Inferência Estatística

Correlação e Regressão

- Coeficiente de Correlação
- **Regressão Linear Simples**

Regressão Linear Simples

- Em muitos problemas há duas ou mais variáveis que são relacionadas, e pode ser importante modelar essa relação.
- Por exemplo, pode-se ter interesse em predizer
  - as vendas futuras de um produto em função do seu preço,
  - a perda de peso de uma pessoa em decorrência do número de dias que se submete a uma determinada dieta,
  - a produção de uma determinada cultura em função da quantidade de nutriente aplicada no solo.

- Outro exemplo, as vendas de um produto podem estar relacionadas ao valor gasto em marketing com esse produto. Assim, é possível construir um modelo relacionando vendas a gastos com marketing, e então pode-se usar esse modelo para fins previsão de vendas.

- Em geral vamos supor que há uma **variável dependente** (ou variável de resposta) **Y** que depende de uma **variável preditora** (ou variável explicativa) **X**.

- A regressão linear simples estima uma equação matemática (ou modelo) que, dado o valor de X (variável preditora), prevê o valor de Y (variável dependente).

- É dito regressão **linear simples**, pois supõe-se tendência linear entre as variáveis e simples por ser uma única variável preditora.

- Modelo de regressão linear simples

$$Y = \beta_0 + \beta_1 X + \varepsilon$$ (erro aleatório)

- O coeficiente β_0 é a *interseção* (valor de Y para $X = 0$).

- O coeficiente β_1 é a *inclinação* da reta, que pode ser positiva, negativa ou nula.

- Se há n pares de dados $(y_1, x_1), \ldots, (y_n, x_n)$ é possível estimar os parâmetros β_0 e β_1 usando o Método dos Mínimos Quadrados.

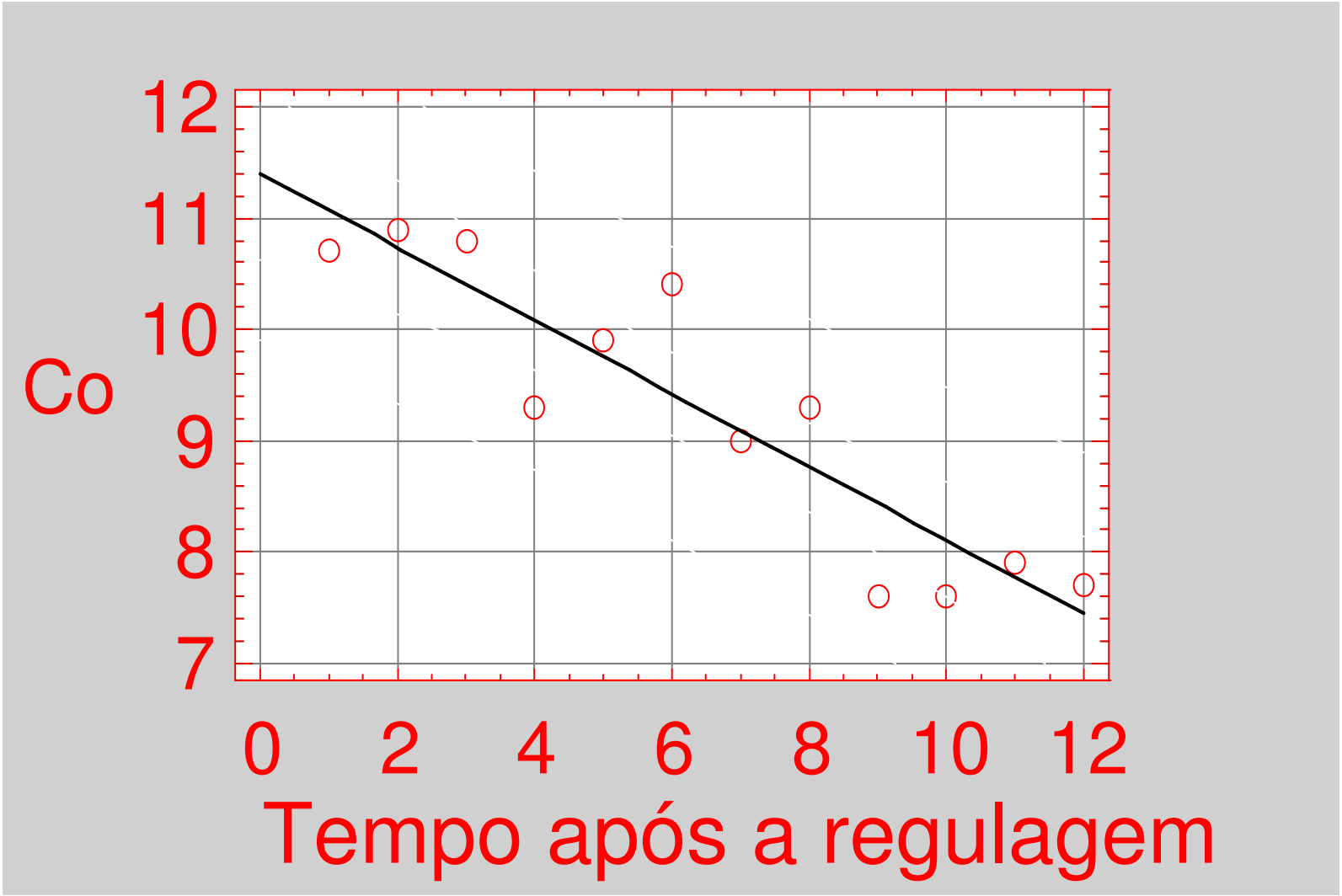
- Temos então b_0 e b_1 , estimativas amostrais de β_0 e β_1.

