Modelo de Regressão Simples

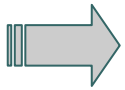
Historia

- História
  - Termo **regressão** foi introduzido por **Francis Galton (1822-1911).** Estudo sobre altura de pais e filhos.
  - **Karl Pearson** coletou mais de mil registros e verificou a "*lei de regressão universal*" de Galton **(1857-1936)**
- Atualmente é uma das técnicas de estimação mais usadas.
  - Aplicações: Industria, Economia, Estudos Biológicos, etc
- Objetivos: descrição de dados, estimação de parâmetros, predição e controle.
- Ampla literatura
  - Modelo de Regressão Linear
  - Modelo de Regressão Não-Linear
  - Modelo Linear Generalizado
  - Entre outros

Exemplos

Aplicação na economia:

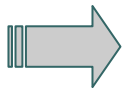
X_1 = renda
X_2 = taxa de juros
X_3 = poupança



Y = consumo

Aplicação no mercado mobiliário (avaliação) :

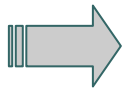
X_1 = área construída
X_2 = custo do m^2
X_3 = localização



Y = preço do imóvel

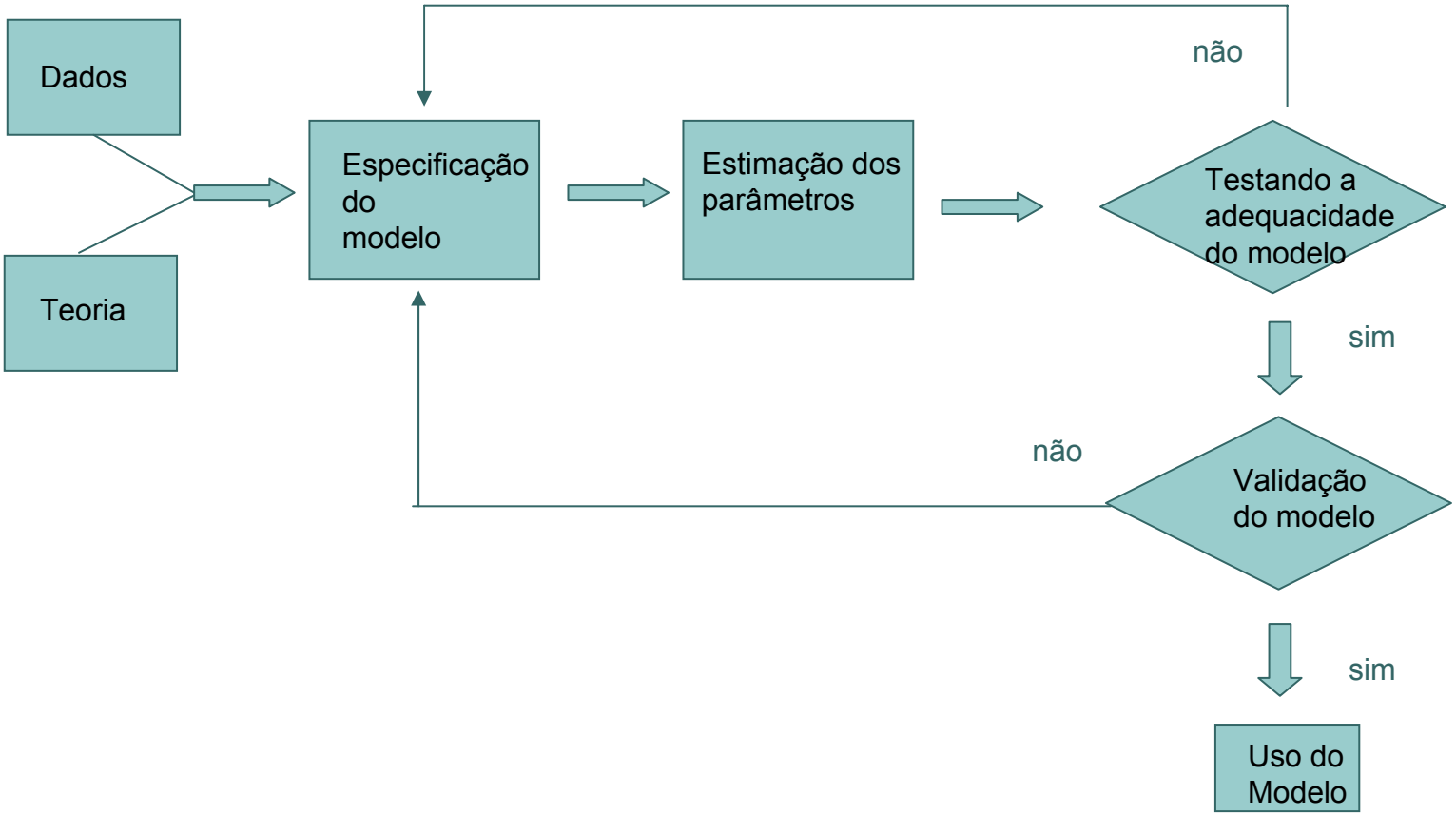
Aplicação na ciência da computação:

X_1 = memória RAM
X_2 = sistema operacional
X_3 = tipo de processador



Y = tempo de resposta

Análise de regressão



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

Relação funcional x Relação estatística

As variáveis podem possuir dois tipos de relações:

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

   Todos os pontos caem na curva da relação funcional

Nesse caso, temos um modelo determinístico.

