



LEIC, LERCI, LEE

Probabilidades e Estatística 2004/05

Colectânea de Exercícios

Capítulo 9

Introdução à regressão linear simples

9.1 Interessa estudar a relação entre a resistência de um determinado tipo de plástico (Y) e o tempo que decorre a partir da conclusão do processo de moldagem até ao momento de medição da resistência (x [horas]). As observações que se seguem foram efectuadas em 12 peças construídas com este plástico, escolhidas aleatoriamente.

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	32	48	72	64	48	16	40	48	48	24	80	56
y_i	230	262	323	298	255	199	248	279	267	214	359	305

- Represente graficamente as observações e desenhe a recta que, no seu entender, melhor se ajusta às observações.
- Considere um modelo de regressão linear simples para explicar as observações. Obtenha a estimativa dos mínimos quadrados dos coeficientes da recta de regressão e desenhe-a no gráfico.
- Com base na recta estimada de regressão dos mínimos quadrados, calcule uma estimativa da diferença esperada da resistência de 2 peças cujos tempos que decorrem a partir da conclusão do processo de moldagem até ao momento de medição da resistência diferem de 4 horas.
- Calcule o coeficiente de determinação e comente o valor obtido.
- Indique, justificando, se é de concluir que o tempo que decorre a partir da conclusão do processo de moldagem até ao momento de medição da resistência tem um impacto significativo no resultado esperado do teste.
- Proceda ao teste da hipótese “O coeficiente angular é nulo”. Qual o interesse desta hipótese? Relacione-o com o resultado obtido em (c).
- Calcule o intervalo de confiança a 95% para o valor esperado da resistência obtida 48 horas depois de concluída a moldagem. Acha legítimo usar o mesmo procedimento tratando-se de um período de 10 horas em vez de 48 horas? Justifique a sua resposta.

- (h) Calcule um intervalo de confiança a 95% para o valor da resistência de uma peça de plástico 48 horas depois de concluída a sua moldagem.

Resolução:

- (b) Considere-se o modelo de regressão linear simples: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, com $\{\varepsilon_i, i = 1, \dots, n\} \stackrel{\text{iid}}{\sim}$, com $E[\varepsilon_i] = 0$, $Var[\varepsilon_i] = \sigma^2, \forall i$. Denote-se por $\hat{\beta}_0$ ($\hat{\beta}_1$) a estimativa dos mínimos quadrados do coeficiente β_0 (β_1). Dado que

$$\begin{aligned} \sum_{i=1}^{12} x_i &= 576 & \sum_{i=1}^{12} y_i &= 3239 & \bar{x} &= 48.0 & \bar{y} &= 269.917 \\ \sum_{i=1}^{12} x_i^2 &= 31488 & \sum_{i=1}^{12} y_i^2 &= 897639 & \sum_{i=1}^{12} x_i y_i &= 164752 \end{aligned}$$

vem que

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{12} x_i y_i - 12\bar{x}\bar{y}}{\sum_{i=1}^{12} x_i^2 - 12\bar{x}^2} = \frac{164752 - 12 \times 48 \times 269.917}{31488 - 12 \times 48^2} = 2.417 \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = 269.917 - 2.417 \times 48 = 153.917. \end{aligned}$$

Logo

$$\hat{y} = \hat{E}[Y|x] = \hat{\beta}_0 + \hat{\beta}_1 x = 153.917 + 2.417x$$

é a estimativa de mínimos quadrados de $E[Y|x]$, para $x \in [16, 80]$.

- (c) Uma estimativa da diferença do valor esperado da resistência de 2 peças cujos tempos que decorrem a partir da conclusão do processo de moldagem até ao momento de medição da resistência diferem de 4 horas é dada por $4 \times \hat{\beta}_1 = 4 \times 2.417 = 9.668$, desde que os valores em questão se encontrem ambos no intervalo $[16, 80]$.
- (d) O valor observado do coeficiente de determinação é

$$\begin{aligned} r^2 &= \frac{(\sum_{i=1}^{12} x_i y_i - 12\bar{x}\bar{y})^2}{(\sum_{i=1}^{12} x_i^2 - 12\bar{x}^2)(\sum_{i=1}^{12} y_i^2 - 12\bar{y}^2)} \\ &= \frac{(164752 - 12 \times 48 \times 269.917)^2}{(31488 - 12 \times 48^2) \times (897639 - 12 \times 269.917^2)} = 0.9593 \end{aligned}$$

pelo que cerca de 95.93% da variabilidade das observações respeitantes à resistência do plástico é explicada pelo modelo de regressão linear simples ajustado pelo método dos mínimos quadrados. Assim sendo, considera-se que este modelo ajusta adequadamente as observações.

