Cap. 9 Introdução à regressão linear simples

Modelos de regressão são modelos usados para estudar a associação/relação entre variáveis.

9.1 Modelo de Regressão Linear Simples (RLS)

- Amostra formada por pares de observações

$$\{(x_i, y_i),\ i = 1, 2, \ldots, n\}$$

estando subjacente uma relação linear afetada por um termo de erro, i.e.

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{(modelo RLS)}$$

onde:

Y_i: variável de resposta (ou dependente) associada à i-ésima prova – variável aleatória

x_i: variável explicativa (ou independente ou regressor – constante conhecida, não-aleatória

β_0: ordenada na origem – parâmetro

β_1: declive ou coef. angular – parâmetro regressor

ε_i: erro aleatório associado à i-ésima prova, verificando: (i) $E(\varepsilon_i) = 0$, $\forall i$

(ii) $Var(\varepsilon_i) = \sigma^2$ (constante indep. de i), $\forall i$

(iii) $Cov(\varepsilon_i, \varepsilon_j) = 0$, $\forall i, j$ com $i \neq j$

$(i = 1, \ldots, n)$

- Observações:

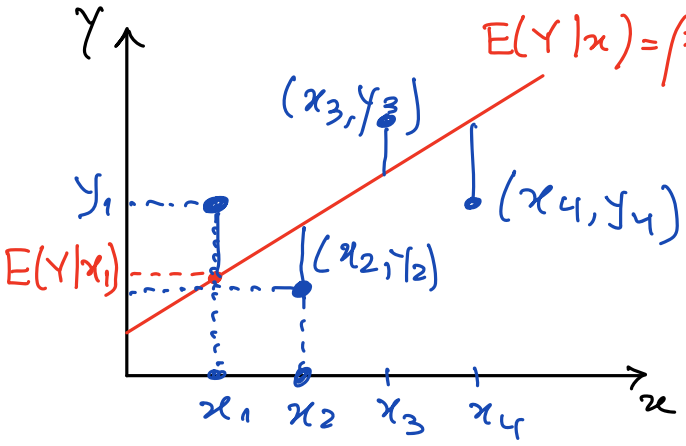
1) $E(\varepsilon_i) = 0 \Rightarrow E[Y|x_i] = \beta_0 + \beta_1 x_i$ é o ponto na reta de regressão $\boxed{Y = \beta_0 + \beta_1 x}$ com ordenada x_i

2) $\beta_0 = E(Y|x=0)$; $\beta_1 = E(Y|x = x^*_{+1}) - E(Y|x=x^*)$

3) O modelo RLS diz-se

- Simples por só haver uma var. explicativa (em vez de por exemplo $Y_i = \beta_0 + \beta_1 (x_1)_i + \beta_2 (x_2)_i + \varepsilon_i$)
- linear por ser linear nos parâmetros ($\beta_0, \beta_1, \ldots$) (ou seja, modelos do tipo $Y = \beta_0 + \beta_1 x^2 + \varepsilon$ ou $\log Y = \beta_0 + \beta_1 \log x + \varepsilon$ também são lineares).

- Método dos Mínimos Quadrados para $\hat{\beta}_0$ e $\hat{\beta}_1$



$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

$(\hat{\beta}_0, \hat{\beta}_1)$ = estimadores de mínimos quadrados de β_0 e β_1

= $\arg\min Q(\beta_0, \beta_1)$

$$\begin{cases} \partial Q/\partial \beta_0 = 0 \\ \partial Q/\partial \beta_1 = 0 \end{cases} \Leftrightarrow \begin{cases} \left(\sum_i Y_i\right) - n\beta_0 - \beta_1 \left(\sum_i x_i\right) = 0 \\ \left(\sum_i x_i Y_i\right) - \beta_0 \left(\sum_i x_i\right) - \beta_1 \left(\sum_i (x_i)^2\right) = 0 \end{cases}$$

$$\Rightarrow \boxed{\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \quad e \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} (x_i)^2 - n\bar{x}^2}}$$

com $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ e $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$

$$\left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

Observações:

1) Quando existe solução ela corresponde de facto a um mínimo (tem Hessiana definida positiva).

2) Existe solução sse $(\sum_i x_i^2) - n\bar{x}^2 = \sum_i (x_i - \bar{x})^2 \neq 0$, sse na amostra existirem pelo menos dois valores distintos de x.

