INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

23 de Janeiro, 2017 **SEGUNDO EXAME 2016-17** **Uma resolução possível**

I

Os dados correspondem a uma tabela de contingências, em que as contagens de pinheiros são classificadas de acordo com dois factores: o estado do pinheiro ao fim de dois anos (com $a = 2$ níveis: sobrevivência ou não) e a família de origem (com $b = 8$ níveis).

1. O teste pedido é um teste de homogeneidade. O número total de pinheiros bravos de cada família foi previamente fixado ($N_{\cdot j} = 60$ em todas as famílias) e pretende-se saber se há homogeneidade na distribuição desses 60 pinheiros pelas duas categorias de sobrevivência/não sobrevivência, ou seja, pretende-se saber se a probabilidade de sobrevivência é igual nas 8 famílias estudadas. Eis os passos do teste pedidos nesta alínea:

   Hipóteses: Represente-se por $\pi_{s|j}$ a probabilidade de sobrevivência, dada a família j (podendo associar-se a j as letras de A a H, como no enunciado, ou os inteiros de 1 a 8). Tem-se:

   Hipótese Nula (H_0): $\pi_{s|A} = \pi_{s|B} = ... = \pi_{s|H} \quad [= \pi_s$, onde π_s indica a probabilidade de sobrevivência comum]. Neste caso, e porque apenas há dois possíveis resultados (sobrevivência ou não), é opcional indicar que também tem de existir igualdade nas probabilidades de não sobrevivência: a igualdade das probabilidades condicionais de sobreviência obriga à igualdade nas probabilidades condicionais de não sobrevivência.

   Hipótese Alternativa (H_1): pelo menos uma das igualdades em H_0 não se verifica.

   Estatística do Teste: É a estatística de Pearson, na forma de contagens bidimensionais: $X^2 = \sum_{i=1}^{2} \sum_{j=1}^{8} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$, onde O_{ij} é a contagem correspondente ao resultado i ($i = 1$ a sobrevivência e $i = 2$ a não sobrevivência), na família j, e $\hat{E}_{ij}$ é o correspondente valor esperado, ao abrigo da hipótese nula de homogeneidade, dado por $\hat{E}_{ij} = \frac{N_{i\cdot} \times N_{\cdot j}}{N}$, onde $N = 480$ é o número total de pinheiros observados, $N_{i\cdot}$ indica o número de pinheiros associado ao resultado i e $N_{\cdot j}$ indica o número total de pinheiros correspondente à família j. A distribuição assintótica desta estatística, caso seja verdade H_0, é $\chi^2_{(a-1)(b-1)}$ (expressão geral num teste de homogeneidade).

   Nível de significância: $\alpha = \text{P[Erro do tipo I]} = \text{P[Rej. } H_0 \mid H_0 \text{ verdade]} = 0.05$

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $\chi^2_{calc} > \chi^2_{\alpha[(a-1)(b-1)]} = \chi^2_{0.05(7)} = 14.0671$.

2. O critério de Cochran sugere que a distribuição assintótica da estatística do teste de Pearson é válida se em nenhuma das células da tabela de contingências o número esperado de observações fôr inferior a 1, e em não mais de 20% das células inferior a 5. Ora, o número esperado (estimado) de observações apenas toma um de dois valores: para qualquer família j, tem-se $\hat{E}_{1j} = \frac{100 \times 60}{480} = 12.5$ e $\hat{E}_{2j} = \frac{380 \times 60}{480} = 47.5$. Logo, está-se em condições de poder admitir a validade da distribuição assintótica χ^2_7.

3. Como $\chi^2_{calc} = 38.804 > 14.0671$, rejeita-se a hipótese nula, ao nível de significância $\alpha = 0.05$ e conclui-se pela heterogeneidade das famílias, ou seja, conclui-se que há famílias de pinheiros

bravos com probabilidades de sobrevivência ao fim de dois anos maiores do que outras. Esta conclusão é inteiramente compatível com o que seria de esperar olhando para a tabela de contagens, uma vez que há famílias com apenas 2 pinheiros sobreviventes e outras com mais de 20 (em 60).