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

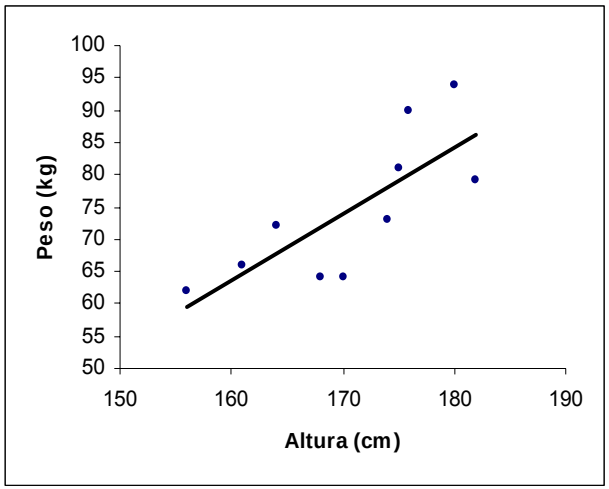
Relação funcional x Relação estatística

Estatística: não é uma relação perfeita como no caso da relação funcional. As observações em geral não caem exatamente na curva da relação.

Nesse caso temos um modelo probabilístico. O modelo captura a aleatoriedade que é parte de um processo do mundo real.

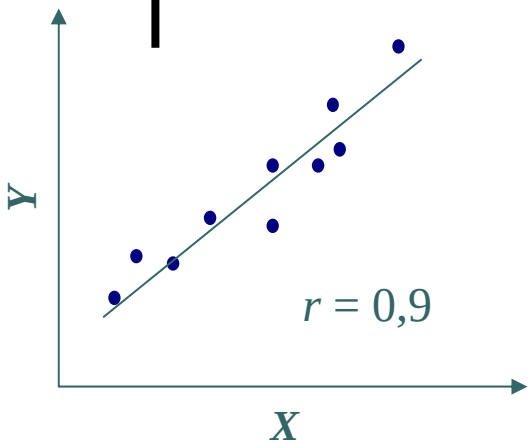
Ex: relação entre tamanho de casa (T) e preço (P). Todas as casas de mesmo tamanho são vendidas pelo mesmo preço?

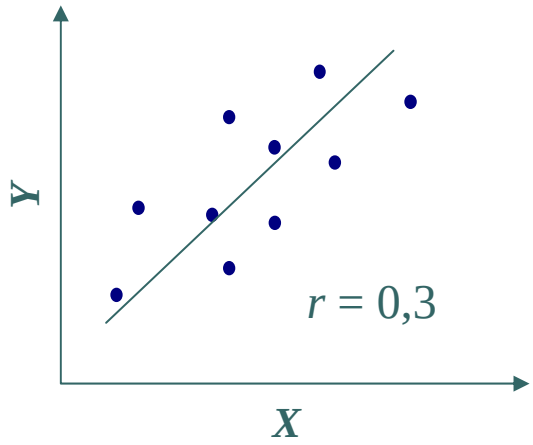
Relação estatística:

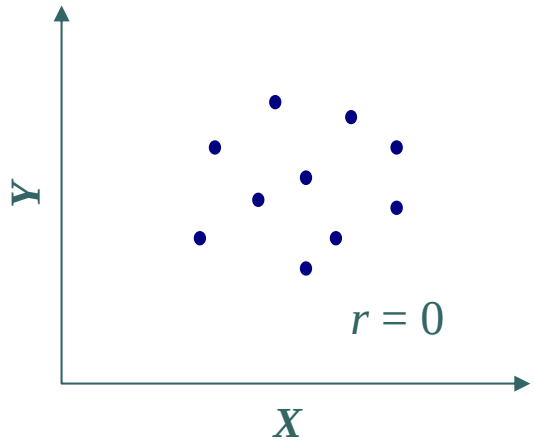


A existência de uma relação estatística entre a variável dependente Y e a variável independente X não implica que Y depende de X, ou que existe uma relação de causa-efeito entre X e Y.

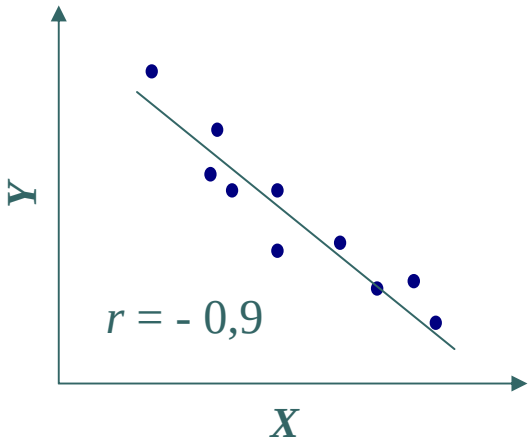
Medida de Associação







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



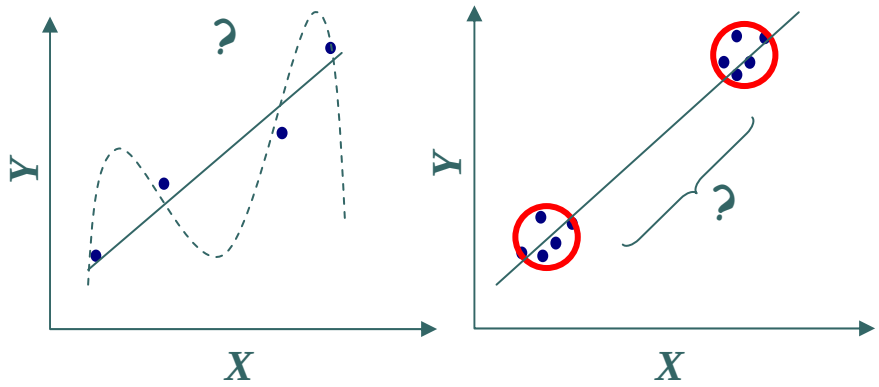
$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)*Var(Y)}} \qquad -1 \le r \le 1$$

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]\left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right]}}$$

