Resolução de dois problemas do Capítulo 9
(Introdução à regressão linear simples)

9.1 a) Fazer no Excel (por exemplo). No gráfico pode ver-se que um modelo em que *y* varia linearmente com *x* (recta) se ajustará razoavelmente às observações (pelo menos na gama de valores observados).

b) Modelo de regressão linear simples: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1,\ldots,12$

com $V(\varepsilon_i) = \sigma^2$ e $\text{cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \forall_{i \neq j}$

Para calcular as estimativas são necessárias as seguintes quantidades ($\sum y_i^2$ só será necessário na alínea c) mas calcula-se já):

$$\sum_{i=1}^{12} x_i = 576 \Rightarrow \bar{x} = 48 \qquad \sum_{i=1}^{12} y_i = 3239 \Rightarrow \bar{y} = \frac{3239}{12}$$

$$\sum_{i=1}^{12} x_i^2 = 31488 \qquad \sum_{i=1}^{12} y_i^2 = 897639 \qquad \sum_{i=1}^{12} x_i y_i = 164752$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{164752 - 12 \times 48 \times \frac{3239}{12}}{31488 - 12 \times 48^2} = 2.41(6)$$

(não esquecer que é importante **não fazer arredondamentos** nos cálculos intermédios)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{3239}{12} - 2.41(6) \times 48 = 153.91(6)$$

c)

$$R^2 = \frac{\left(\sum x_i y_i - n\bar{x}\bar{y}\right)^2}{\left(\sum x_i^2 - n\bar{x}^2\right) \times \left(\sum y_i^2 - n\bar{y}^2\right)} =$$

$$= \frac{\left(164752 - 12 \times 48 \times \frac{3239}{12}\right)^2}{\left(31488 - 12 \times 48^2\right) \times \left(897639 - 12 \times \left(\frac{3239}{12}\right)^2\right)} = 0.9593$$

Como o valor de R^2 está bastante próximo de 1 conclui-se que a recta estimada se ajusta bem aos pontos observados, o que já era qualitativamente visível pelo gráfico. Pode também afirmar-se que 95.93% da variação observada em *y* é explicada pela variável *x*.

d) Hipóteses. H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$ (uma vez que não há nenhuma indicação para escolher uma alternativa unilateral).

Estatística de teste: $T_0 = \dfrac{\hat{\beta}_1 - 0}{\sqrt{\dfrac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}}}$. Sob H_0 $T_0 \sim t_{10}$ se $\varepsilon_i \underset{iid}{\sim} N(0, \sigma^2)$

(assume-se a última condição, normalidade dos erros, como hipótese de trabalho, numa situação real deve ser verificado se tal é razoável, por análise dos resíduos)

Para calcular o valor observado da estatística de teste é necessário primeiro calcular a estimativa da variância dos erros:

$$\hat{\sigma}^2 = \frac{1}{n-2}\left[\left(\sum y_i^2 - n\bar{y}^2\right) - \left(\hat{\beta}_1\right)^2\left(\sum x_i^2 - n\bar{x}^2\right)\right] =$$

$$= \frac{1}{10}\left[\left(897639 - 12 \times \left(\frac{3239}{12}\right)^2\right) - \left(2.41(6)\right)^2\left(31488 - 12 \times 48^2\right)\right] = 95.225$$

Valor observado da estatística do teste:

$$t_0 = \frac{2.41(6)}{\sqrt{\dfrac{95.225}{31488 - 12 \times 48^2}}} = \frac{2.41(6)}{0.15747} = 15.35$$

Consultando a tabela da distribuição *t* verifica-se que o valor observado t_0 é maior que o maior valor que vem na tabela para 10 graus de liberdade, 4.587, a que corresponderia um nível de significância de 0.1%=0.001= $2 \times (1 - 0.9995)$. Pode afirmar-se que o valor-p é inferior a 0.001, pelo que se rejeita H_0 para os níveis de significância usuais (geralmente entre 1% e 5%).

Comentários:

- O interesse deste teste reside em que quando se rejeita a hipótese nula (e isso é uma conclusão "forte") significa que os dados indicam de forma significativa que a variável *x* é importante na explicação da variável *y*.
- A conclusão desta alínea está de acordo com o resultado obtido em c).

e) O que se pretende é um intervalo de confiança a 95% para $E(Y|x = 48) = \mu_{Y|x=48}$.

A estimativa pontual deste valor esperado é dada por $\hat{\mu}_{Y|x=48} = \hat{\beta}_0 + \hat{\beta}_1 \times 48 = 269.92$.