4. Tem-se $O_{11}=2$, $O_{21}=58$ e (como se viu acima) $\hat{E}_{11}=12.5$, $\hat{E}_{21}=47.5$. Logo, a contribuição para o valor de X^2_{calc} das duas parcelas correspondentes à família A é:

$$\frac{(2-12.5)^2}{12.5}+\frac{(58-47.5)^2}{47.5} \quad = \quad 11.14105$$

Não sendo, por si só, um valor suficiente para levar à rejeição de H_0, vemos apesar de tudo que a família A contribui de forma importante para essa rejeição, uma vez que lhe corresponde quase um terço (cerca de 29%) do valor de X^2_{calc}.

II

1. Considerem-se regressões lineares simples.

   (a) A melhor regressão linear simples corresponderá à utilização, como preditor, da variável mais correlacionada (em valor absoluto) com a variável resposta `taninos`. A partir da matriz de correlações disponível no enunciado, verifica-se que essa será a variável preditora `polifenois`, com $r=0.486$. Sabemos que, numa regressão linear, a proporção de variabilidade dos valores observados da variável resposta que é explicada pelo modelo é dada pelo coeficiente de determinação R^2. Sabemos ainda que, numa regressão linear simples, esse valor é o quadrado do coeficiente de correlação entre a variáveis resposta e o preditor. Logo, para a regressão linear de `taninos` sobre `polifenois`, tem-se $R^2=0.486^2=0.236196$, ou seja, a regressão explicará menos de 24% da variabilidade observada no teor de taninos. Trata-se dum valor bastante baixo, e muito pouco satisfatório.

   (b) Eis o teste de ajustamento global pedido:

   Hipóteses: $H_0 : \mathcal{R}^2=0$ *vs.* $H_1 : \mathcal{R}^2>0$, onde $\mathcal{R}^2$ indica o coeficiente de determinação populacional.

   Estatística do Teste: $F \;=\; (n-2)\,\frac{R^2}{1-R^2} \;\cap\; F_{(1,n-2)}$, sob H_0.

   Nível de significância: $\alpha=$ P[Erro do tipo I] = P[Rej. H_0 | H_0 verdade] $=0.05$.

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc}>f_{0.05[1,28]}=4.20$.

   Conclusões: Tem-se $F_{calc}=28\times\frac{0.236196}{1-0.236196}=8.658619$. Logo, rejeita-se H_0. Este resultado pode parecer surpreendente, dado o baixo valor do coeficiente de determinação amostral, mas sublinha a ideia frequentemente repetida de que o teste de ajustamento global compara a nossa regressão linear simples com o Modelo Nulo onde não existe qualquer preditor. O resultado do teste de ajustamento global permite afirmar que o ajustamento do modelo é significativamente melhor (ao nível $\alpha=0.05$) que o do Modelo Nulo. Mas essa afirmação não equivale a dizer que seja um bom modelo.

   (c) Sabemos que a recta de regressão ajustada tem equação $y=b_0+b_1\,x$, onde $b_1=\frac{cov_{xy}}{s^2_x}=r_{xy}\frac{s_y}{s_x}$ e $b_0=\overline{y}-b_1\overline{x}$. No enunciado são dados os desvios padrão e as médias das duas variáveis, bem como o coeficiente de correlação entre elas. Substituindo nas fórmulas, tem-se $b_1=0.486\times\frac{0.78275}{17.92042}=0.0212281$ e $b_0=6.94150-0.0212281\times 251.29287=1.607029$. Logo, a

recta de regressão ajustada tem equação $y = 1.607029 + 0.0212281\,x$. Pode interpretar-se este valor do declive afirmando que, por cada unidade a mais no Índice de Folin com o qual se mediram os polifenóis totais, é esperado um aumento de 0.0212281 g l^{-1} no teor de taninos.