Coeficiente de Correlação

Interpretações errôneas dos coeficientes de correlação

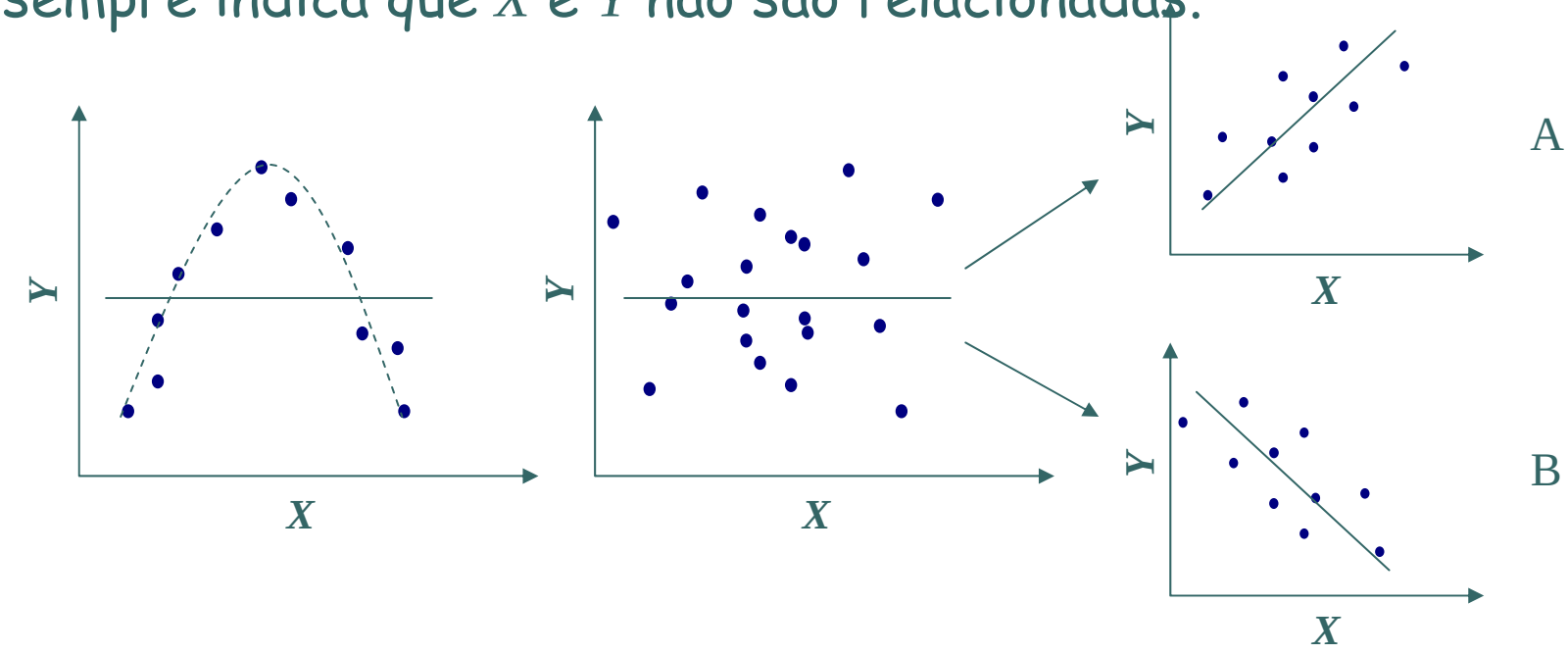
1. Um alto coeficiente de correlação nem sempre indica que a equação de regressão estimada está bem ajustada aos dados.