$$b_1 = \frac{\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)/n}{\sum x_i^2 - \left(\sum x_i\right)^2/n} \qquad b_0 = \overline{Y} - b_1 \overline{X}$$

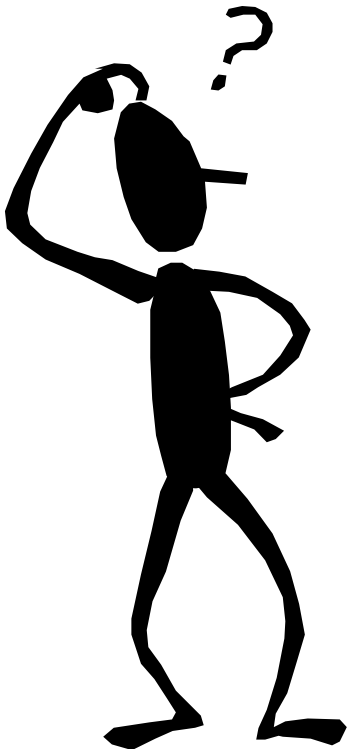
Exemplo: Após uma regulagem eletrônica um veículo apresenta um rendimento ideal no que tange a consumo de combustível. Contudo, com o passar do tempo esse rendimento vai se degradando. Os dados a seguir representam o rendimento medido mês a mês após a regulagem. Ajuste um modelo linear a esses dados.

X:meses após a regulagem	1	2	3	4	5	6
Y : rendimento	10,7	10,9	10,8	9,3	9,5	10,4
X:meses após a regulagem	7	8	9	10	11	12
Y : rendimento	9,0	9,3	7,6	7,6	7,9	7,7

Rendimento de combustível



Cálculos iniciais



Meses(X)	Rendimento(Y)	X^2	Y^2	X*Y
1	10,7	1	114,49	10,7
2	10,9	4	118,81	21,8
3	10,8	9	116,64	32,4
4	9,3	16	86,49	37,2
5	9,5	25	90,25	47,5
6	10,4	36	108,16	62,4
7	9	49	81	63
8	9,3	64	86,49	74,4
9	7,6	81	57,76	68,4
10	7,6	100	57,76	76
11	7,9	121	62,41	86,9
12	7,7	144	59,29	92,4
78	**110,7**	**650**	**1039,55**	**673,1**
6,5	**9,225**			