(d) É pedido um intervalo a 95% de confiança para $E[Y|X = \overline{x} = 251.29287]$, onde Y indica o teor de taninos e X o teor de polifenóis total. A expressão para este tipo de intervalos de confiança corresponde à dos intervalos de predição para Y (disponíveis no formulário), mas sem a primeira parcela debaixo da raiz quadrada (o "1+"). Além disso, como o valor de X para o qual é pedido o IC é igual ao valor médio nessa variável preditora, a última parcela debaixo da raiz quadrada anula-se, ficando apenas o intervalo:

$$\left] \, (b_0 + b_1\,x) - t_{\frac{\alpha}{2}(n-2)} \sqrt{\frac{QMRE}{n}} \quad , \quad (b_0 + b_1\,x) + t_{\frac{\alpha}{2}(n-2)} \sqrt{\frac{QMRE}{n}} \, \right[$$

Já foram indicados na alínea anterior os valores de b_0 e b_1. Tem-se $n = 30$, pelo que $t_{0.025(28)} = 2.048$. Finalmente, é necessário o valor de $QMRE = \frac{SQRE}{n-2}$. Ora, sabemos que $SQT = (n-1)\,s_y^2 = 29 \times 0.78275^2 = 17.76823$. Assim, $SQRE = SQT - SQR = SQT\,(1-R^2) = 17.76823 \times (1 - 0.236196) = 13.57144$. Logo, $QMRE = \frac{13.57144}{28} = 0.4846945$. Substituindo estes valores na expressão acima, tem-se o intervalo] 6.681182 , 7.201817 [. Trata-se dum intervalo de amplitude relativamente pequena que podemos afirmar, com 95% de confiança, contém o teor médio em taninos, para plantas com índice 251.29287 de polifenóis totais.

2. Neste ponto estuda-se a regressão linear múltipla com $p = 7$ preditores.

(a) A afirmação não é verdadeira. O coeficiente estimado do preditor `polifenois` é $b_2 = 0.055241$. Em termos estatísticos, não é possível afirmar se o valor "*é quase nulo*" e se afecta ou não de forma significativa o ajustamento, sem estudar se o parâmetro populacional estimado por b_2, ou seja, β_2, pode ser considerado nulo. Essa possibilidade pode ser estudada através dum teste t bilateral, com a hipótese nula $H_0 : \beta_2 = 0$. No enunciado tem-se a informação necessária para concluir que essa hipótese nula deve ser rejeitada: o valor calculado da estatística desse teste ($t_{calc} = 17.677$) tem um valor de prova (*p-value*) muito próximo de zero ($p = 1.73 \times 10^{-14}$), pelo que há uma claríssima rejeição da hipótese nula $H_0 : \beta_2 = 0$, em favor da hipótese alternativa $H_0 : \beta_2 \neq 0$. Tendo em conta que a estatística de teste é dada pela razão entre a estimativa e o correspondente erro padrão, ou seja, que $T_{calc} = \frac{b_2 - 0}{\hat{\sigma}_{\hat{\beta}_2}}$, é fácil perceber que o facto de o erro padrão $\hat{\sigma}_{\hat{\beta}_2} = 0.003125$ ser muito mais pequeno do que a estimativa b_2 é decisivo para que *não* se possa considerar a estimativa $b_2 = 0.055241$ como sendo "*quase nula*". Tenha-se ainda em conta que os valores da variável `polifenois`, que são multiplicados por b_2 oscilam entre um mínimo de 217.125 e 293.106, pelo que as parcelas $b_2\,x_2$ estarão entre 12 e 16, o que representa uma contribuição importante para os valores ajustados da variável resposta `taninos`.

(b) A questão nesta alínea tem pontos de contacto com a da alínea anterior: o preditor cuja exclusão menos afectará o ajustamento será o preditor ao qual corresponda, num teste às hipóteses $H_0 : \beta_j = 0$ vs. $H_0 : \beta_j \neq 0$, o maior valor de prova. Essa exclusão não alterará o ajustamento de forma signficativa, ao nível α, caso esse valor de prova exceda α. É evidente que nesta regressão linear, o preditor `rend`, cujo valor de prova é $p = 0.8419$, não provoca uma diminuição significativa da qualidade de ajustamento para nenhum dos níveis de significância habituais. Pode calcular-se o coeficiente de determinação do submodelo com $p - 1 = 6$ preditores, resultante de excluir o preditor `rend`, recordando que o quadrado