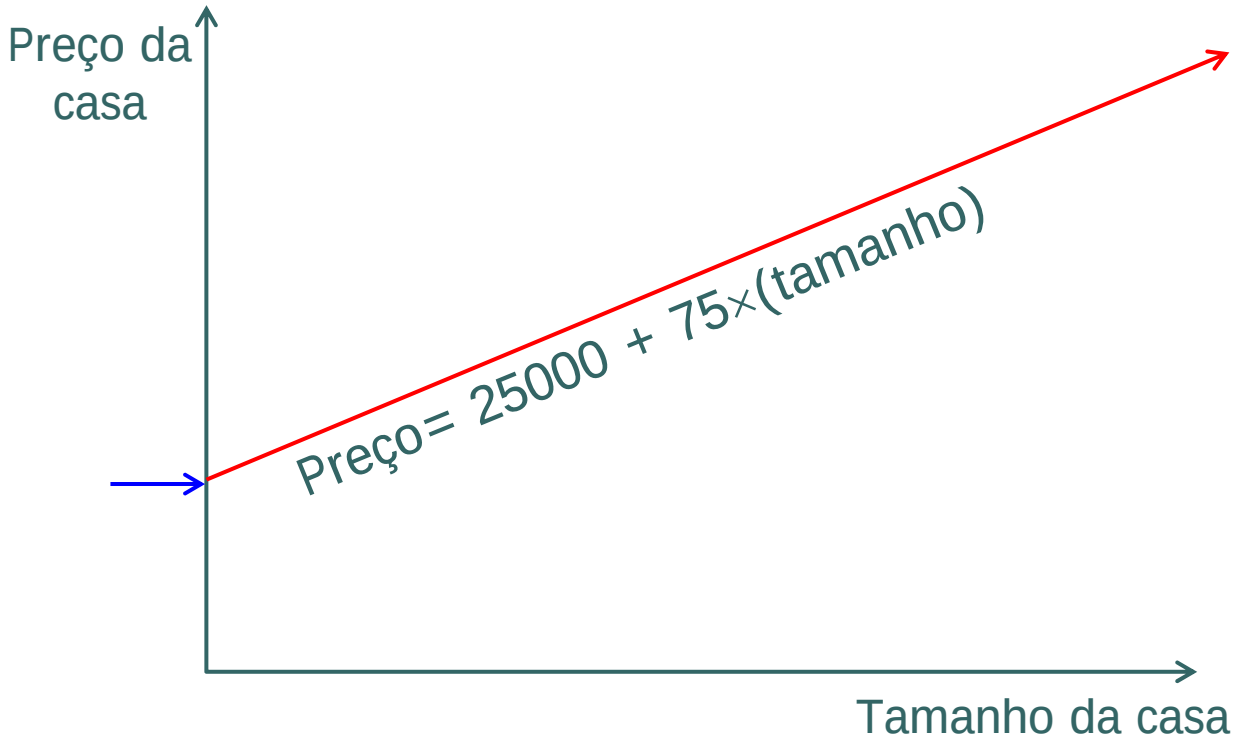
Coeficiente de Correlação

Interpretações errôneas dos coeficientes de correlação

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

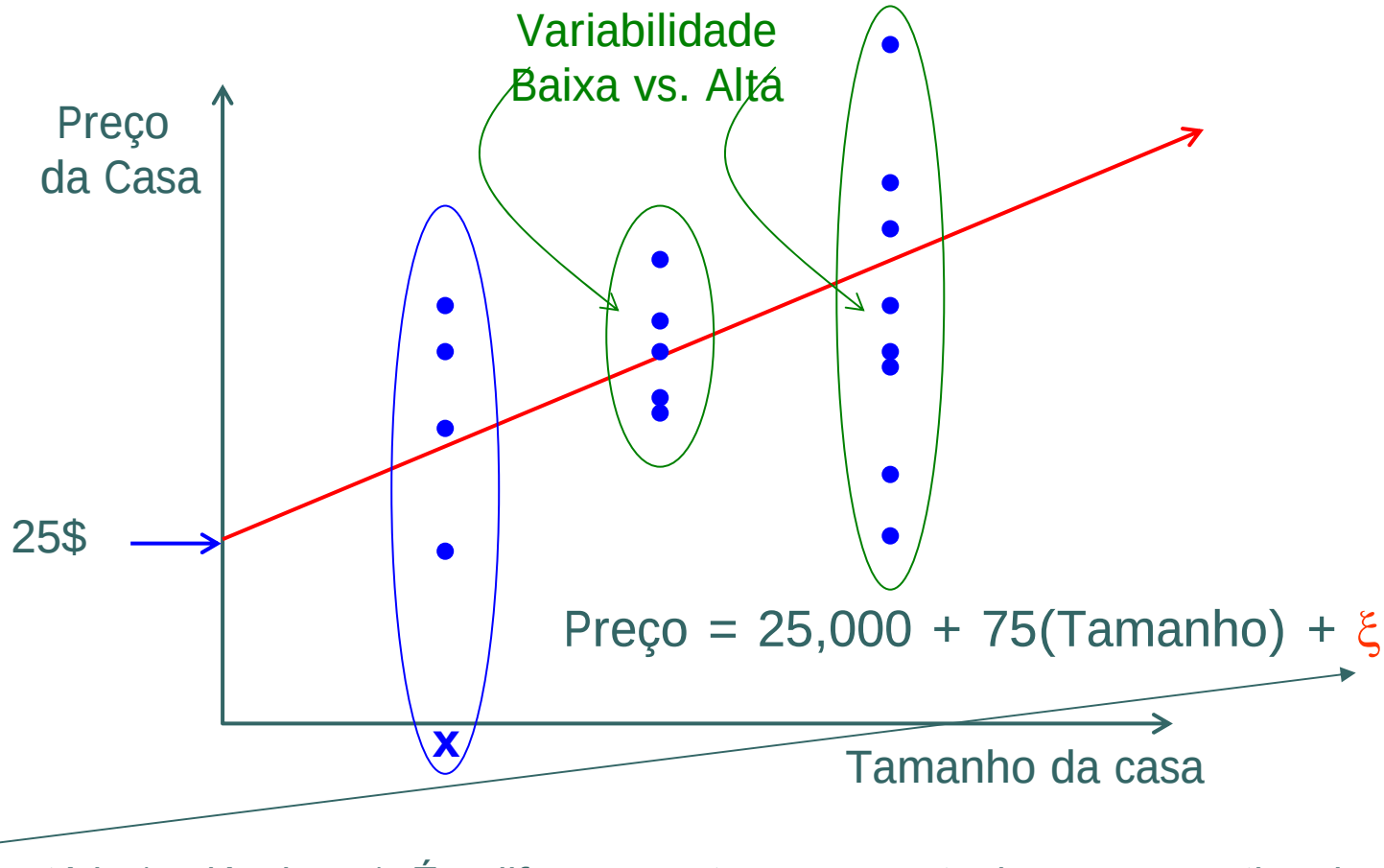


Um modelo determinístico



Neste modelo, o preço da casa é completamente determinado pelo tamanho.

Um modelo estatístico



É o termo aleatório (variável erro). É a diferença entre o preço atual e o preço estimado baseando-se no tamanho da casa.

Análise de Regressão

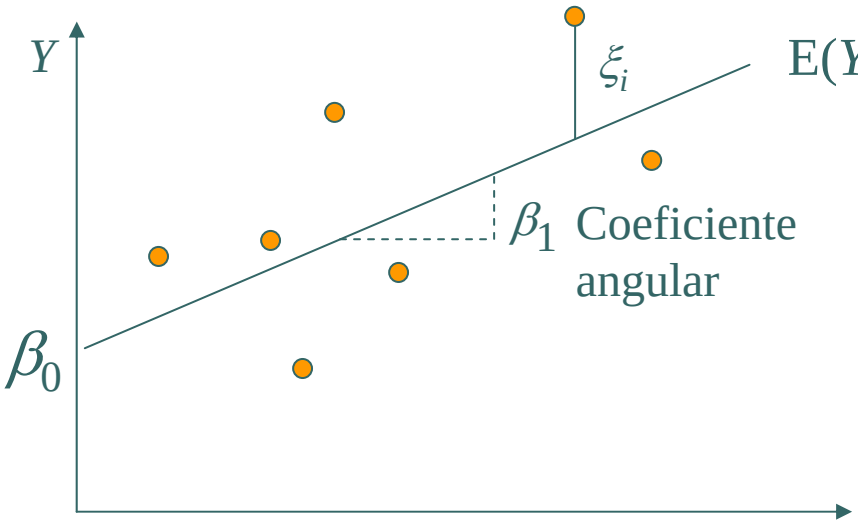
1. Determinar como duas ou mais variáveis se relacionam.
2. Estimar a função que determina a relação entre as variáveis.
3. Usar a equação ajustada para prever valores da variável dependente.

