INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO EXPERIMENTAL

31 de outubro, 2024 **PRIMEIRO TESTE 2024-25** **Uma resolução possível**

I

1. Os valores em falta A, B e C:
   B - o valor da estimativa do desvio padrão do estimador do parâmetro β_1 é a raíz quadrada do elemento (2,2) da matriz de (co-) variâncias estimadas dos estimadores dos parâmetros do modelo (matriz *vcov* do R, dada no enunciado), ou seja, $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{35978.63} = 189.68$;
   A - o valor da estimativa do coeficiente associado ao preditor Pb, $b_1 = 3.963 \times 189.68 = 751.70$;
   C - o coeficiente de determinação modificado, $R^2_{mod} = 1 - (1 - 0.8043) \cdot \frac{42-1}{35} = 0.7708$.

2. O teste de F ajustamento global do modelo pode ser formulado assim:

   Hipóteses: $H_0 : \mathcal{R}^2 = 0$ vs. $H_1 : \mathcal{R}^2 > 0$.
   Estatística do teste: $F = \frac{QMR}{QMRE} = \frac{n-(p+1)}{p} \frac{R^2}{1-R^2} \cap F_{(p,n-(p+1))}$, sob H_0.
   Nível de significância: $\alpha = 0.05$.
   Região Crítica (Unilateral direita): Rej. H_0 se $F_{calc} > f_{\alpha\,(p,n-(p+1))} = f_{0.05(6,35)} \approx 2.3$.
   Conclusões: O valor calculado da estatística é dado na listagem produzida pelo R ($F_{calc} = 23.98$). Logo, rejeita-se a hipótese nula, que corresponde à hipótese dum modelo inútil. Esta conclusão também resulta directamente da análise do valor de prova (*p-value*) associado à estatística de teste calculada: $p - value = 4.788e - 11 < \alpha$, o que corresponde a uma rejeição de H_0 aos níveis usuais de α, entre os quais 0.05.

   O valor obtido para o coeficiente de determinação $R^2 = 0.8043$ significa que 80.43% da variabilidade observada para o teor de fenóis total é explicada pela regressão sobre os 6 preditores.

3. Trata-se de um teste de hipóteses ao parâmetro β_6 :

   Hipóteses: $H_0 : \beta_6 = 1$ vs. $H_1 : \beta_6 \neq 1$
   Estatística do Teste: $T = \frac{\hat{\beta}_6 - 1}{\hat{\sigma}_{\hat{\beta}_6}} \cap t_{(n-(p+1))}$, sob H_0.
   Nível de significância: $\alpha = 0.05$.
   Região Crítica: (Bilateral) Rejeitar H_0 se $|T_{\text{calc}}| > t_{0.025(35)} \approx 2.0$.
   Conclusões: Tem-se $T_{\text{calc}} = \frac{b_6 - 1}{\hat{\sigma}_{\hat{\beta}_6}} = \frac{0.9531-1}{0.1817} = -0.258$. $|T_{\text{calc}}| = 0.258 < 2.0$, assim, não se rejeita a hipótese nula, ao nível de significância de 0.05, pelo que é admissível a afirmação.

4. (a) A descrição do modelo em notação matricial, incluindo os pressupostos associados aos erros aleatórios, foram explicados nas aulas teóricas e estão descritos nas páginas 100 e 101 das folhas de apoio, assim como nos slides das aulas teóricas.
   (b) Prova feita nas aulas téoricas e nas páginas 108 e 109 das folhas de apoio.

5. Num algoritmo de exclusão sequencial com base em testes T aos parâmetros do modelo (para $\alpha = 0.10$), a variável $Brix$ será a primeira variável preditora a excluir do modelo. Segundo os resultados apresentados, ao nível $\alpha = 0.10$, existem 3 variáveis candidatas a sair (isto é, não se rejeita $H_0 : \beta_j = 0$). Nestas condições, a primeira variável a excluir é a variável cujo teste ao respetivo coeficiente apresenta maior p-value (isto é, a variável $Brix$; $p - value > 0.10$ e não se rejeita H_0).