- (a) Doravante, para além das hipóteses acima mencionadas, assume-se igualmente que $\varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$. Pretende-se testar $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. Para tal considera-se a estatística de teste

$$T = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{\frac{\sum_{i=1}^{12} x_i Y_i - 12\bar{x}\bar{Y}}{\sum_{i=1}^{12} x_i^2 - 12\bar{x}^2}}{\frac{\hat{\sigma}^2}{\sum_{i=1}^{12} x_i^2 - 12\bar{x}^2}} \stackrel{H_0}{\sim} t_{12-2}$$

onde

$$\hat{\sigma}^2 = \frac{(\sum_{i=1}^{12} Y_i^2 - 12\bar{Y}^2) - \hat{\beta}_1^2 (\sum_{i=1}^{12} x_i^2 - 12\bar{x}^2)}{12 - 2}.$$

Concretizando para a amostra observada: $\hat{\sigma}^2 = 94.606$, $s_{\hat{\beta}_1} = 0.157$, e o valor observado da estatística de teste é $t_{obs} = 15.398$. O valor- p é igual a $2 \times F_{t_{10}}(-15.398) \approx 0.0$, pelo que se rejeita a hipótese de que o coeficiente angular seja nulo para todos os níveis de significância, concluindo-se pois pela significância estatística do modelo de regressão linear utilizado.

- (f) Nesta alínea pretende-se testar exactamente as mesmas hipóteses que na alínea anteriores, pelo que o resultado decorre da alínea (e). Note-se que a conclusão da significância estatística do modelo de regressão linear utilizado está em consonância com o resultado obtido na alínea (d), uma vez que o elevado valor observado do coeficiente de determinação indicia já que o modelo de regressão linear explica bem a variabilidade dos dados.
- (g) Pretende-se obter um intervalo de confiança a 95% para $E[Y|x = 48] = \mu_{Y|x=48}$. Note-se que $48 \in [\min x_i; \max x_i] = [16, 80]$ pelo que se considera válido o modelo de regressão linear para $x = 48$. A variável fulcral adequada para a determinação do intervalo de confiança é

$$T = \frac{\hat{\beta}_0 + 48\hat{\beta}_1 - \mu_{Y|x=48}}{S_{\hat{\beta}_0}} \sim t_{12-2}$$

onde $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$ e

$$S_{\hat{\beta}_0} = \sqrt{\hat{\sigma}^2 \left(\frac{1}{12} + \frac{(48 - \bar{x})^2}{\sum_{i=1}^{12} x_i^2 - 12\bar{x}^2} \right)}$$

Considerando então um grau de confiança 95%, vem que

$$P(-2.228 \leq T \leq 2.228) = 0.95$$

pelo que

$$\text{ICA}_{0.95}(\mu_{Y|x=48}) =]\hat{\beta}_0 + 48\hat{\beta}_1 - 2.228S_{\hat{\beta}_0}, \hat{\beta}_0 + 48\hat{\beta}_1 + 2.228S_{\hat{\beta}_0}[.$$

Concretizando para a amostra em questão, vem que $\text{IC}_{0.95}(\mu_{Y|x=48}) =]263.677; 276.189[$.

Se em vez de se considerar $x = 48$ tivermos $x = 10$, o procedimento acima adoptado já não é adequado uma vez que o modelo de regressão linear só é válido para valores de x no intervalo $x \in [\min x_i, \max x_i] = [16, 80]$.

- (h) Para construir um intervalo de confiança aleatório a 95% para $Y|x = 48$ utiliza-se um procedimento semelhante ao da alínea anterior excepto que a variável fulcral é

$$T = \frac{\hat{\beta}_0 + 48\hat{\beta}_1 - (Y|x = 48)}{\sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{12} + \frac{(48 - \bar{x})^2}{\sum_{i=1}^{12} x_i^2 - 12\bar{x}^2} \right)}} \sim t_{12-2}$$

pelo que $\text{IC}_{0.95}(Y|x = 48) =]247.377; 292.489[$.