Regressão Linear Simples

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

$$\mathrm{E}(\xi_i) = 0;\ \mathrm{Var}\,(\xi_i) = \sigma^2 \text{ e } \mathrm{COV}\,(\xi_i, \xi_j) = 0$$

Modelo de Regressão Linear Simples



Inclinação populacional

Intercepto populacional

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 e β_1.
- 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:

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

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

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

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

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

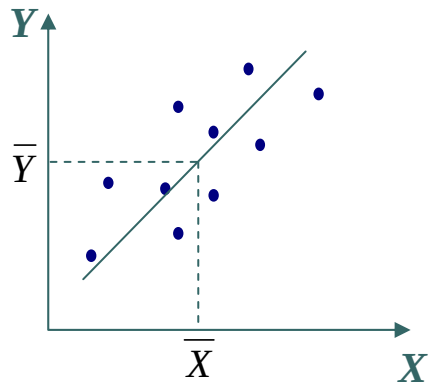
Propriedades da equação de regressão

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 $(\overline{X}, \overline{Y})$



Predição

- Um dos objetivos da análise de regressão
- Para um determinado valor x_0 de X, queremos prever o valor que deverá ser assumido por Y.

$$\hat{y} = \hat{\alpha} + \hat{\beta} x_0$$

Inferência em Análise de Regressão

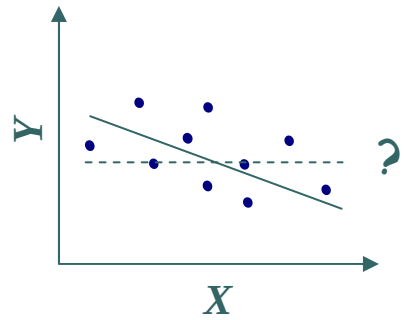
Considere o modelo:

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

$$\xi \sim N(0;\ \sigma^2) \text{ e } \mathrm{COV}(\xi_i, \xi_j) = 0$$

IC para β_0 e β_1, IC para Y_{novo}

$\beta_0 = 0$? $\beta_1 = 0$? (teste de hipótese)

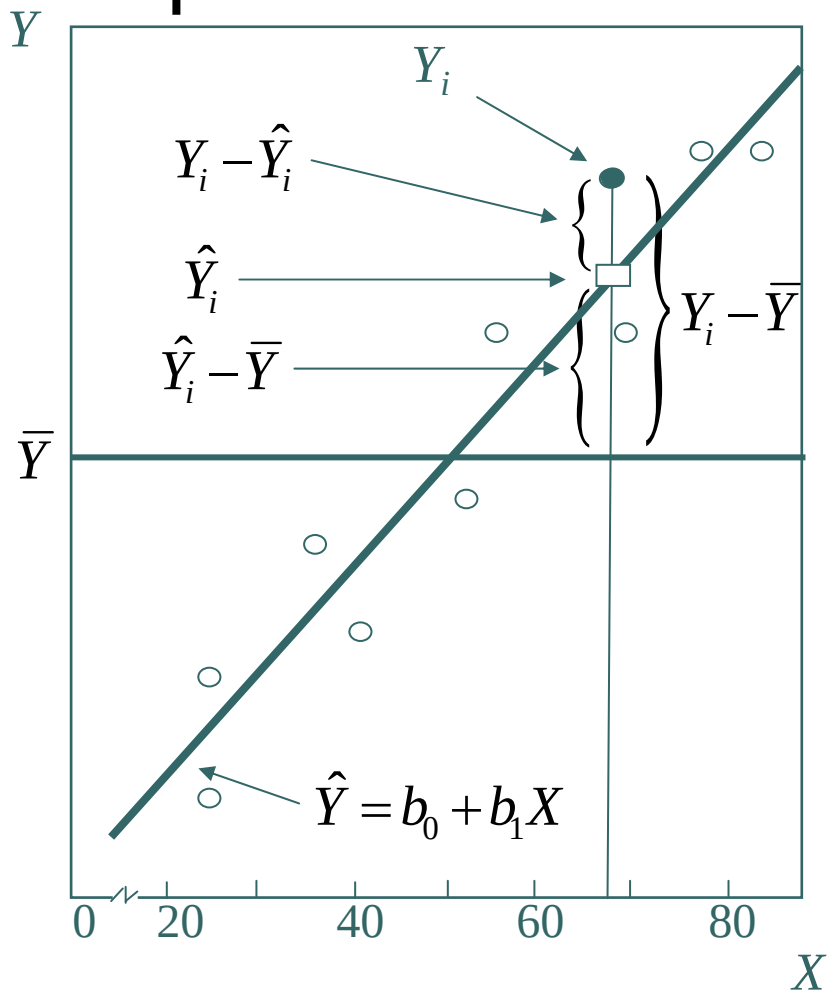


$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 < 0$$

$$t = \frac{b_1 - \beta_1}{s(b_1)} \sim t_{n-2} \qquad t = \frac{b_1}{s(b_1)} \sim t_{n-2} \qquad s^2(b_1) = \frac{\mathrm{QMRes}}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Precisão do modelo



$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

SQTo = SQReg + SQRes

$$R^2 = \frac{\text{SQReg}}{\text{SQTo}}$$

$$= \frac{\text{SQTo - SQRes}}{\text{SQTo}}$$

$$= 1 - \frac{\text{SQRes}}{\text{SQTo}}$$

Coeficiente de determinação

$0 \le R^2 \le 1$

Interpretação: R^2 mede a fração da variação total de Y explicada pela regressão.

Considerações sobre o coeficiente de determinação

- O coeficiente de determinação deve ser usado com cautela.
- Embora o coeficiente não pode diminuir quando mais regressores são adicionados no modelo, isto não significa que o novo modelo é melhor do que o anterior.
- O coeficiente depende do range de variabilidade de x. Um alto valor do coeficiente pode ser porque x teve um grande range de variação não realístico. Por outro lado, um valor pequeno do coeficiente pode ser porque x teve um pequeno range de variação que não permitiu que a sua relação com y seja detectada..
- A média dos quadrados dos resíduos é uma medida adequada de qualidade do ajuste.

