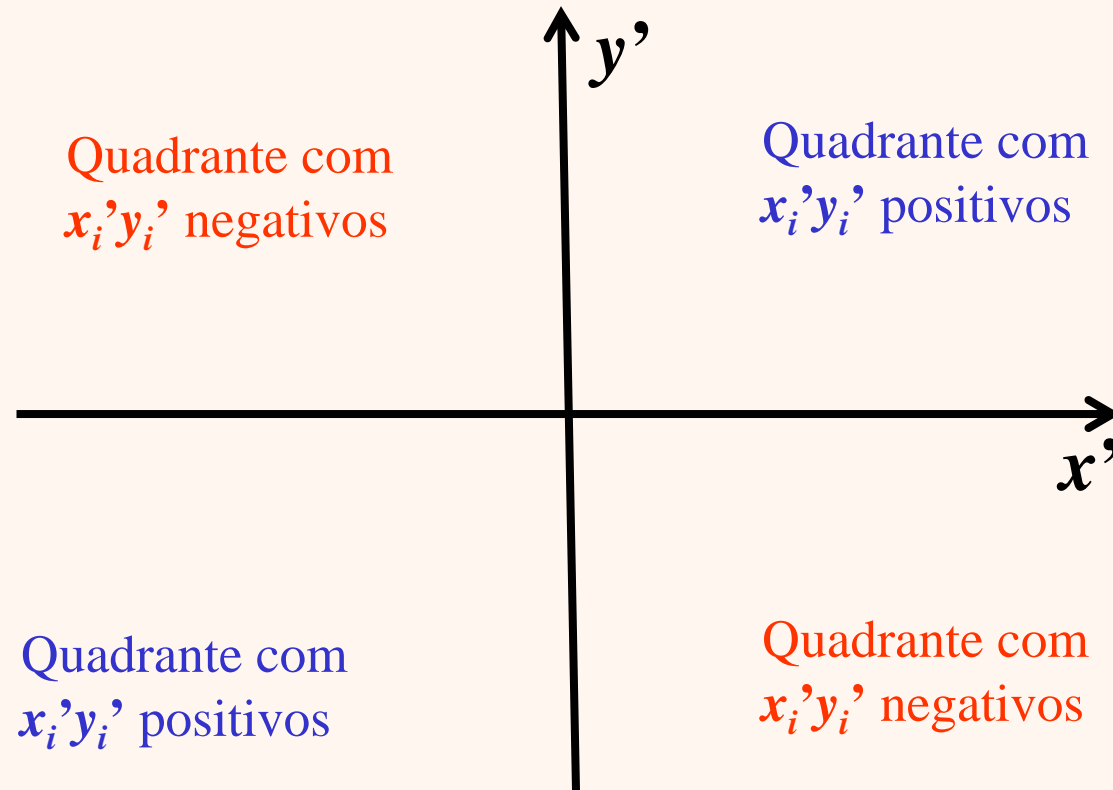
# Idéia de construção do Coef. de Correlação de Pearson

$$x'_i = \frac{x_i - \bar{x}}{s_x} \quad y'_i = \frac{y_i - \bar{y}}{s_y} \quad (i = 1, 2, \dots, n)$$

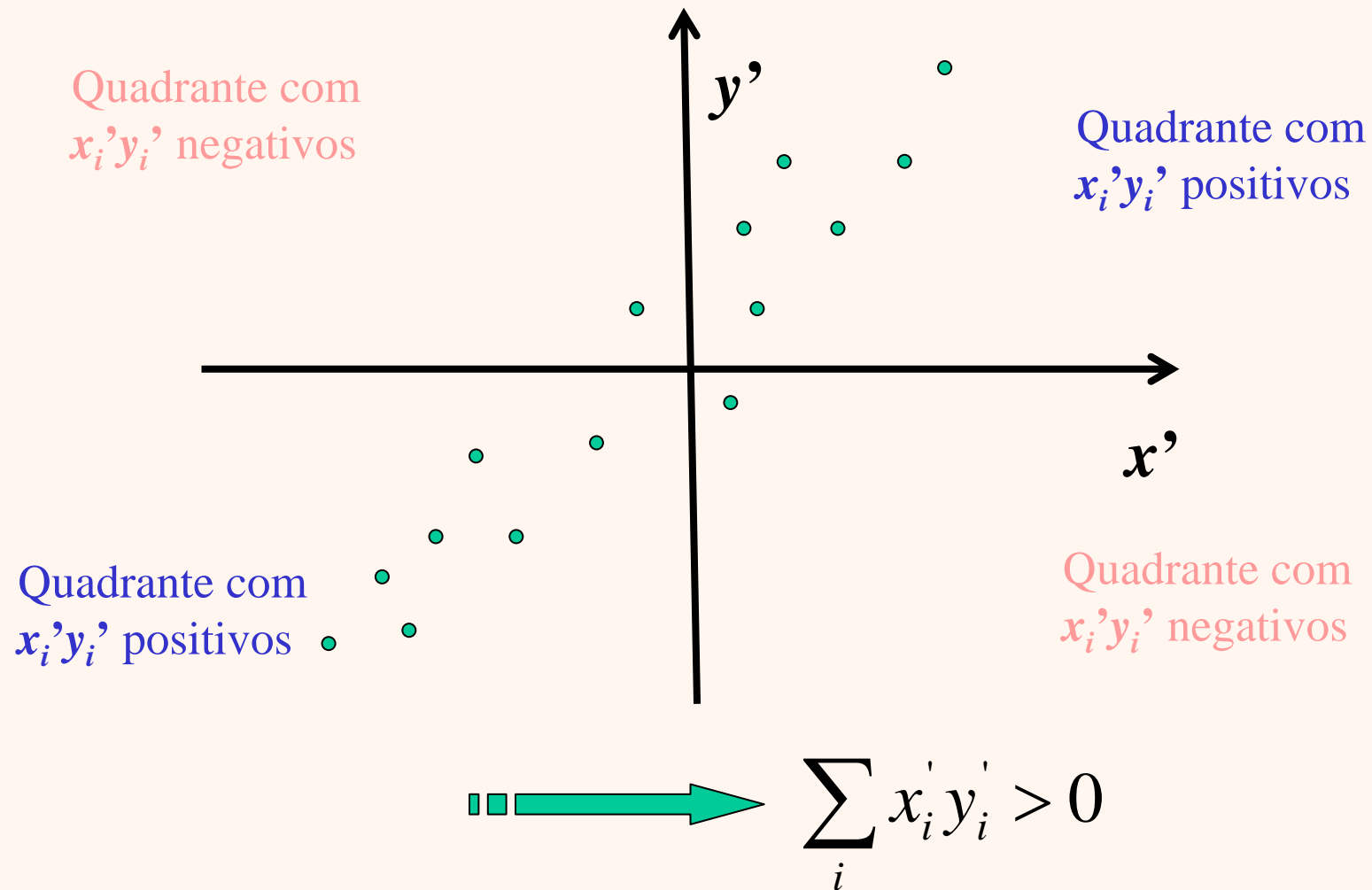
Considere os produtos dos valores padronizados:

$$x'_i y'_i$$

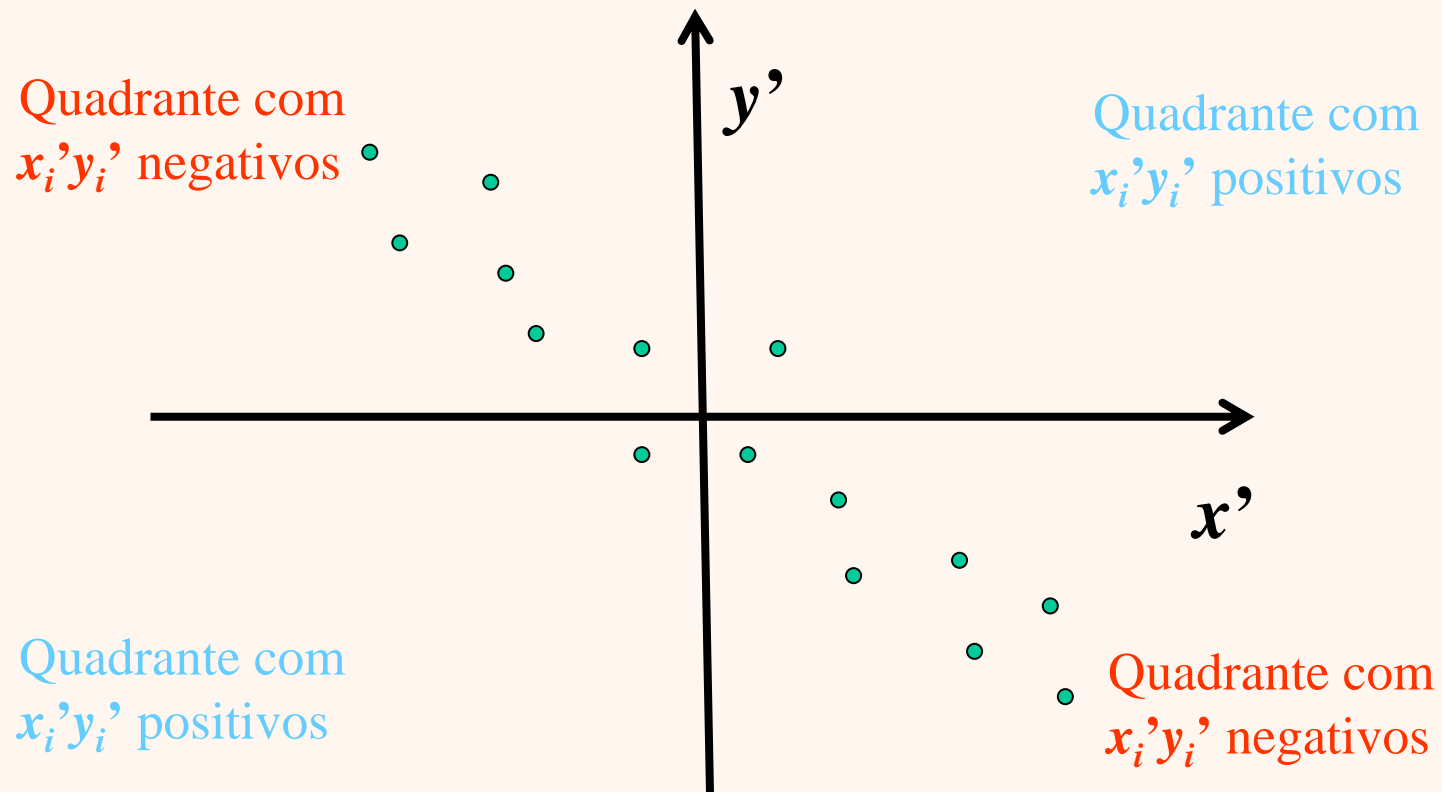
# Sinais dos produtos dos valores padronizados:



# Sinais dos produtos dos valores padronizados:

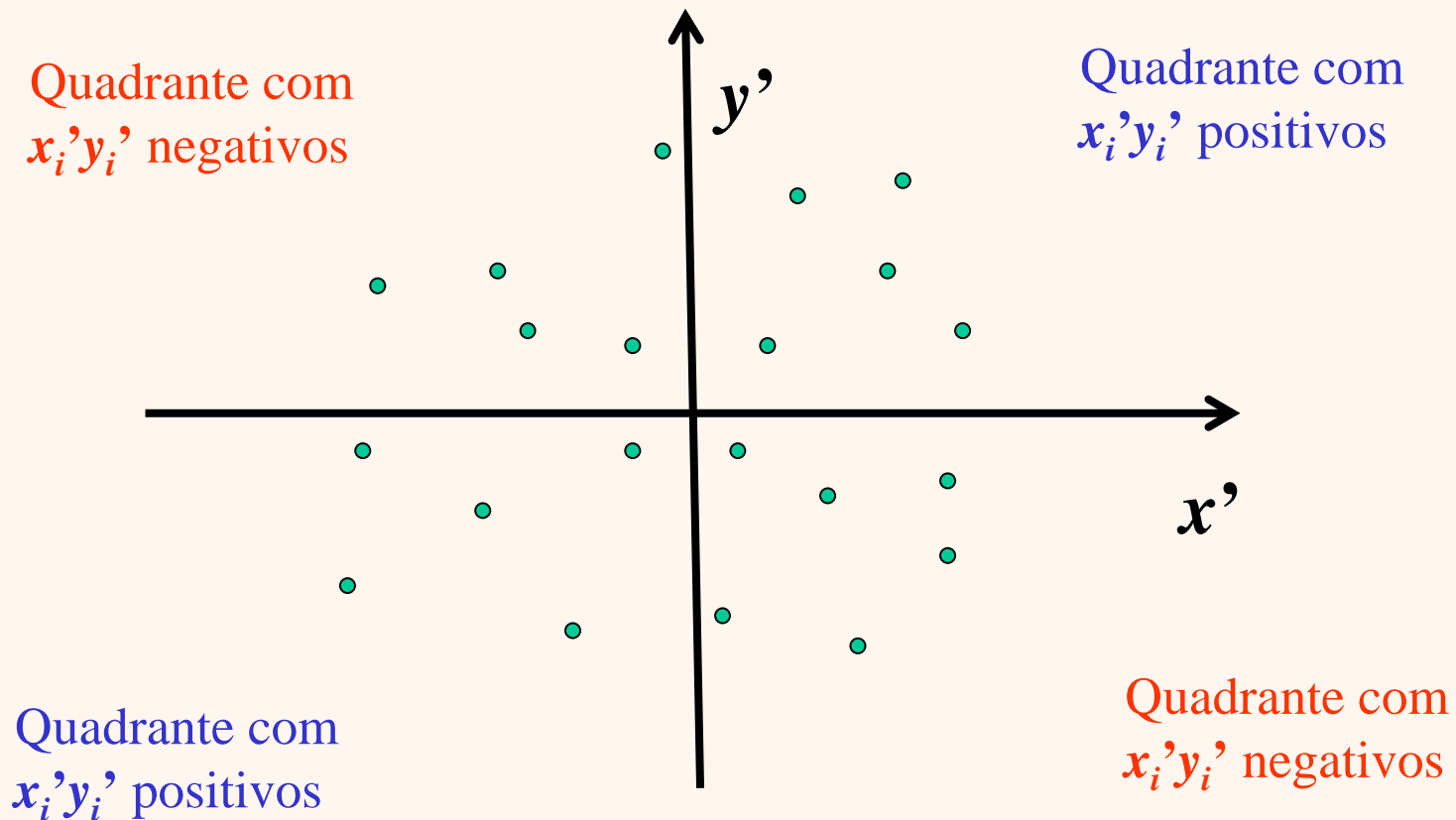


# Sinais dos produtos dos valores padronizados:



⇒  $\sum_i x_i' y_i' < 0$

# Sinais dos produtos dos valores padronizados:



⇒  $\sum_i x'_i y'_i \approx 0$

## Idéia de construção do Coef. de Correlação de Pearson

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

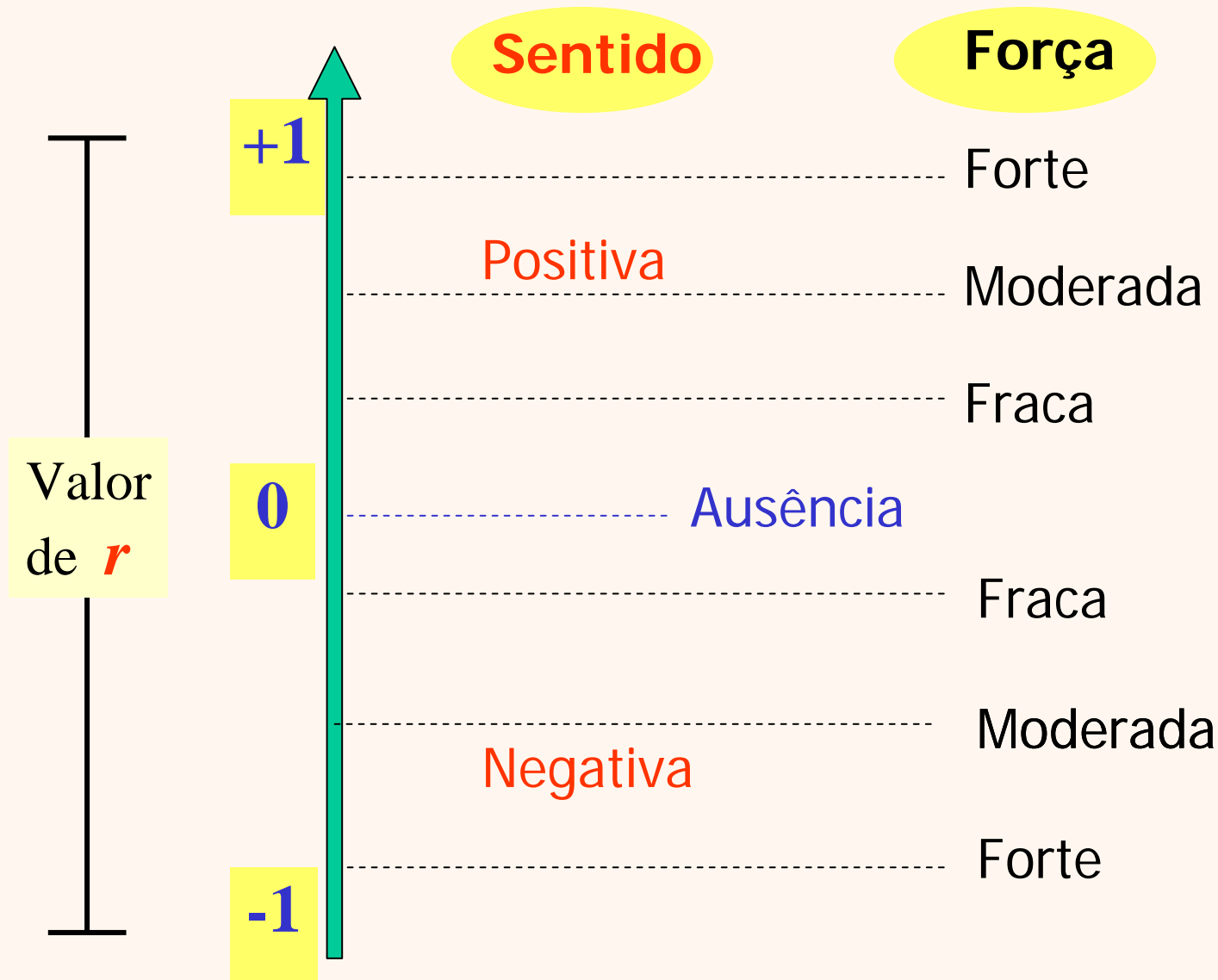
$$x'_i = \frac{x_i - \bar{x}}{s_x} \quad y'_i = \frac{y_i - \bar{y}}{s_y} \quad (i = 1, 2, \dots, n)$$

Coef. de Correlação de Pearson:

$$r = \frac{\sum_{i=1}^n (x'_i y'_i)}{n - 1}$$



# Valores possíveis de $r$ e interpretação da correlação



## Exemplo 11.1. Matriz de correlações

	retração linear	resistência mecânica	absorção de água
retração linear	1,00	0,75	-0,88
resistência mecânica	0,75	1,00	-0,84
absorção de água	-0,88	-0,84	1,00

Interpretar.

## Outra forma de calcular $r$

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

# Coeficiente de correlação populacional

$$\rho = \text{Corr}(X, Y) = E \left\{ \left( \frac{X - \mu_X}{\sigma_X} \right) \cdot \left( \frac{Y - \mu_Y}{\sigma_Y} \right) \right\}$$

$$\mu_X = E(X) \qquad \sigma_X = \sqrt{V(X)}$$

$$\mu_Y = E(Y) \qquad \sigma_Y = \sqrt{V(Y)}$$



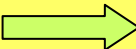
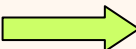

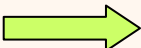
# Inferência sobre $\rho$

- Dada uma amostra aleatória simples  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  do par de variáveis aleatórias  $(X, Y)$ , o coeficiente  $r$  pode ser considerado uma *estimativa* do verdadeiro e desconhecido coeficiente  $\rho$

# Teste de significância de $\rho$

- $H_0: \rho = 0$  (as variáveis  $X$  e  $Y$  são *não correlacionadas*)
- $H_1: \rho \neq 0$  (as variáveis  $X$  e  $Y$  são *correlacionadas*)  
(pode também ser unilateral)
- Admitindo  $(X, Y)$  com distribuição normal bivariada, a Tabela 10 apresenta o valor absoluto mínimo de  $r$  para se rejeitar  $H_0$ .
  - Ver continuação do Exemplo 11.1 no livro.