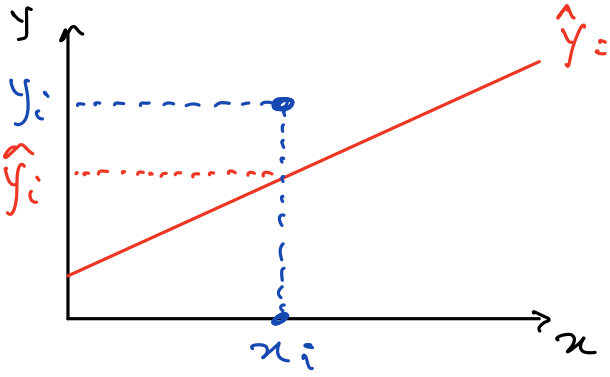
- Exercício 9.1 [...]. Cinco medições conduziram a $\sum_{i=1}^{5} x_i = 12$, $\sum_{i=1}^{5} (x_i)^2 = 32.5$, $\sum_{i=1}^{5} y_i = 0.177$, $\sum_{i=1}^{5} y_i^2 = 0.006789$ e $\sum_{i=1}^{5} x_i y_i = 0.4685$. Temos então que

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{5} x_i y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{5} x_i^2 - n\bar{x}^2} = \frac{0.4685 - 5\left(\frac{12}{5}\right)\left(\frac{0.177}{5}\right)}{32.5 - 5\left(\frac{12}{5}\right)^2} = \frac{0.0602}{3.7}$$

$= 0.01627$ e $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{0.177}{5} - 0.01627 \times \frac{12}{5} = -0.003648$

- **Coeficiente de Determinação**

O coeficiente de determinação, R^2, é uma medida descritiva indicadora da qualidade do ajustamento da reta estimada.



Resíduos: $e_i = y_i - \hat{y}_i$

com $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Observação: $\boxed{\sum_{i=1}^{n} e_i = \left(\sum_{i=1}^{n} y_i\right) - \sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) =}$

$= n\bar{y} - n\hat{\beta}_0 - \hat{\beta}_1 n\bar{x} = n\bar{y} - n(\bar{y} - \hat{\beta}_1 \bar{x}) - n\hat{\beta}_1 \bar{x} = 0$

$$\Rightarrow \underbrace{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}_{SST} = \underbrace{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}_{SSE} + \underbrace{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}_{SSR}$$

SST = Soma dos quadrados total

SSR = Soma dos quadrados da regressão

SSE = Soma dos quadrados dos resíduos

$= \sum_{i=1}^{n} (e_i)^2$

Def.: O coeficiente de determinação é definido por

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2}$$

$$\Rightarrow \boxed{R^2 = \frac{\left(\left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{Y}\right)^2}{\left(\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2\right)\left(\left(\sum_{i=1}^{n} y_i^2\right) - n\bar{Y}^2\right)}} \quad (\text{cf. formulário})$$

Observações

1) $0 \leq R^2 \leq 1$ e $(R^2 \times 100)\%$ representa a % da variabilidade total que é explicada pelo modelo de RLS.

2) $\boxed{r^2 = 1 \Leftrightarrow \hat{y}_i = y_i, \forall i.}$ Todos os pontos da amostra estão sobre a reta, pelo que o modelo RLS explica toda a variabilidade observada. O modelo é ótimo!

3) $\boxed{r^2 = 0 \Leftrightarrow \hat{y}_i = \bar{Y}, \forall i \Leftrightarrow \hat{\beta}_1 = 0 \ (\Leftrightarrow \hat{\beta}_0 = \bar{Y})}$

A reta é horizontal, pelo que o modelo RLS não explica nada da variabilidade observada. O modelo é péssimo!

• Exercício 9.4: Y = freq. cardíaca em batimentos por minuto

x = temperatura corporal em °C

130 medições independentes deram os seguintes resultados:

$$\sum_{i=1}^{130} x_i = 4784.7 \; ; \; \sum_{i=1}^{130} x_i^2 = 176\,121.67 \; ; \; \sum_{i=1}^{130} y_i = 9589 \; ;$$

$$\sum_{i=1}^{130} y_i^2 = 713\,733 \; ; \; \sum_{i=1}^{130} x_i y_i = 353018.5 \; .$$

Ajuste um modelo RLS a este conjunto de dados, calcule o coeficiente de determinação e comente a utilidade do modelo ajustado.

- $\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = 353018.5 - 130 \times \frac{4784.7}{130} \times \frac{9589}{130}$

$= 91.7$

- $\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 176121.67 - 130\left(\frac{4784.7}{130}\right)^2$

$= 18.9$

- $\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 713733 - 130 \times \left(\frac{9589}{130}\right)^2$

$= 6433.6$

- $\hat{\beta}_1 = \frac{91.7}{18.9} = 4.85$ e $\hat{\beta}_0 = \frac{9589}{130} - 4.85 \times \frac{4784.7}{130}$

$= -104.74$

- $R^2 = \frac{(91.7)^2}{18.9 \times 6433.6} = 0.069$

$\Rightarrow$ modelo só explica cerca de 7% da variabilidade, pelo que é muito pouco útil.