Análise de variância: teste de significância do modelo

- SQT tem n-1 graus de liberdade
- SQR tem n-2 graus de liberdade
- SQM tem 1 grau de liberdade

H_0: $\beta_1=0$

	Soma de quadrados	Graus de liberdade	Média	F_0
Regressão	SQM	1	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	$\left[\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2\right]/\sigma^2$ / 1
Residual	SQR	n-2	$\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$	$\left[\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/\sigma^2}{n-2}\right]$
Total	SQT	n-1	$\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}$	

If H_0 é verdadeira
SQR/n-2 tem distribuição qui-quadrado com n-2 graus de liberdade.
SQM/1 tem distribuição qui-quadrado com 1 grau de liberdade.
SQR e SQM são independentes. Por definição, F_0 segue uma distribuição F-Snedecor com 1 e n-2 graus de liberdade. Rejeita H_0 $F_0 > F_{1,n2}$

Considerações

- Os modelos de regressão são construídos baseando-se no range de valores dos regressores.
- A equação dos mínimos quadrados é fortemente afetada por pontos extremos da distribuição de x.
- Os métodos de mínimos quadrados são influenciados por outliers (pontos aberrantes).
- Porque a regressão indicou forte correlação entre duas variáveis não significa que exista uma relação de causa e efeito.

Modelos Linearizáveis

Modelo Padrão: $Y_i = \beta_0 + \beta_1 X_i + \xi_i$

exponencial

$$Y_i = \beta_0 e^{\beta_1 X_i} \xi_i \qquad \ln Y_i = \ln \beta_0 + \beta_1 X_i + \ln \xi_i \qquad Y_i' = \beta_0' + \beta_1 X_i + \xi_i'$$

potencial

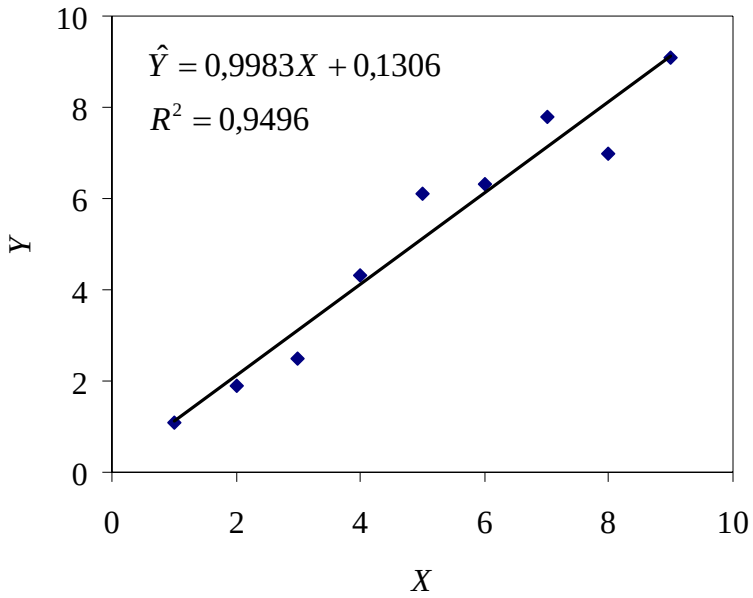
$$Y_i = \beta_0 X_i^{\beta_i} \xi_i \qquad \ln Y_i = \ln \beta_0 + \beta_1 \ln X_i + \ln \xi_i \qquad Y_i' = \beta_0' + \beta_1 X_i' + \xi_i'$$

