# Regressão Linear Simples (Seção 9.5)

## Coeficiente de Correlação Linear ($\rho_{XY}$)

Dadas duas variáveis quantitativas X e Y, medidas na mesma unidade amostral, avalia o grau de relação linear existente entre as duas variáveis. Ex: Peso e Altura, Salário e Anos de Escolaridade, Gasto com energia e Tamanho da residência

Dada uma amostra de n pares de observações

$(x_1, y_1)\ (x_2, y_2)\ \dots\ (x_n, y_n)$

$$\rho_{XY} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{j=1}^{n} y_j^2 - n\bar{y}^2\right)}}$$

$-1 \leq \rho_{XY} \leq 1$

$\rho_{XY} \simeq 0$ inexistência de relação linear

$\rho_{XY} = 1$ perfeita relação linear crescente

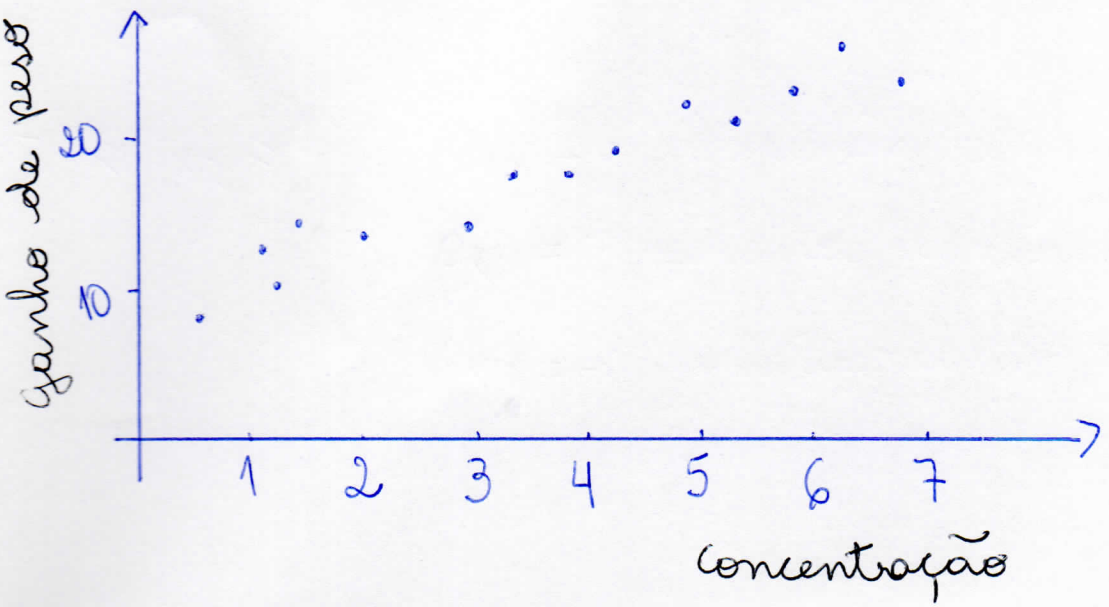
$\rho_{XY} = -1$ perfeita relação linear decrescente

# Exemplo 9.19 pág 348 Magalhães e Lima

Acredita-se que a utilização de uma certa substância no pasto produz um maior ganho de peso no gado. Para verificar o fato, foram escolhidos 15 bois de mesma raça e idade e cada animal recebeu uma determinada concentração da substância (variável X, em mg/l). O ganho de peso após 30 dias (variável Y, em kg) foi anotado. Os resultados foram

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| X | 0,2 | 0,5 | 0,6 | 0,7 | 1,0 | 1,5 | 2,0 | 2,5 |
| Y | 9,4 | 11,4 | 12,3 | 10,2 | 11,9 | 13,6 | 14,2 | 16,2 |
| X | 3,0 | 3,5 | 4,0 | 4,5 | 5,0 | 5,5 | 6,0 | |
| Y | 16,2 | 17,7 | 18,8 | 19,9 | 22,5 | 24,7 | 23,1 | |

## Diagrama de Dispersão



$\sum x_i y_i = 785,55 \qquad \bar{x} = 2,7 \qquad \bar{y} = 16,14$

$\sum x_i^2 = 163,39 \qquad \sum y_i^2 = 4239,43$

$$\rho_{xy} = \frac{785,55 - 15 \times 2,7 \times 16,14}{\sqrt{(163,39 - 15.2,7^2)\,(4239,43 - 15.16,14^2)}} = 0,99$$

Forte relação linear crescente.

Quanto maior a concentração da substância, maior o ganho de peso.

Os pontos tendem a se concentrar em torno de uma reta crescente.