$\Sigma x = 78$

$\Sigma y = 110,7$

$\Sigma x\,y = 673,1$

$\overline{X} = 6,50$

$\overline{Y} = 9,225$

$\Sigma x^2 = 650$

$\Sigma y^2 = 1039,55$

Cálculos

$\Sigma x = 78$ $\quad \overline{X} = 6{,}50$ $\quad \Sigma x^2 = 650$

$\Sigma y = 110{,}7$ $\quad \overline{Y} = 9{,}225$ $\quad \Sigma xy = 637{,}1$

$$S_{XY} = \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)/n = 673{,}1 - (78 \times 110{,}70)/12 = -46{,}45$$

$$S_{XX} = \sum x_i^2 - \left(\sum x_i\right)^2/n = 650 - (78)^2/12 = 143$$

Estimativa dos parâmetros:

$$b_1 = \frac{\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)/n}{\sum x_i^2 - \left(\sum x_i\right)^2/n}$$

$b_1 = -46{,}45 / 143{,}00 = -0{,}325$

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

$b_0 = 9{,}225 - (-0{,}325)\ 6{,}50 = 11{,}34$

Equação de regressão

$Y = 11{,}34 - 0{,}325\ X$

⇐ Interpretar e utilizar!!!

Coeficiente de Determinação

r^2 é conhecido como Coeficiente de Determinação

r^2 = quadrado do coeficiente de correlação r

r^2 equivale a proporção da variância dos valores de Y que pode ser atribuída à regressão com a variável X.

- Para o exemplo, resultou $r^2 = (-0{,}907)^2 = 0{,}82$, ou seja, 82% da variabilidade nos resultados de rendimento de combustível pode ser devida ao tempo decorrido após a regulagem.

18% da variabilidade total é devido a outros fatores que não foram investigados.

Intervalos de Confiança e Testes de Hipótese

Como os resíduos de *Y* supostamente seguem a distribuição normal, e como os valores de b_0 e b_1 são funções lineares de *Y*:

$$b_0 \rightarrow N\left(\beta_0, \sigma^2_{b0}\right) \qquad b_1 \rightarrow N\left(\beta_1, \sigma^2_{b1}\right)$$

Esses resultados podem ser usados em testes de hipótese. Por exemplo, se a hipótese é:

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

então calcula-se:

$$t = b_1 / S_{b1}$$

⇨ H_0 é rejeitada se $|t| > t_{\alpha/2,n-2}$.

O intervalo de confiança para β_1 virá dado por

$$b_1 - t_{\alpha/2} S_{b1} < \beta_1 < b_1 + t_{\alpha/2} S_{b1}$$

- Usando os dados do problema do consumo de combustível, vamos construir um intervalo de confiança para a inclinação b_1 e verificar a hipótese.

$$S_{xx} = 143 \qquad S_{yy} = 18{,}34 \qquad S_{xy} = -46{,}45$$

$$SQR = S_{YY} - b_1 S_{XY} = 3{,}24$$

$$S^2 = \frac{SQR}{n-2} = 0{,}324 \qquad S_{b1}^2 = \frac{S^2}{S_{XX}} = 0{,}00227 \Rightarrow S_{b1} = 0{,}0476$$

Intervalo de confiança para β_1

$$b_1 - t_{\alpha/2} S_{b1} < \beta_1 < b_1 + t_{\alpha/2} S_{b1}$$

$t_{0,025;10} = 2,228$

$$-0,325 - 2,228\,(0,0476) < \beta_1 < -0,325 + 2,228\,(0,0476)$$

$$-0,431 < \beta_1 < -0,219$$

Como esse intervalo não inclui o zero, a hipótese $\beta_1 = 0$ é rejeitada, ou seja, existe uma relação entre o consumo de combustível e o tempo decorrido após a regulagem.