# Regressão linear simples

Variável independente, <b>X</b>		Variável dependente, <b>Y</b>	
			
Temperatura do forno (°C)		Resistência mecânica da cerâmica (MPa)	
Quantidade de aditivo (%)		Octanagem da gasolina	
Renda (R\$)		Consumo (R\$)	
Memória RAM do computador (Gb)		Tempo de resposta do sistema (s)	
Área construída do imóvel (m²)		Preço do imóvel (R\$)	

## Exemplo 11.2:

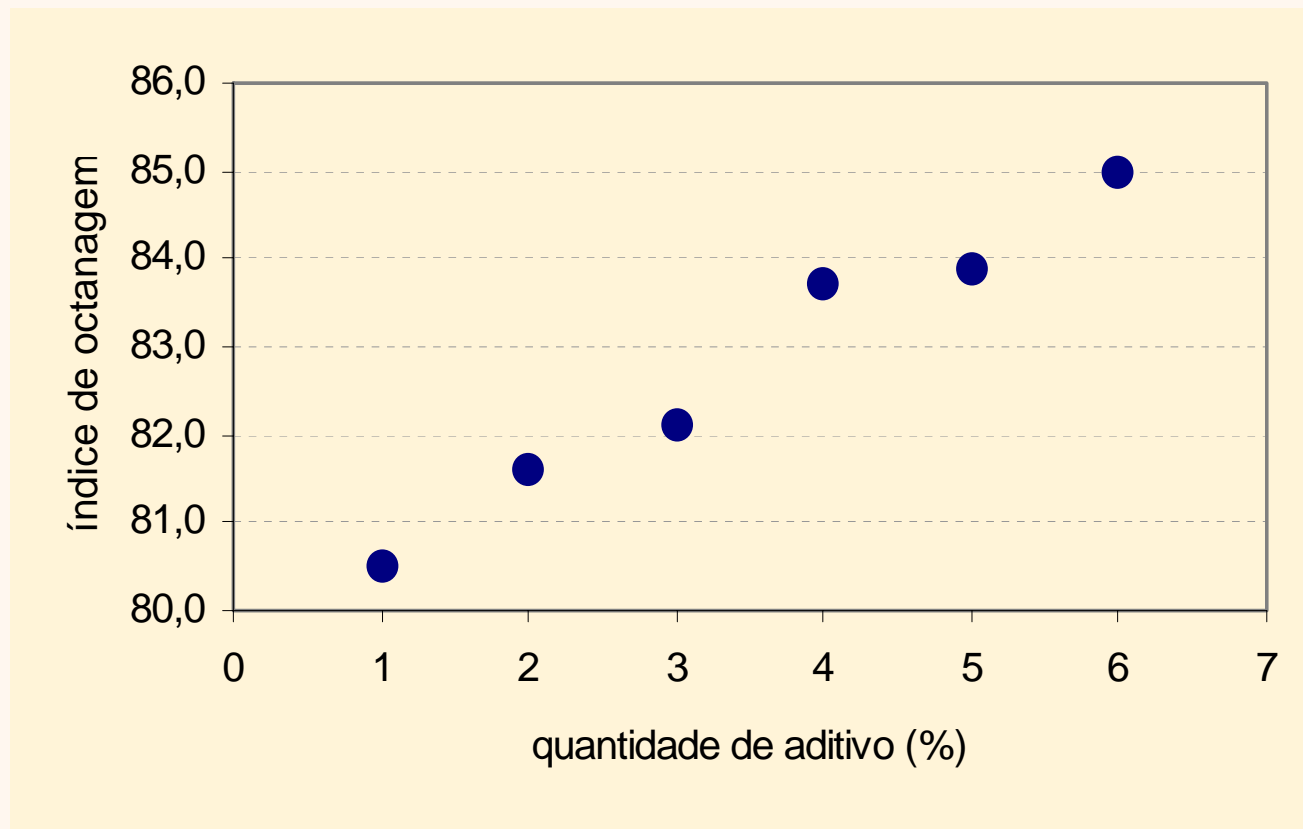
- $X$  = % de aditivo
- $Y$  = Índice de octanagem da gasolina

Resultados de  $n = 6$   
ensaios experimentais:

$X$	$Y$
1	80,5
2	81,6
3	82,1
4	83,7
5	83,9
6	85,0



## Exemplo 11.2:



# Regressão - Modelo

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

$$y_i = \alpha + \beta \cdot x_i + e_i$$

Regressão  
Linear  
Simples

Parâmetros

# Modelo de regressão linear simples

- Em termos das variáveis:  $E\{Y\} = \alpha + \beta X$
- Em termos dos dados:  $Y_i = \alpha + \beta x_i + \varepsilon_i$
- Suposições:
  - os termos de *erro* ( $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ ) são variáveis aleatórias *independentes*;
  - $E\{\varepsilon_i\} = 0$ ;
  - $V\{\varepsilon_i\} = \sigma^2$ ; e
  - $\varepsilon_i$  tem distribuição normal ( $i = 1, 2, \dots, n$ ).

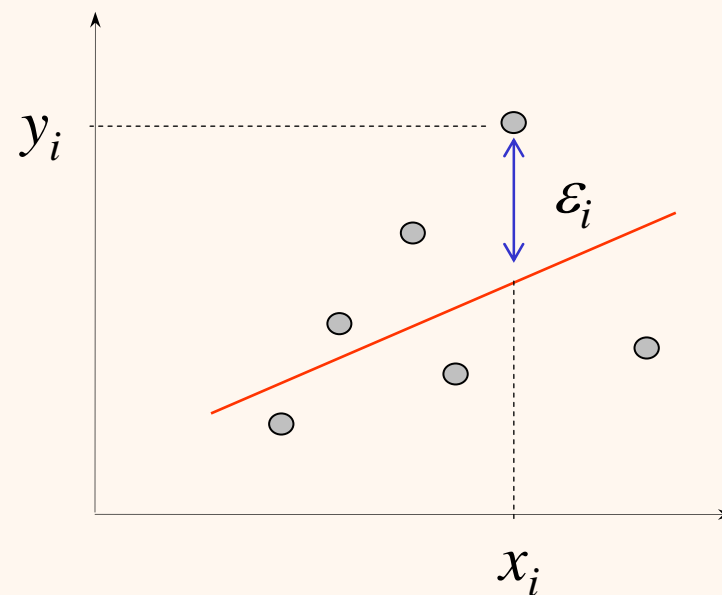
# Método dos mínimos quadrados para estimar $\alpha$ e $\beta$

- Minimizar em relação a  $\alpha$  e  $\beta$  :

$$S = \sum \varepsilon_i^2 = \sum \{Y_i - (\alpha + \beta x_i)\}^2$$

$$\frac{\partial S}{\partial \alpha} = 0$$

$$\frac{\partial S}{\partial \beta} = 0$$



# Método dos mínimos quadrados para estimar $\alpha$ e $\beta$

- Resultado das derivadas parciais:

Estimativa de  $\beta$ : 
$$b = \frac{n \cdot \sum (x_i y_i) - (\sum x_i) \cdot (\sum y_i)}{n \cdot \sum x_i^2 - (\sum x_i)^2}$$

Estimativa de  $\alpha$ : 
$$a = \frac{\sum y_i - b \sum x_i}{n}$$

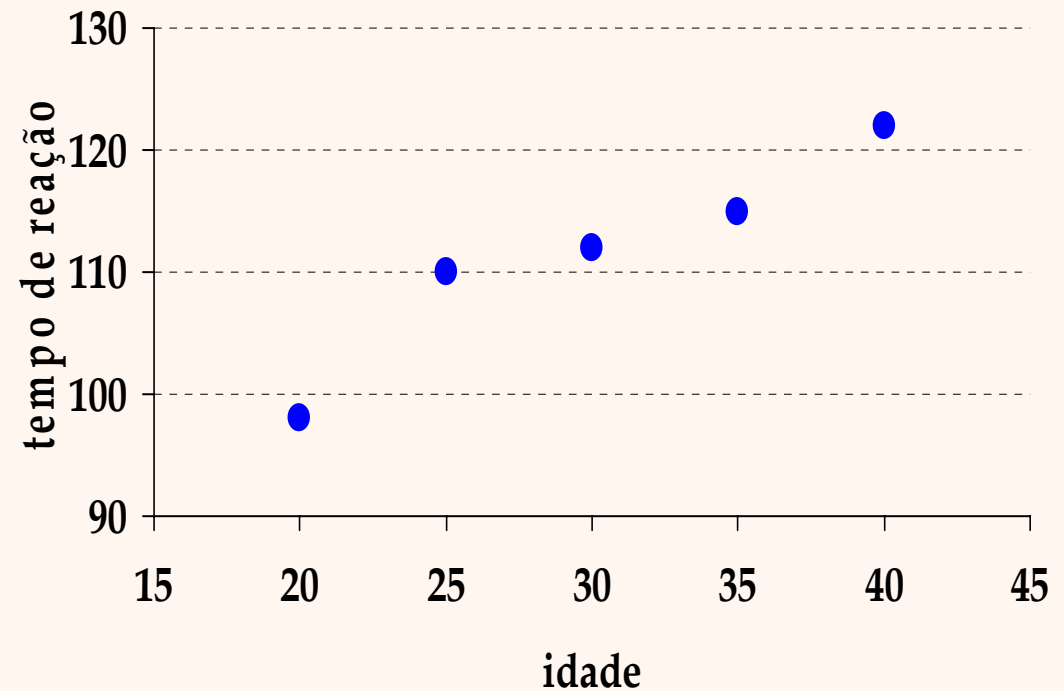
Reta de regressão construída com os dados:

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

# Exemplo numérico

$i$	$x_i$	$y_i$
1	20	98
2	25	110
3	30	112
4	35	115
5	40	122

Diagrama de dispersão



## Exemplo numérico

i	$x_i$	$y_i$	$x_i^2$	$x_i y_i$
1	20	98	400	1960
2	25	110	625	2750
3	30	112	900	3360
4	35	115	1225	4025
5	40	122	1600	4880
$\Sigma$	150	557	4750	16975

reta de regressão:  
 $\hat{y} = a + b.x$

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

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

## Exemplo numérico

$\Sigma x_i$	$\Sigma y_i$	$\Sigma x_i^2$	$\Sigma x_i y_i$
150	557	4750	16975

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

$$b = \frac{5 \cdot (16975) - (150) \cdot (557)}{5 \cdot (4750) - (150)^2}$$

$$\mathbf{b = 1,06}$$



## Exemplo numérico

$\Sigma x_i$	$\Sigma y_i$	$\Sigma x_i^2$	$\Sigma x_i y_i$
150	557	4750	16975

$$b = 1,06$$

$$a = \frac{557 - (1,06).(150)}{5} = 79,6$$

reta de regressão:

$$\hat{y} = a + b.x$$

$$\hat{y} = 79,6 + 1,06x$$

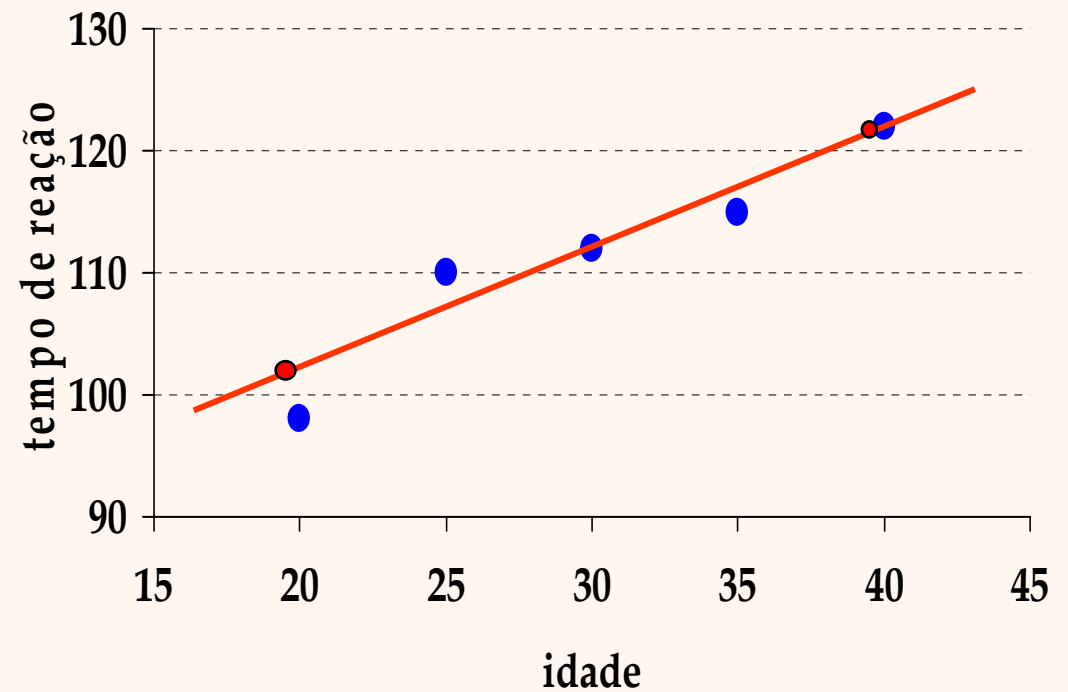
# Exemplo numérico

$$\hat{y} = 79,6 + 1,06x$$

$$x = 20 \Rightarrow \hat{y} = 100,8$$

$$x = 40 \Rightarrow \hat{y} = 122,0$$

Diagrama de dispersão



# Qualidade do ajuste

- Ajustou-se uma equação de regressão entre **X** e **Y**. E a qualidade do ajuste?
  - análise de variância do modelo
  - análise dos resíduos

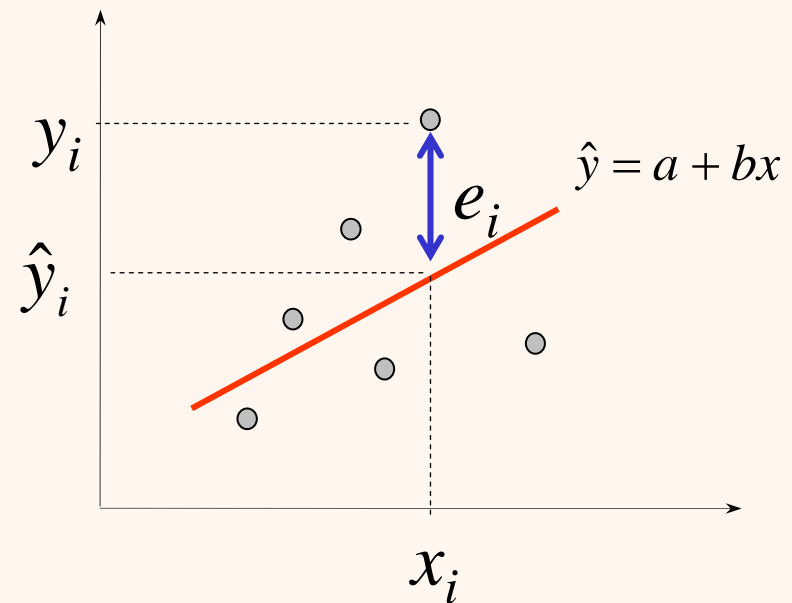
# Reta de regressão e resíduos

- Valores preditos:

$$\hat{y}_i = a + bx_i$$

- Resíduos:

$$e_i = y_i - \hat{y}_i$$



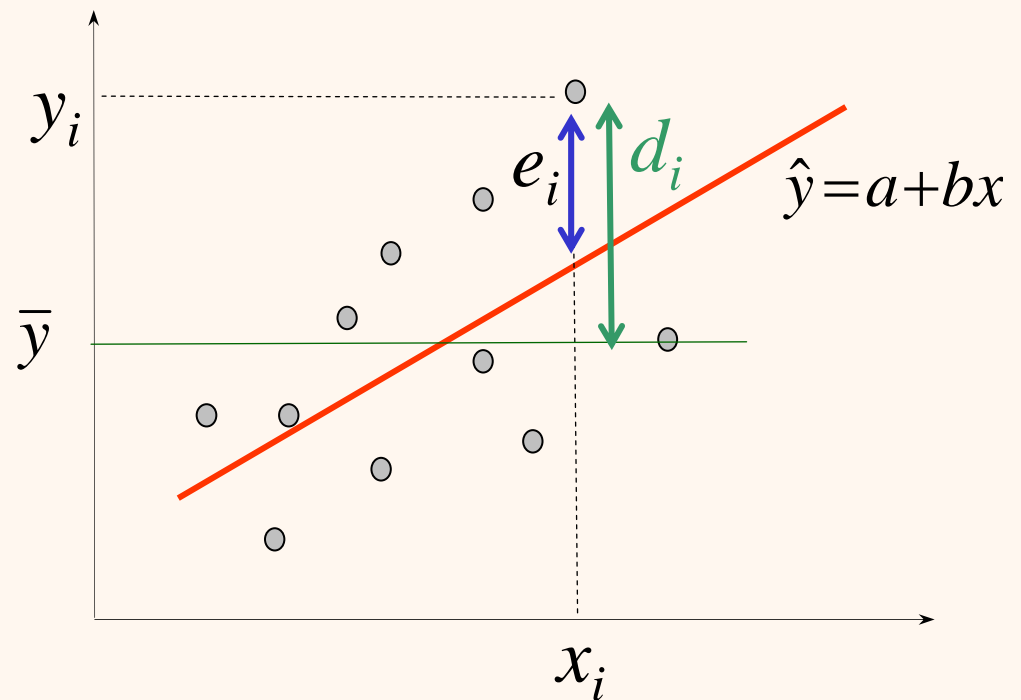
# Análise de variância do modelo

Desvio em relação à  
média aritmética:

$$d_i = y_i - \bar{y}$$

Desvio em relação à  
reta de regressão  
(resíduo da regressão):

$$e_i = y_i - \hat{y}_i$$



# Somas de quadrados

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

SQT

variação total

SQR

variação explicada  
pela equação de  
regressão

SQE

variação não  
explicada

# Somas de quadrados

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

$$SQE = \sum (y_i - \hat{y}_i)^2 = \sum y_i^2 - a \sum y_i - b \sum x_i y_i$$

$$SQR = SQT - SQE$$

Coeficiente de determinação:

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}$$

# Medida da qualidade do ajuste:

Coeficiente de determinação ( $R^2$ )