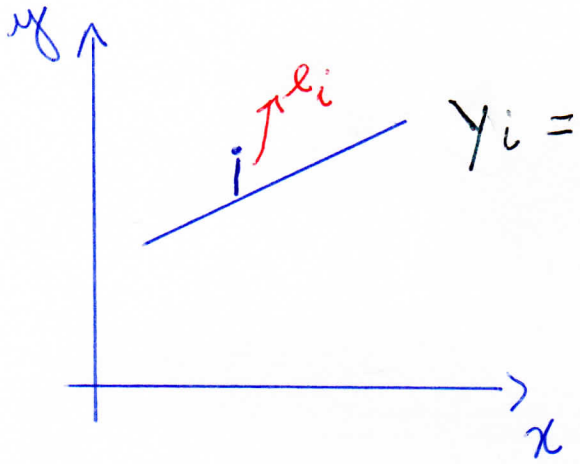
Acredita-se que para cada valor $x_i$ da variável X, tem-se

$$Y_i = \underbrace{\alpha + \beta x_i}_{\text{parte de } Y_i \text{ que é explicada por X}} + \underbrace{e_i}_{\text{erro do modelo (parte de } Y_i \text{ não explicada por X)}}$$

↓

Variáveis externas não consideradas

Variabilidade natural entre indivíduos




$\alpha$ e $\beta$ parâmetros desconhecidos

$Y_i = \alpha + \beta x_i + \ell_i$ é um modelo de regressão linear simples. Suposições

i) $\ell_i$, $i = 1, 2 \dots n$ são v. a. independentes

$\ell_i \sim N(0, \sigma^2)$

ii) Os valores de $x_i$ são fixados e portanto não aleatórios. Como consequência

$Y_i$, $i = 1, 2 \dots n$ são v. a. independentes e

$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$

Y - variável resposta ou variável dependente

X - variável independente, explicativa ou preditora.

$E(Y|X=x)$ : média de $Y$ quando $X=x$

Ex: $E(Y|X=4)$ → ganho médio de peso para bois submetidos à concentração 4,0 mg/l.

Admite-se que $E(Y_i|X=x_i) = \alpha + \beta x_i$

Significado de $\beta$

$E(Y|X=x+1) = \alpha + \beta(x+1) = \alpha + \beta x + \beta$

$E(Y|X=x) = \alpha + \beta x$

$\beta = E(Y|X=x+1) - E(Y|X=x)$

↳ acréscimo na média de $Y$ quando $X$ aumenta em uma unidade (suposto constante $\forall$ o valor de $X$)

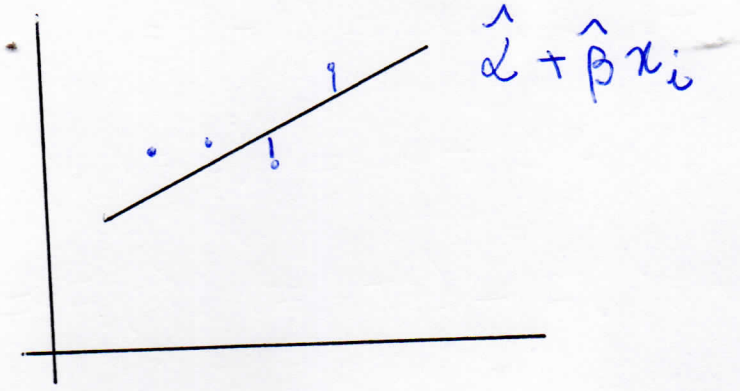
$\alpha = E(Y|X=0)$

↳ só tem significado prático se fizer sentido $X=0$.

# Estimadores de Mínimos Quadrados de $\alpha$ e $\beta$

Dada a amostra de $n$ pares de observações, os estimadores de mínimos quadrados de $\alpha$ e $\beta$, $\hat{\alpha}$ e $\hat{\beta}$, são os valores de $\alpha$ e $\beta$ que minimizam

$$SQ(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 .$$




A reta de mínimos quadrados $\hat{y} = \hat{\alpha} + \hat{\beta} x$ é aquela que dentre todas as possíveis retas minimiza a soma dos quadrados das distâncias verticais dos pontos $(x_i, y_i)$ à reta.

$\hat{y} = \hat{\alpha} + \hat{\beta} x$ é denominada reta de regressão de $Y$ em $X$.

$$\left.\begin{array}{l} \dfrac{\partial SQ(\alpha,\beta)}{\partial \beta} = 0 \\ \dfrac{\partial SQ(\alpha,\beta)}{\partial \alpha} = 0 \end{array}\right\} \Rightarrow$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

No exemplo

$n = 15 \qquad \sum x_i y_i = 785{,}55 \qquad \sum x_i^2 = 163{,}39$

$\bar{x} = 2{,}7 \qquad \bar{y} = 16{,}14$

$$\hat{\beta} = \frac{785{,}55 - 15 \cdot 2{,}7 \cdot 16{,}14}{163{,}39 - 15 \cdot 2{,}7^2} = 2{,}44$$

$$\hat{\alpha} = 16{,}14 - 2{,}44 \cdot 2{,}7 = 9{,}55$$

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = 9{,}55 + 2{,}44x$$

Obtida a reta de regressão

$$\hat{y} = 9{,}55 + 2{,}44x$$

estima-se que, com o aumento em uma unidade da concentração da substância, o ganho médio de peso aumenta em 2,44 kg.