da estatística t, no teste bilateral a $\beta_7 = 0$ é a estatística do teste F parcial que compara o modelo completo original e o submodelo resultante da exclusão do preditor `rend`. Assim, tem-se:

$$\begin{aligned} T^2_{calc} \;=\; F_{calc} \;&\Leftrightarrow\; (-0.202)^2 \;=\; \frac{n-(p+1)}{p-(p-1)}\,\frac{R^2_c - R^2_s}{1-R^2_c} \\ &\Leftrightarrow\; 0.040804 \;=\; (30-8)\,\frac{0.9443 - R^2_s}{1-0.9443} \\ &\Leftrightarrow\; R^2_s \;=\; 0.9443 - 0.0001033083 \;=\; 0.9441967 \end{aligned}$$

Assim, o modelo resultante da exclusão do preditor `rend` apenas explica menos cerca de 0.01% da variabilidade observada no teor de taninos.

(c) Considera-se agora um modelo com apenas $k = 2$ preditores, `ant` e `polifenois`, cujo coeficiente de determinação é ainda bastante elevado: $R^2 = 0.9261$.

i. Pede-se um teste F parcial para comparar o submodelo com o modelo completo original, de $p = 7$ preditores. Tem-se:

Hipóteses: $H_0 : \mathcal{R}^2_c = \mathcal{R}^2_s \quad vs. \quad H_1 : \mathcal{R}^2_c > \mathcal{R}^2_s$, onde $\mathcal{R}^2_c$ e $\mathcal{R}^2_s$ indicam os coeficientes de determinação populacional, respectivamente do modelo completo e do submodelo.

Estatística do Teste: $F \;=\; \frac{n-(p+1)}{p-k}\,\frac{R^2_c - R^2_s}{1-R^2_c} \;\cap\; F_{(p-k,n-(p+1))}$, sob H_0.

Nível de significância: $\alpha = \mathrm{P}[\text{ Erro do tipo I }] = \mathrm{P}[\text{ Rej. } H_0 \mid H_0 \text{ verdade }] = 0.05$.

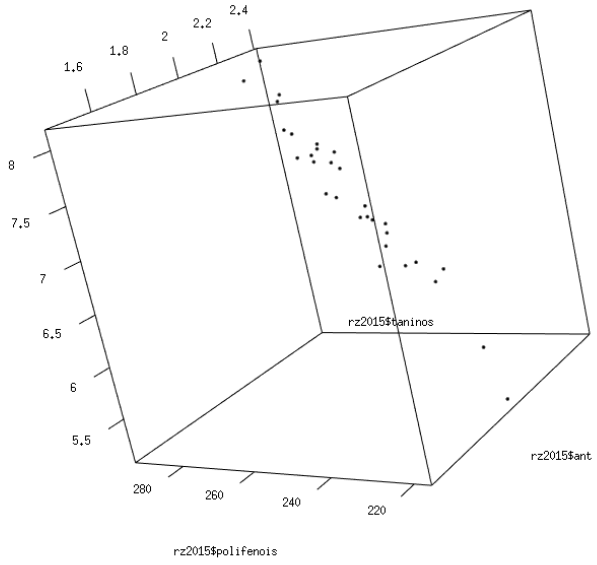
Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05[5,22]} = 2.66$.

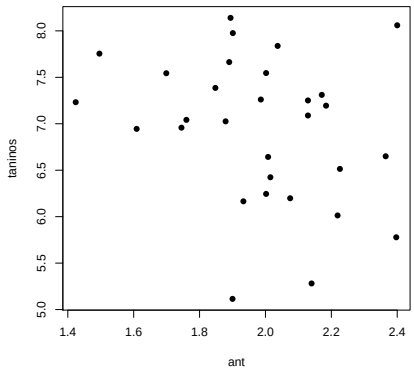
Conclusões: Tem-se $F_{calc} = \frac{22}{5}\,\frac{0.9443-0.9261}{1-0.9443} = 1.437702$. Logo, não se rejeita H_0, i.e., considera-se que a qualidade de ajustamento do modelo completo não é significativamente melhor (ao nível $\alpha = 0.05$) do que a do submodelo, facto compatível com uma diferença de menos de 0.02 no valor dos coeficientes de determinação. Assim, o mais parcimonioso submodelo parece ser a opção mais adequada.