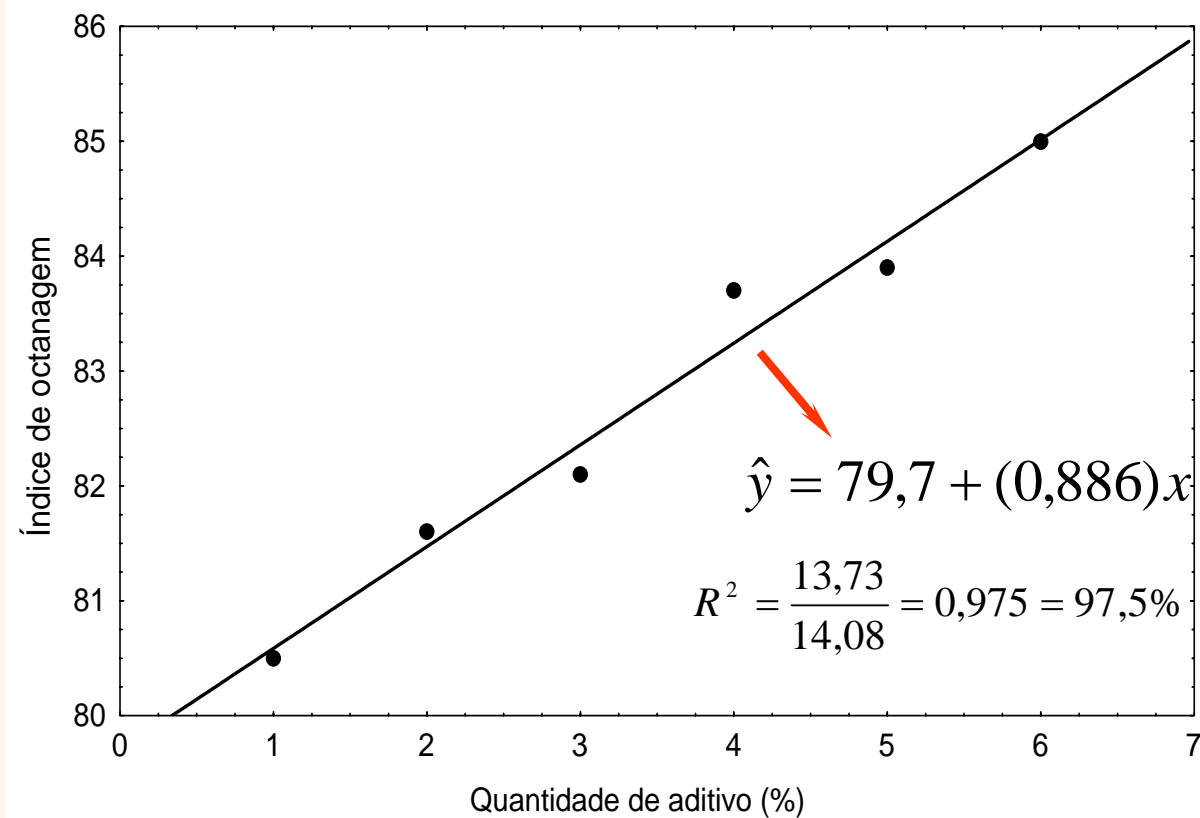
$$R^2 = \frac{\text{Variação explicada}}{\text{Variação total}} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

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

Matematicamente,  $R^2$  é o quadrado do Coef. de Correlação de Pearson.



## Exemplo 11.2:



Interpretar.

# Análise de variância do modelo

Fonte de variação	<i>gl</i>	<i>SQ</i>	<i>QM</i>	Razão <i>f</i>
Regressão	1	$SQR = \sum (\hat{y}_i - \bar{y})^2$	$QMR = SQR / 1$	$f = QMR / QME$
Erro	$n - 2$	$SQE = \sum (y_i - \hat{y}_i)^2$	$QME = SQE / (n - 2)$	
Total	$n - 1$	$SQT = \sum (y_i - \bar{y})^2$		

# Teste de significância do modelo

$$E\{Y\} = \alpha + \beta.X$$

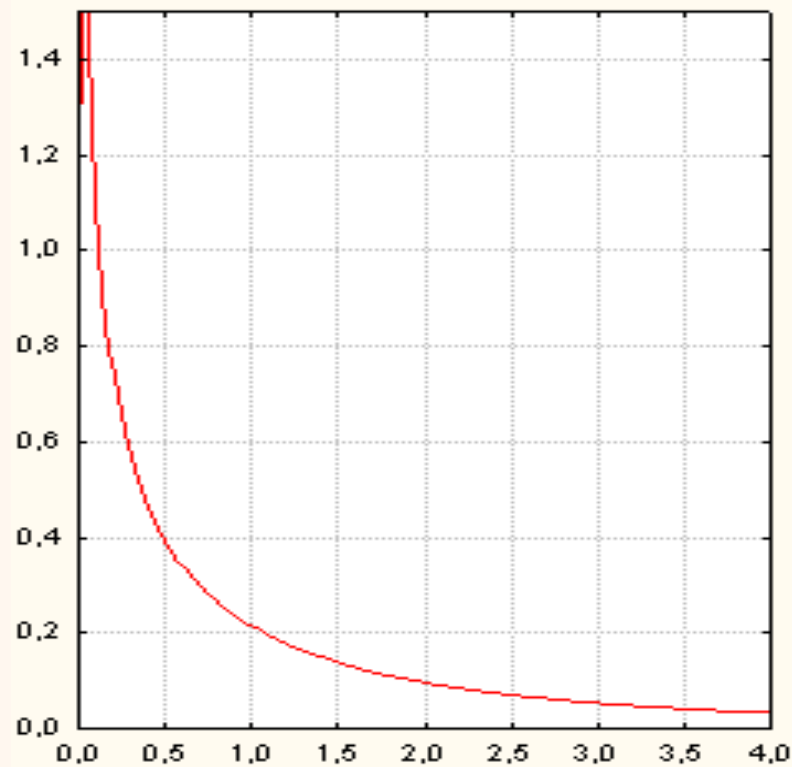
- $H_0: \beta = 0$  e  $H_1: \beta \neq 0$
- Distribuição de referência para a razão  $f$ :  
*distribuição F* com  $gl = 1$  no numerador e  
 $gl = n - 2$  no denominador (Tabela 6).

## Exemplo 11.2:

Fonte de variação	$gl$	$SQ$	$MQ$	Razão $f$
Regressão	1	13,73	13,729	156,26
Erro	4	0,35	0,088	
Total	5	14,08		

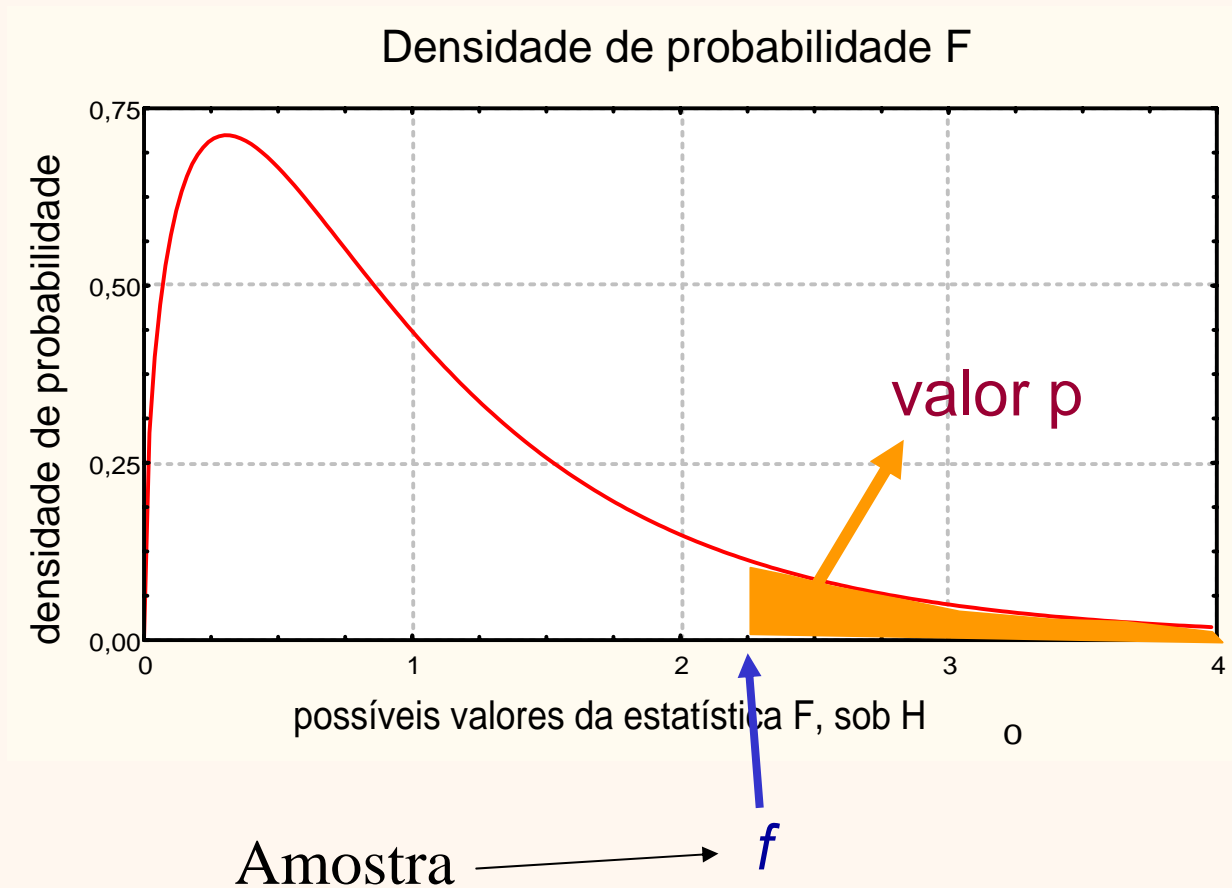
Usar a Tabela 6 e fazer o teste de significância do modelo.