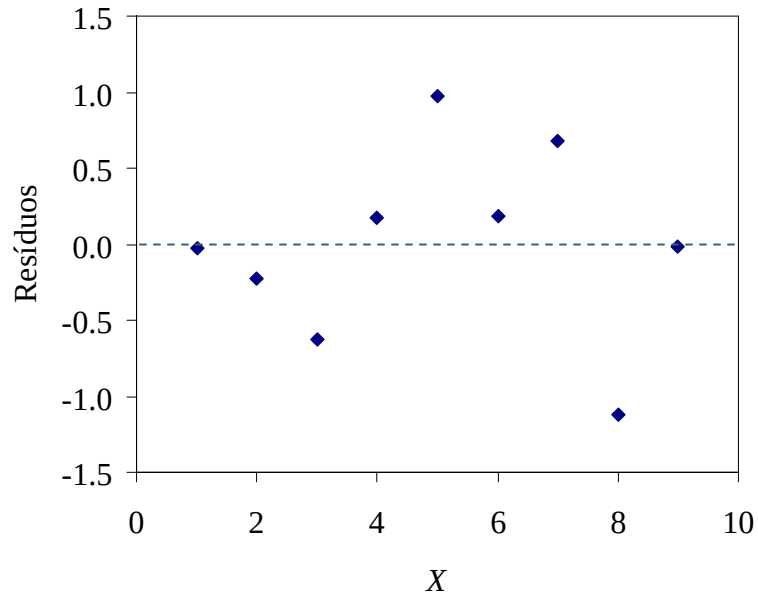
$$\xi_i' \sim N(0, \sigma^2)$$

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

- logaritmo
- potência
- inverso

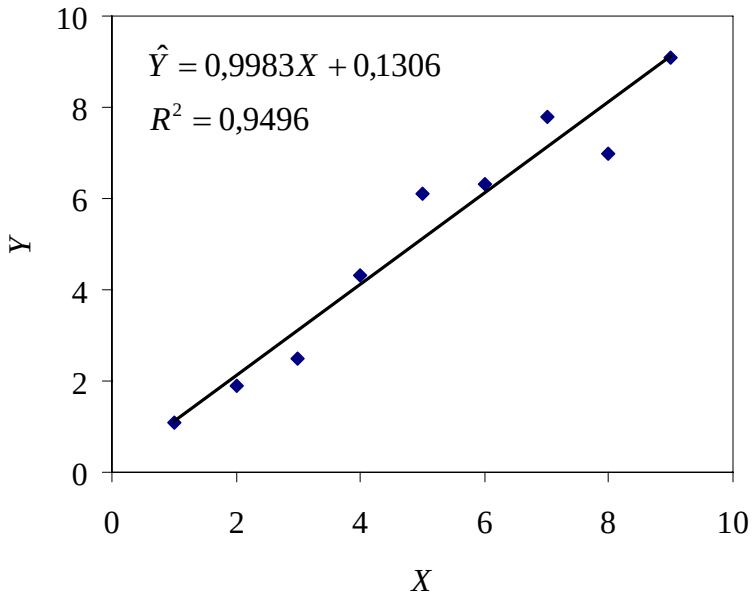
Análise de Resíduos

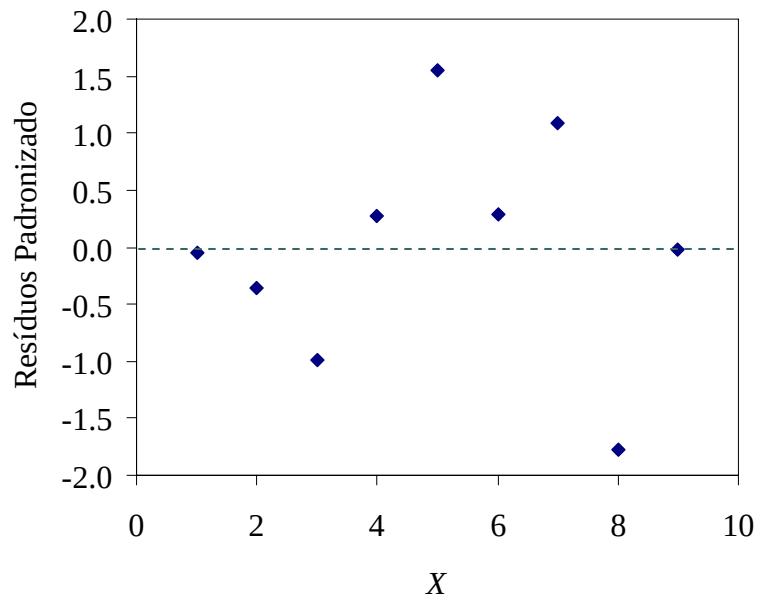




Resíduo = $e_i = Y_i - \hat{Y}_i$

Análise de Resíduos

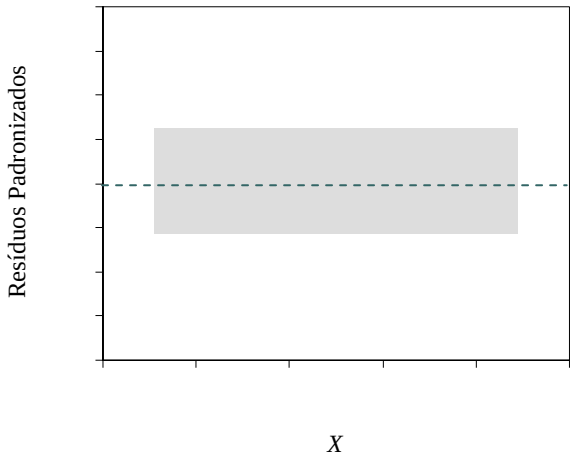




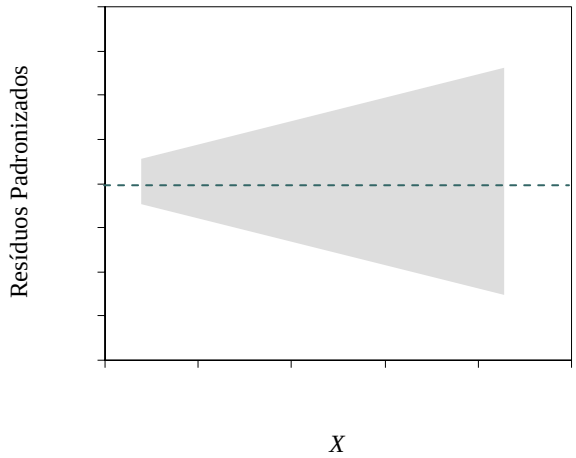
Resíduo Padronizado = $e_i \Big/ \sqrt{MQRes}$

Análise de Resíduos

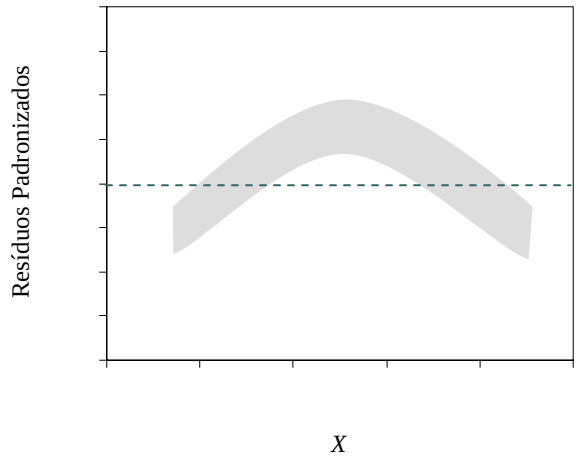
"ideal"