Para obter o intervalo pedido usa-se a variável aleatória fulcral seguinte:

$$T = \frac{\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) - \left(\beta_0 + \beta_1 x_0\right)}{\sqrt{\left(\frac{1}{n} + \frac{\left(\bar{x} - x_0\right)^2}{\sum x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} = \frac{\hat{\mu}_{y|x_0} - \mu_{y|x_0}}{se\left(\hat{\mu}_{y|x_0}\right)} \sim t_{n-2}$$

Procedendo como é habitual para obter um intervalo de confiança:

$$P(-a \le T \le a) = 0.95 \Leftrightarrow$$

$$\Leftrightarrow P\left(\mu_{y|x_0} \in \left[\hat{\mu}_{y|x_0} - a \times se\left(\hat{\mu}_{y|x_0}\right); \hat{\mu}_{y|x_0} + a \times se\left(\hat{\mu}_{y|x_0}\right)\right]\right) = 0.95 \quad \text{com } a = t_{n-2,0.975},$$

neste caso $a = t_{10,0.975} = 2.228$ e

$se\left(\hat{\mu}_{y|x=48}\right) = \sqrt{\left(\frac{1}{12} + \frac{(48-48)^2}{\sum x_i^2 - n\bar{x}^2}\right)95.225} = 2.817$, pelo que a concretização do intervalo aleatório deduzido dá:

$$I.C._{95\%}\left(\mu_{y|x=48}\right) = \left[\hat{\mu}_{y|x=48} - a \times se\left(\hat{\mu}_{y|x=48}\right); \hat{\mu}_{y|x=48} + a \times se\left(\hat{\mu}_{y|x=48}\right)\right] =$$

$$= [269.92 - 2.228 \times 2.817; 269.92 + 2.228 \times 2.817] =$$

$$= [263.64; 276.20]$$

Não é legítimo usar o mesmo procedimento para *x* = 10 horas porque 10 não pertence ao intervalo de variação dos valores de *x* observados: $\left[\min x_i; \max x_i\right] = [16;80]$ e não há nenhuma garantia de que o modelo seja válido fora deste intervalo. Antes pelo contrário, pois a resistência não pode crescer nem decrescer de forma ilimitada.

9.3 a)

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{637.1 - 10 \times 12.4 \times 5.21}{1560 - 10 \times 12.4^2} = -0.3991$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 5.21 - (-0.3991) \times 12.4 = 10.1589$$

Donde $\hat{\mu}_{y|x} = \hat{\beta}_0 + \hat{\beta}_1 x = 10.1589 - 0.3991x$.

Para obter o intervalo de confiança pedido usa-se a variável aleatória fulcral

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}}} \sim t_{n-2}$$

e procede-se de forma semelhante à da alínea (e) do problema 9.1 obtendo-se

$$I.C._{90\%}(\beta_1) = \left[\hat{\beta}_1 - a \times se\left(\hat{\beta}_1\right); \hat{\beta}_1 + a \times se\left(\hat{\beta}_1\right)\right] \text{ com } a = t_{n-2,0.95} = t_{8,0.95} = 1.86 \text{ e}$$

$$se\left(\hat{\beta}_1\right) = \sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}}$$

Cálculos:

$$\hat{\sigma}^2 = \frac{1}{n-2}\left[\left(\sum y_i^2 - n\bar{y}^2\right) - \left(\hat{\beta}_1\right)^2\left(\sum x_i^2 - n\bar{x}^2\right)\right] =$$

$$= \frac{1}{8}\left[\left(275.13 - 10 \times 5.21^2\right) - (-0.3991)^2\left(1560 - 10 \times 12.4^2\right)\right] = 0.0151228$$

$$se\left(\hat{\beta}_1\right) = \sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}} = \sqrt{\frac{0.0151228}{1560 - 10 \times 12.4^2}} = 0.02598$$

$$I.C._{90\%}(\beta_1) = \left[\hat{\beta}_1 - a \times se\left(\hat{\beta}_1\right); \hat{\beta}_1 + a \times se\left(\hat{\beta}_1\right)\right] =$$

$$= [-0.3991 - 1.86 \times 0.02598; -0.3991 + 1.86 \times 0.02598] =$$

$$= [-0.4474; -0.3508]$$

b) $\hat{\mu}_{y|x=10} = \hat{\beta}_0 + \hat{\beta}_1 \times 10 = 10.1589 - 0.3991 \times 10 = 6.1$

Para x = 20 não se pode fazer nenhuma predição usando este modelo pois isso corresponderia a uma extrapolação (ver a resposta à alínea e) do problema 9.1).