# Distribuição f com $gl = 1$ e 4

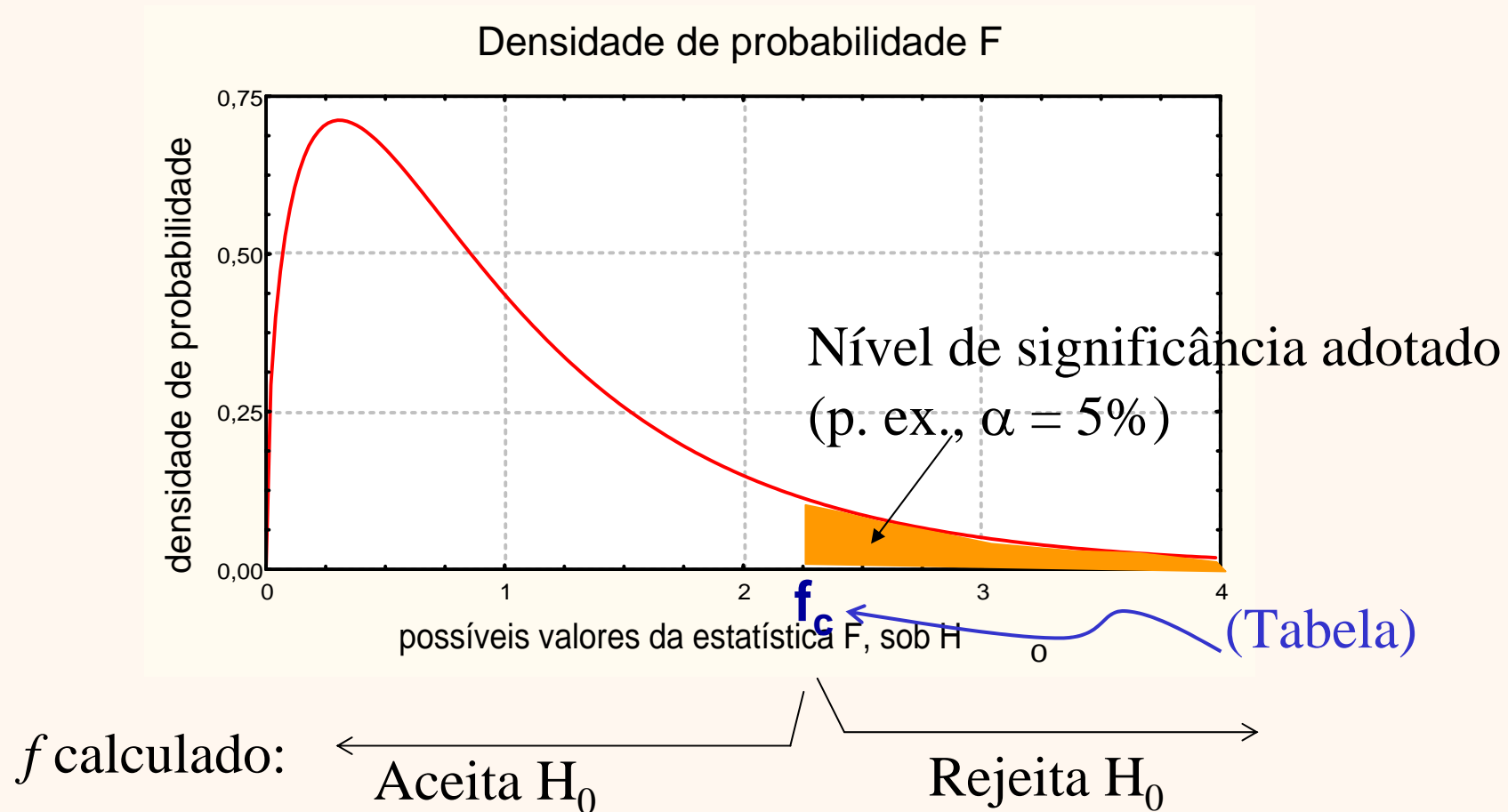


Possíveis valores de  $f$ , sob  $H_0$

# Valor p na distribuição F

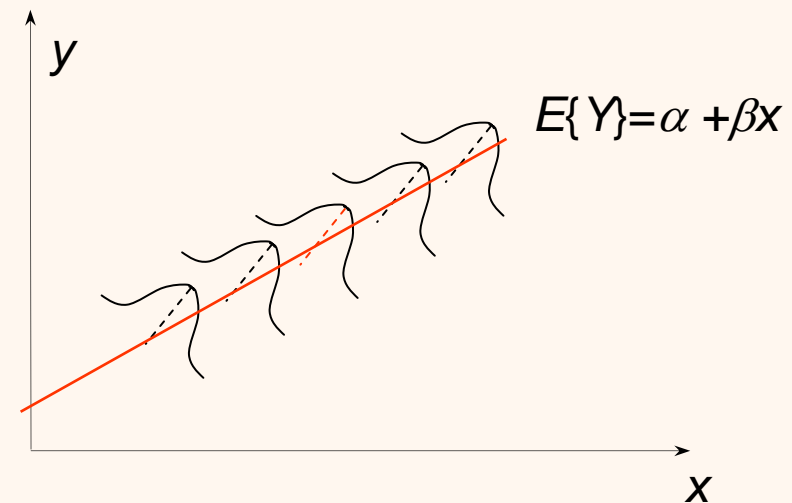


# Abordagem clássica: regra de decisão



# Suposições do modelo

- Modelo:  $Y_i = \alpha + \beta x_i + \varepsilon_i$ 
  - os termos de *erro* ( $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ ) são variáveis aleatórias independentes;
  - $E\{\varepsilon_i\} = 0$ ;
  - $V\{\varepsilon_i\} = \sigma^2$ ; e
  - $\varepsilon_i$  tem distribuição normal ( $i = 1, 2, \dots, n$ ).





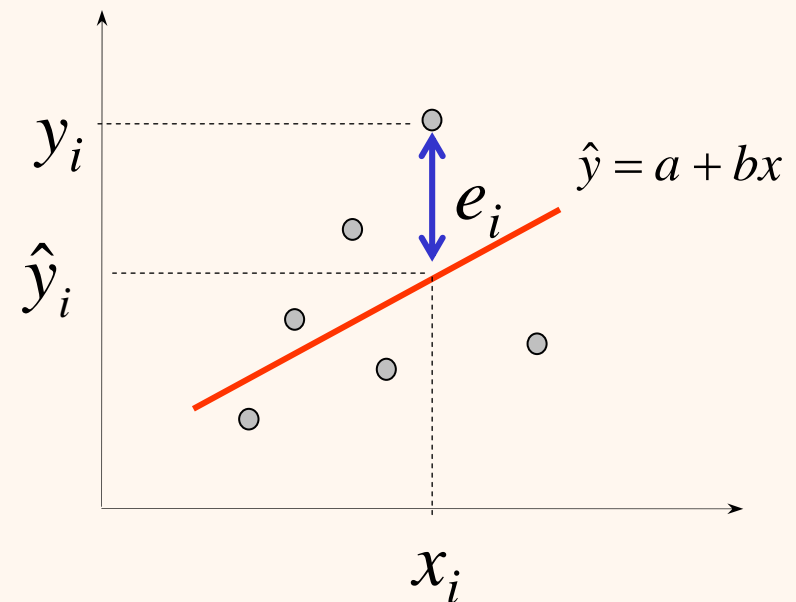
# Análise dos resíduos: um diagnóstico das suposições do modelo

- Valores preditos:

$$\hat{y}_i = a + bx_i$$

- Resíduos:

$$e_i = y_i - \hat{y}_i$$



# Análise dos resíduos

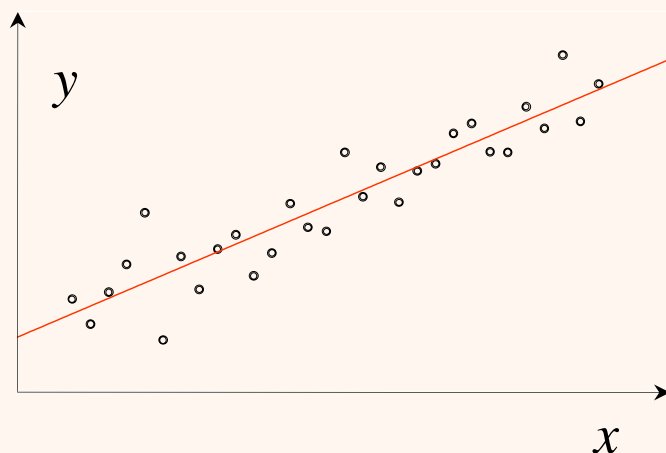


Gráfico dos dados:

$(x_i, y_i)$

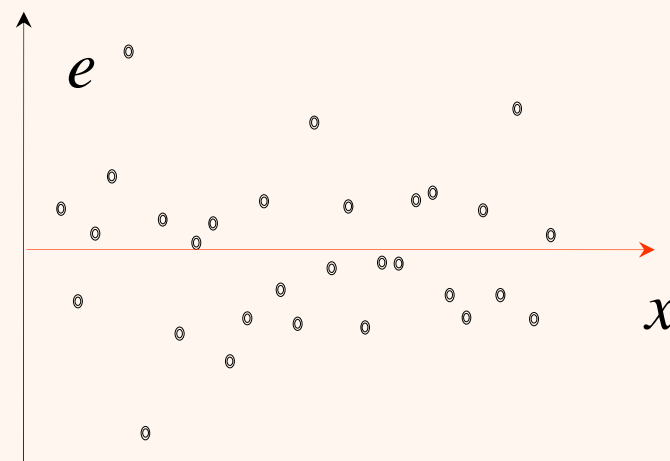


Gráfico dos resíduos:

$(x_i, e_i)$

As suposições do modelo parecem satisfeitas?

# Análise dos resíduos

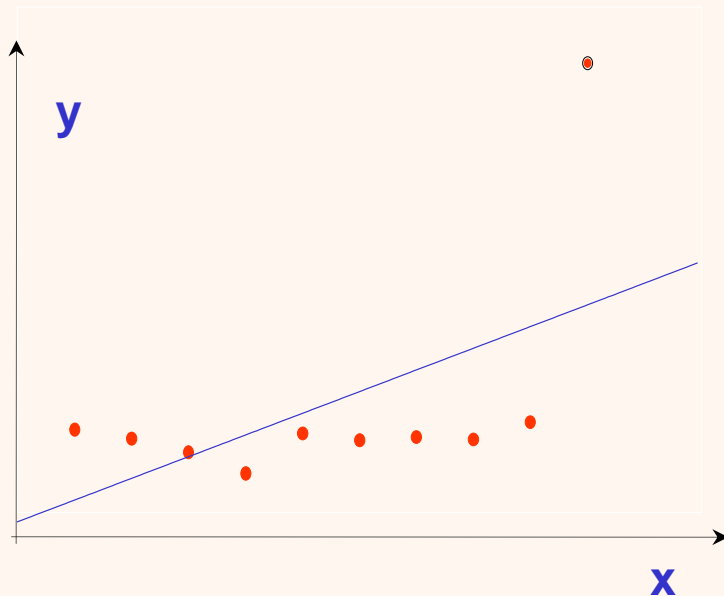


Gráfico dos dados:

$(x_i, y_i)$



Gráfico dos resíduos:

$(x_i, e_i)$

As suposições do modelo parecem satisfeitas?  
O que pode ser feito? (Ver livro)