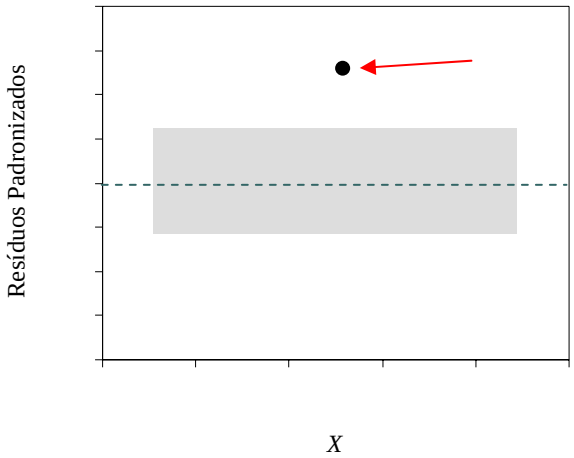
σ^2 não constante



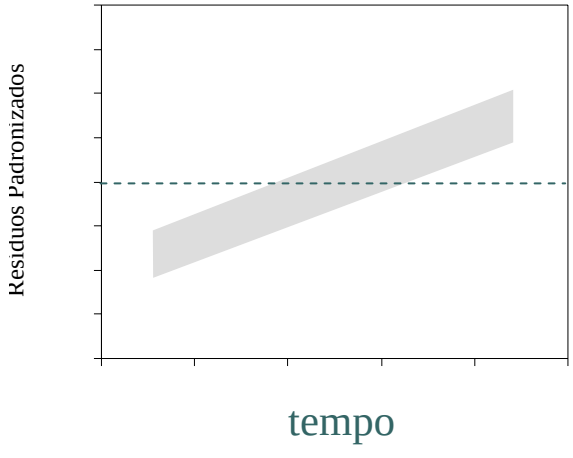
não linearidade



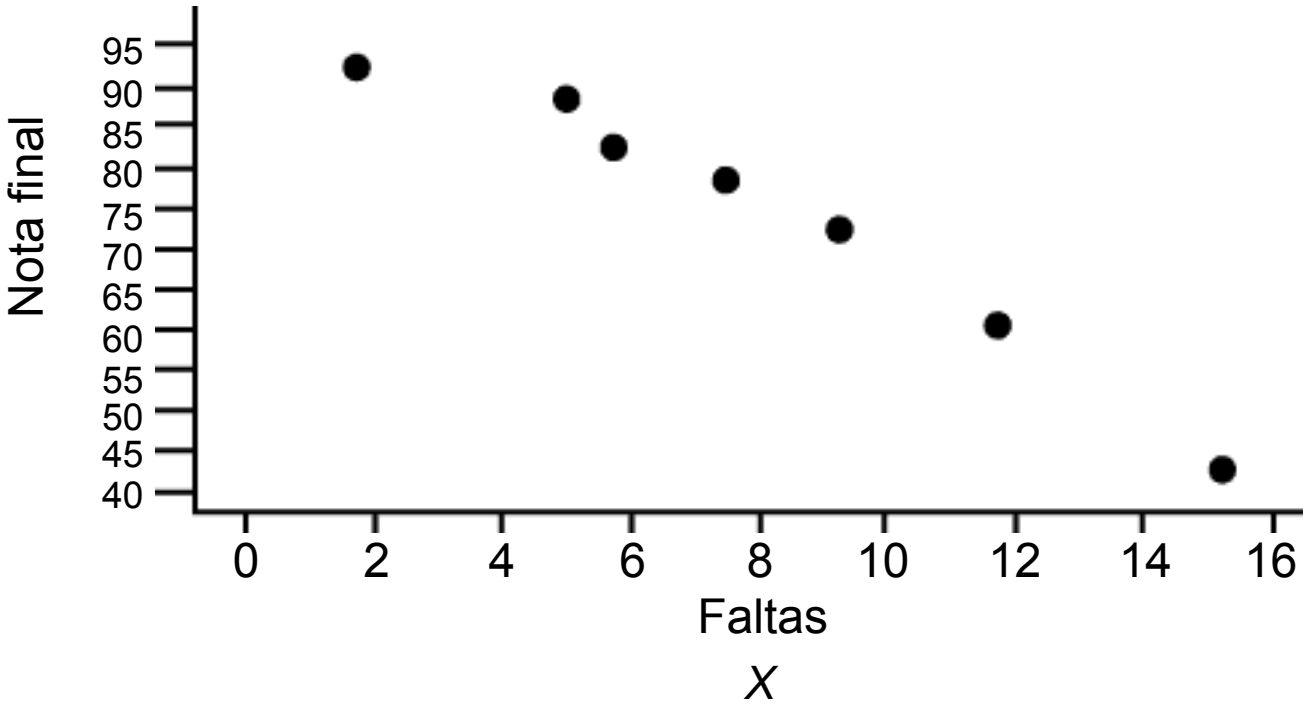
"outlier"



não independência



Aplicação



Faltas	Nota final
x	*y*
8	78
2	92
5	90
12	58
15	43
9	74
6	81

Cálculo de *r*

	x	y	xy	x^2	y^2
1	8	78	624	64	6.084
2	2	92	184	4	8.464
3	5	90	450	25	8.100
4	12	58	696	144	3.364
5	15	43	645	225	1.849
6	9	74	666	81	5.476
7	6	81	486	36	6.561
	57	**516**	**3.751**	**579**	**39.898**

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^{\,2} - \left(\sum_{i=1}^{n} y_i\right)^2}} = \frac{-3155}{\sqrt{804}\sqrt{13030}} = -0.975$$

Escreva a equação da reta de regressão com ***x* = número de faltas** e ***y* = nota final**.

	x	*y*	*xy*	x^2	y^2
1	8	78	624	64	6.084
2	2	92	184	4	8.464
3	5	90	450	25	8.100
4	12	58	696	144	3.364
5	15	43	645	225	1.849
6	9	74	666	81	5.476
7	6	81	486	36	6.561
	57	**516**	**3.751**	**579**	**39.898**

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = -3{,}924$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 105{,}667$$

A equação de regressão é dada por:

$$\hat{y} = 105{,}667 - 3{,}924x_i$$

Prevendo Valores

Com a reta de regressão, é possível prever valores de *y* correspondentes aos valores de *x*.

Usando a equação de regressão podemos prever a nota esperada de um aluno com:

(a) 3 faltas **(b) 12 faltas**

(a) $\hat{y} = -3{,}924(3) + 105{,}667 = 93{,}895$

(b) $\hat{y} = -3{,}924(12) + 105{,}667 = 58{,}579$