6. (a) Trata-se de comparar o modelo completo com um seu submodelo (teste F parcial). Neste contexto, temos:

Hipóteses: $H_0 : \mathcal{R}_c^2 = \mathcal{R}_s^2$ vs. $H_1 : \mathcal{R}_c^2 > \mathcal{R}_s^2$

Estatística do Teste: $F = \frac{n-(p+1)}{p-k} \cdot \frac{R_c^2 - R_s^2}{1-R_c^2} \cap F_{(p-k,n-(p+1))}$, sob H_0

Nível de significância: $\alpha = 0.05$

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{\text{calc}} > f_{\alpha(p-k,n-(p+1))}$

Conclusões: no enunciado encontra-se o valor $F_{\text{calc}} = 0.1971$. Uma vez que também é dado o p-value associado ao valor calculado da estatística do teste, verifica-se que $p - value = 0.8977 > 0.05$, logo, não se rejeita H_0, ou seja, modelo e submodelo não diferem significativamente ao nível 0.05 (ou a qualquer um dos outros níveis habitualmente utilizados), pelo que se deve optar pelo submodelo (modelo mais parcimonioso).

(b) A estimativa da variância dos erros aleatórios do submodelo com 3 preditores ($QMRE_S$) pode ser obtida a partir da informação listada no enunciado. Nos resultados apresentados encontram-se os valores das $SQRE$ para os 2 modelos, sob a designação `RSS`. O valor da $SQRE$ referente ao submodelo é o valor mais elevado ($SQRE_S = 85421$). Assim, $QMRE_S = \frac{SQRE_S}{g.l.(SQRE_S)} = \frac{85421}{38} = 2247.92$.
Os valores das 3 somas de quadrados: $SQT = (n-1)\, s_y^2 = 41 \times 10470.8364 = 429304.29$; $SQRE_S = 85421$; $SQR = 429304.29 - 85421 = 343883.29$.

(c) Prova feita nas aulas teóricas e na página 130 das folhas de apoio.

II

1. Pode ver-se na matriz de correlações apresentada no enunciado que o coeficiente de correlação entre teor de fenóis total e antocianas é $r_{Fen,Ant} = 0.8270$. Assim, o coeficiente de determinação desta regressão linear simples é $R^2 = 0.8270^2 = 0.6839$.

2. Sabemos que a expressão para o IC a $(1-\alpha) \times 100\%$ para o declive da reta populacional (β_1) é:

$$\Big] \; b_1 - t_{\alpha/2\,[n-2]} \cdot \hat{\sigma}_{\hat{\beta}_1} \quad , \quad b_1 + t_{\alpha/2\,[n-2]} \cdot \hat{\sigma}_{\hat{\beta}_1} \; \Big[\; .$$

Tendo em conta o *output* do `R` apresentado no enunciado, o valor do lado esquerdo do intervalo de IC a 95% ($b_1 - t_{0.025\,(40)} \cdot \hat{\sigma}_{\hat{\beta}_1}$) é 0.0007595435.
Então, sabendo que o valor estimado do declive é $b_1 = 0.0009631$, obtém-se

$$t_{0.025\,(40)} \cdot \hat{\sigma}_{\hat{\beta}_1} = 0.0009631 - 0.0007595435 = 0.0002035565.$$

O valor do extremo direito do IC a 95% será então: $b_1 + t_{0.025\,(40)} \cdot \hat{\sigma}_{\hat{\beta}_1} = 0.0009631 + 0.0002035565 = 0.001166657$.

Temos 95% de confiança em como o verdadeiro valor de β_1 (isto é, da variação esperada no logaritmo do teor de fenóis total quando o teor de antocianas aumenta uma 1 mg/l) está compreendido entre 0.0007595435 e 0.001166657.

3. Uma vez que foi apenas a variável resposta nesta regressão linear que foi logaritmizada, a relação não linear entre as variáveis originais é uma relação exponencial. Concretamente, e exponenciando a recta ajustada, tem-se:

$$\begin{aligned} \ln(y) &= 6.4237374 + 0.0009631 \cdot x \\ \Leftrightarrow \quad y &= e^{6.4237374+0.0009631\cdot x} = e^{6.4237374} \cdot e^{0.0009631\cdot x} \\ \Leftrightarrow \quad y &= 616.3022 \cdot e^{0.0009631\cdot x} \end{aligned}$$

em que y é a variável `Fen` e x é a variável `Ant`.