# Análise dos resíduos

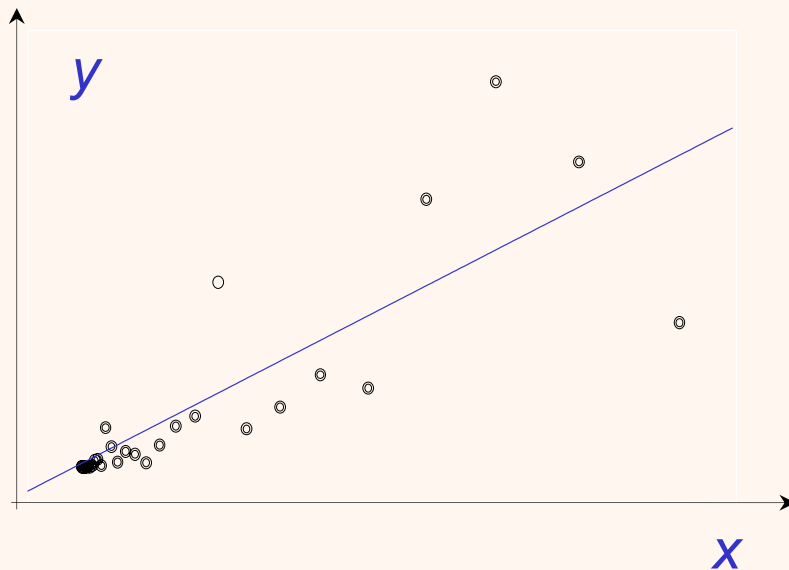


Gráfico dos dados:

$(x_i, y_i)$

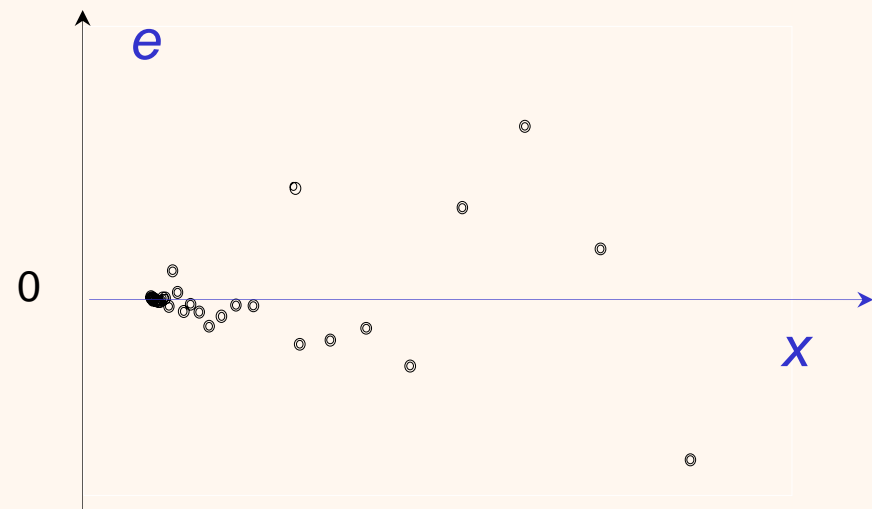


Gráfico dos resíduos:

$(x_i, e_i)$

As suposições do modelo parecem satisfeitas?  
O que pode ser feito? (Ver livro)

# Análise dos resíduos

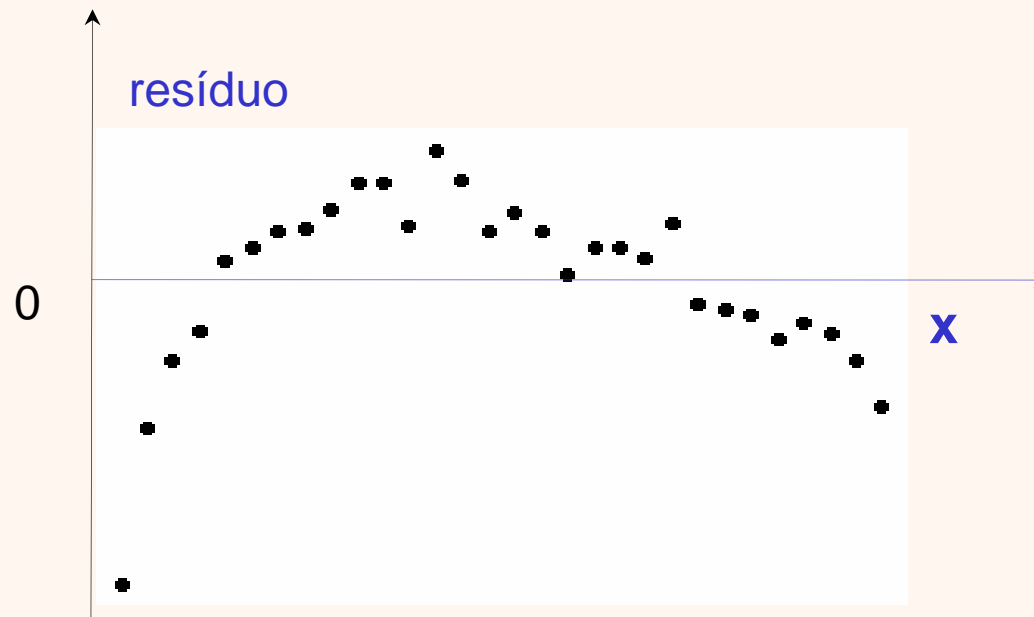


Gráfico dos resíduos:  $(x_i, e_i)$

As suposições do modelo parecem satisfeitas?  
O que pode ser feito? (Ver livro)

# Análise dos resíduos

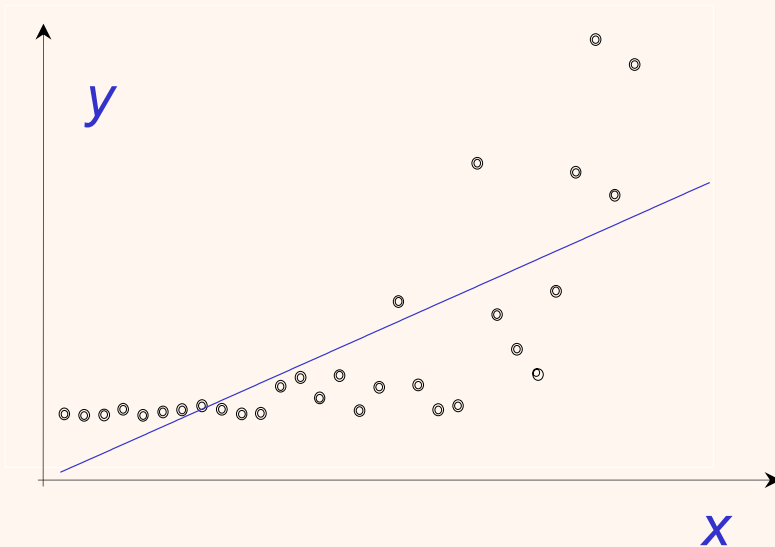


Gráfico dos dados:

$(x_i, y_i)$

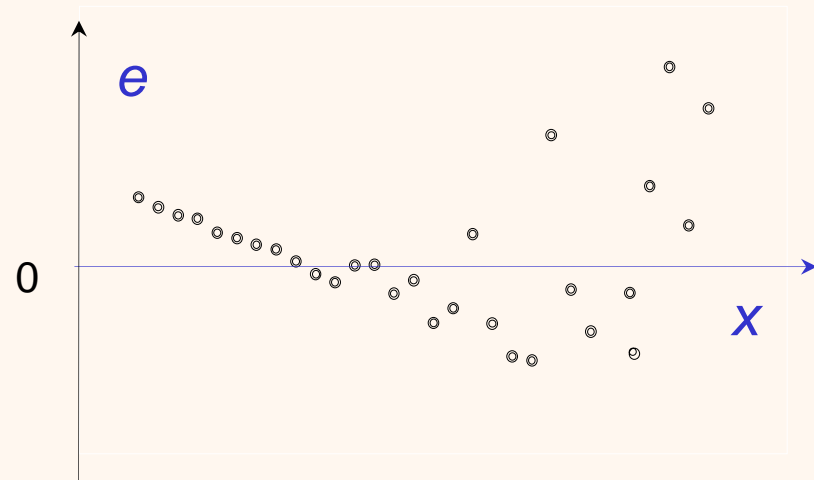


Gráfico dos resíduos:

$(x_i, e_i)$

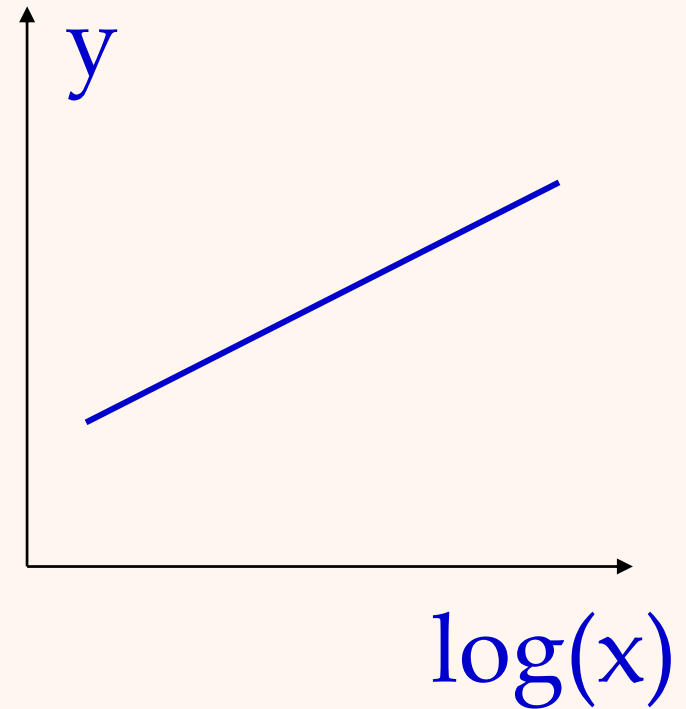
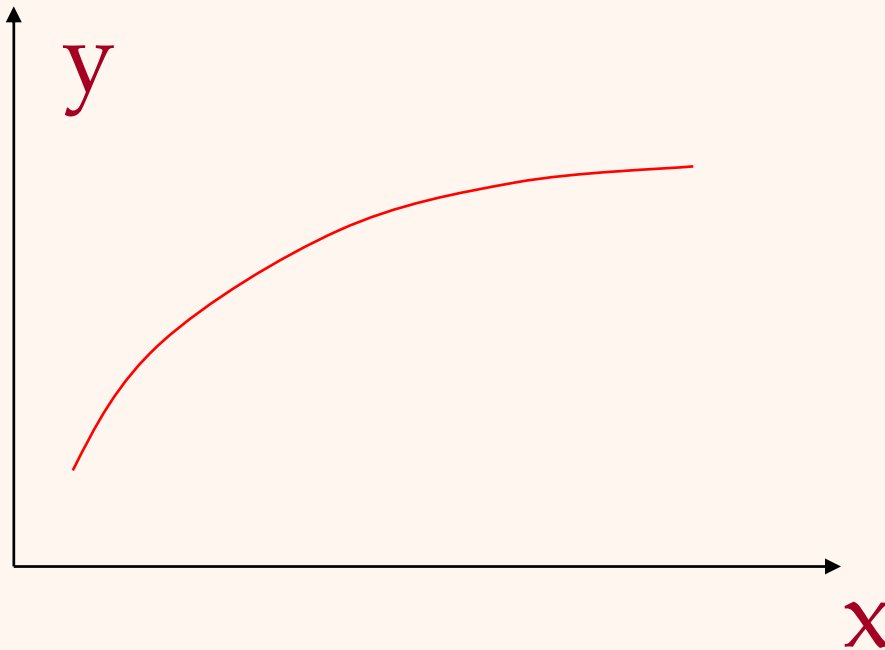
As suposições do modelo parecem satisfeitas?  
O que pode ser feito? (Ver livro)