Estima-se que o ganho médio de peso para bois que não recebem a substância é 9,55 kg.

A estimativa do ganho médio de peso de um boi sujeito a 6,2 mg/l da substância segundo o modelo ajustado é

$$\hat{y} = 9{,}55 + 2{,}44 \cdot 6{,}2 = 24{,}678$$

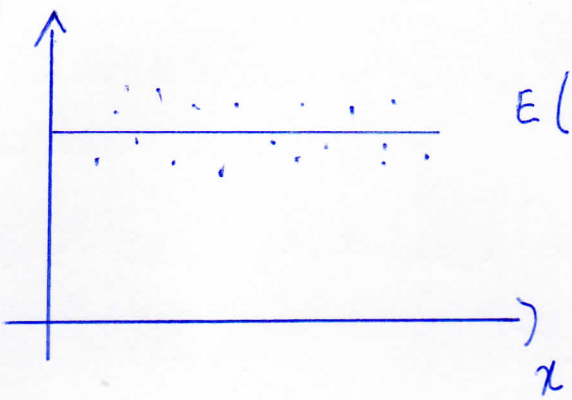
Não extrapolar muito.

# Teste de hipóteses para o parâmetro $\beta$

$H_0: \beta = 0$

$H_a: \beta \neq 0$

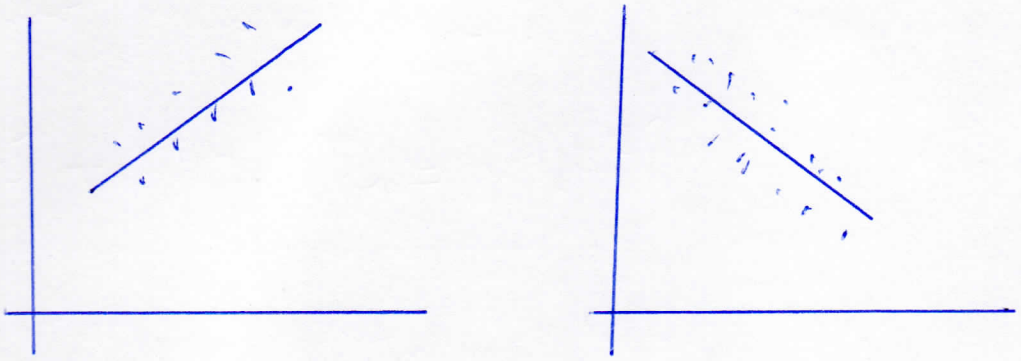
$\beta = 0$




$E(Y|X=x) = \alpha$

Não há influência da variável explicativa na média da variável resposta.

$\beta \neq 0$



Há influência da variável explicativa na média da variável resposta.

Verifica-se que $Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

$\sigma^2$ desconhecido.

Estimador não viciado de $\sigma^2$:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ → soma de quadrados do resíduo

Estatística de teste

$$T = \frac{\hat{\beta}}{\sqrt{\frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}}}$$

Rejeita-se $H_0$ para $T \geq t_c$ ou $T \leq -t_c$

$t_c$ | $P(T \geq t_c) = \alpha/2$ $\quad T \sim t_{n-2}$

No exemplo

| $i$ | $(y_i - \hat{y}_i)^2$ | | $(x_i - \bar{x})^2$ |
|---|---|---|---|
| 1 | 0,41 | $= (9,4 - 9,55 - 2,44 \cdot 0,2)^2$ | 6,25 |
| 2 | 0,39 | | 4,84 |
| 3 | 1,65 | | 4,41 |
| 4 | 1,12 | | 4,00 |
| ⋮ | | | |
| 15 | 1,20 | | 10,89 |
| Total: | 10,09 | = Soma de Quadrados do resíduo | 54,04 |
| | | | $\Sigma (x_i - \bar{x})^2$ |

$$\hat{\sigma}^2 = \frac{10,09}{13} = 0,776$$

$$T = \frac{2,44}{\sqrt{\frac{0,776}{54,04}}} = 20,375$$

13 gl

$\alpha = 0,05$

Teste bicaudal

$H_0: \beta = 0$

$H_a: \beta \neq 0$

RC : $T \leq -2,16$ ou $T \geq 2,16$

Rejeita-se $H_0$.

Os dados sugerem que o ganho médio de peso é influenciado pela concentração da substância.

Adota-se o modelo $\hat{y} = 9{,}55 + 2{,}44x$

Se $H_0$ fosse aceita, a conclusão seria que os dados sugerem que o ganho médio de peso não é influenciado pela concentração da substância. Não se utilizaria o modelo

$\hat{y} = 9{,}55 + 2{,}44x$