# Busca de um modelo adequado

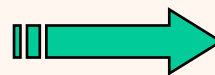
- Suposição de linearidade entre  $x$  e  $y$ : uso de transformações;
- Suposição de variância constante: transformações para estabilizar a variância ou uso do método dos mínimos quadrados generalizados;
- Suposição de independência entre as observações: transformações, uso do método dos mínimos quadrados generalizados ou aplicação de técnicas de séries temporais;
- Suposição de distrib. normal para os erros: uso de transformações.

# Regressão

## Modelos Linearizáveis



$$y = \alpha + \beta \log(X)$$

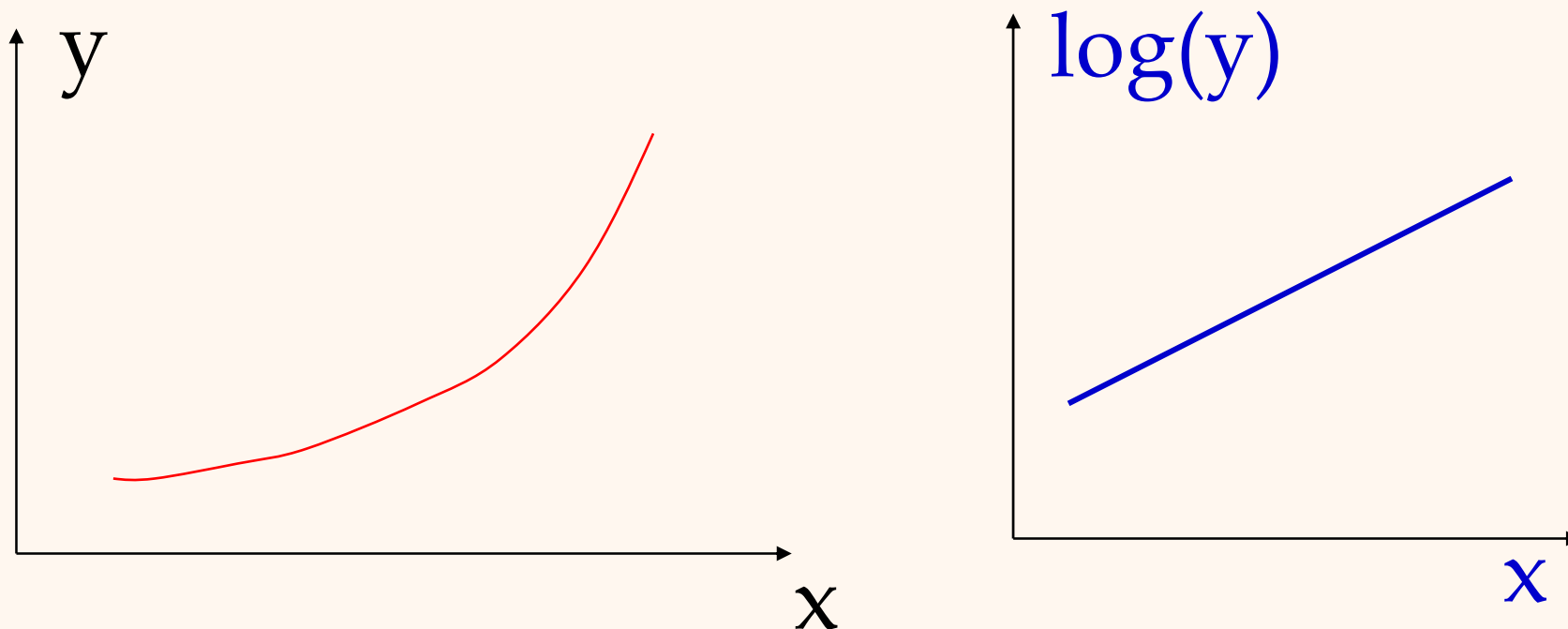


$$y = \alpha + \beta \cdot \log(x)$$

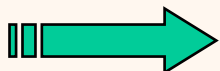


# Regressão

## Modelos Linearizáveis



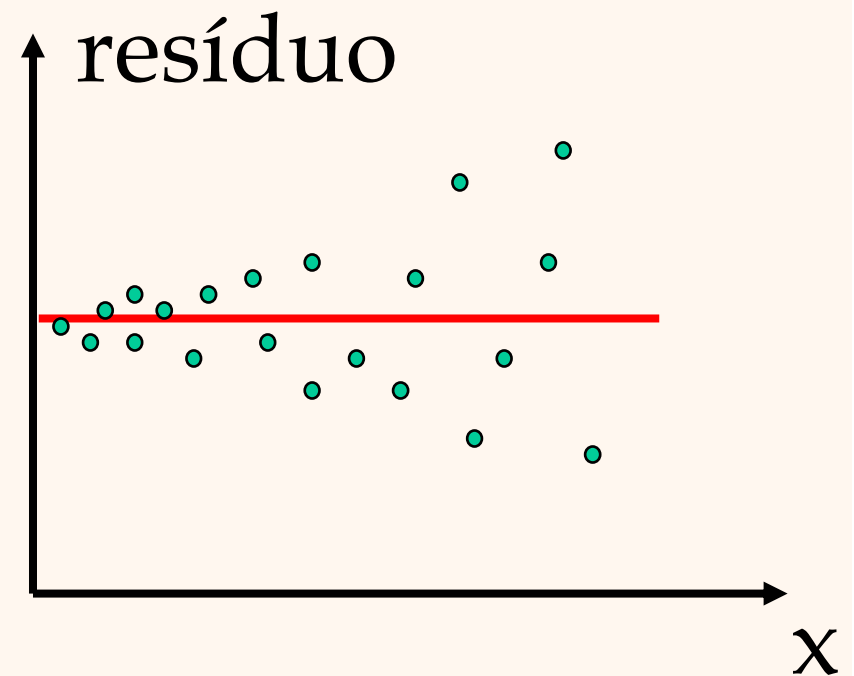
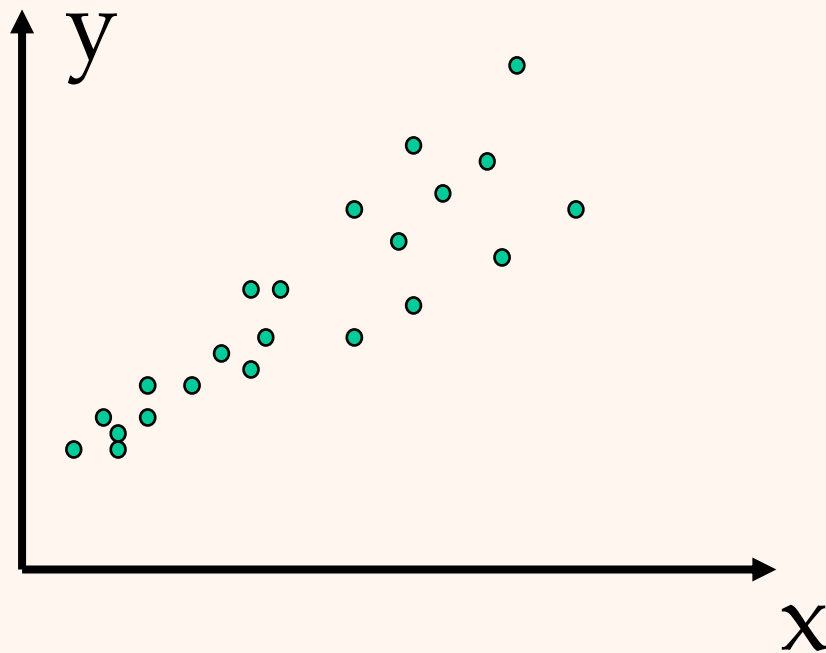
$$y = \alpha \cdot \beta^x$$



$$\log(y) = \log(\alpha) + \log(\beta) \cdot x$$

# Regressão

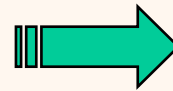
## Transformações para estabilizar a variância



# Regressão

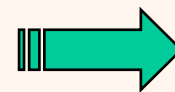
Transformações para estabilizar a variância:  
alguns resultados teóricos

$y$  com distrib. de Poisson



$$y' = \sqrt{y}$$

$y$  com distrib. binomial

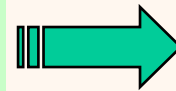


$$y' = \text{sen}^{-1}(\sqrt{y})$$

# Regressão

## Transformações para estabilizar a variância

Se o desvio padrão de  $y$  aumenta proporcionalmente em relação ao valor esperado de  $y$  ( $\sigma_y \approx \mu_y$ )



$$y' = \log(y)$$