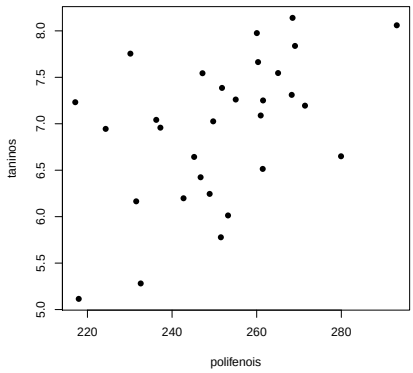
ii. Como se viu acima, na regressão linear simples de `taninos` sobre `polifenois`, o coeficiente de determinação é de apenas $R^2 = 0.236196$. A regressão linear simples de `taninos` sobre `ant` tem um coeficiente de determinação bastante mais baixo: $R^2 = (r_{ant,taninos})^2 = (-0.266)^2 = 0.070756$. No entanto, à regressão linear múltipla de `taninos` sobre os dois preditores referidos corresponde o valor muito apreciável de $R^2 = 0.9261$. Este exemplo ilustra uma ideia importante: *não é possível prever a qualidade dum modelo de regressão linear múltipla apenas com base nos coeficientes de correlação entre cada preditor individual e a variável resposta.*

iii. Podemos representar as $n = 30$ plantas observadas, como uma nuvem de 30 pontos no espaço a 3 dimensões cujos eixos correspondem às três variáveis associadas a este modelo. A boa qualidade do modelo de regressão linear de `taninos` sobre os preditores `ant` e `polifenois` significa que os n pontos da nuvem se dispersam essencialmente em torno dum plano de equação $y = 0.269530 - 3.706825\,x_1 + 0.055789\,x_2$ (onde y indica `taninos`, x_1 indica `ant` e x_2 indica `polifenois`). No entanto, a projecção da nuvem de pontos, quer sobre o plano coordenado $x_1 0 y$, quer sobre o plano coordenado $x_2 0 y$ não produz uma nuvem de pontos de forma (sequer aproximadamente) linear, dados os baixos valores dos coeficientes de determinação dos respectivos modelos. Isto significa que a inclinação do plano em $\mathbb{R}^3$ inicialmente referido não se alinha com os planos

coordenados referidos. Nas figuras que se seguem, pode ver-se o gráfico de 30 pontos em $\mathbb{R}^3$, com uma inclinação do sistema de eixos que evidencia que a nuvem se dispõe em torno dum plano. Seguidamente, veêm-se as nuvens dos mesmos 30 pontos nos planos coordenados definidos pela variável resposta `taninos` e cada um dos preditores, podendo confirmar-se os baixos valores de R^2 já calculados.







III

1. Trata-se dum delineamento factorial a dois factores, sendo a variável resposta Y a altura aos dois anos (em cm); o primeiro factor (A) a proveniência, com $a = 5$ níveis e o segundo factor (B) o local do estudo (com $b = 2$ níveis). O delineamento é equilibrado, uma vez que em cada uma das $ab = 10$ células (situações experimentais) existem $n_c = 6$ observações, num total de $n = n_c ab = 60$ observações. Uma vez que existem repetições nas células, é possível (e desejável) estudar a existência de eventuais efeitos de interacção.

   O modelo ajustado é o modelo ANOVA a dois factores, com efeitos de interacção. Admite-se que os níveis de cada factor estão ordenados pela ordem com que aparecem no enunciado.

- $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, para qualquer $i = 1, 2, 3, 4, 5$, $j = 1, 2$ e $k = 1, 2, 3, 4, 5, 6$, sendo μ_{11} a altura esperada (aos dois anos) dos pinheiros gregos em Sines; α_i o efeito principal (acréscimo à altura) associado à proveniência i (com $\alpha_1 = 0$); β_j o efeito principal (acréscimo à altura) associado a $j = 2$ (dada a restrição $\beta_1 = 0$); $(\alpha\beta)_{ij}$ o efeito de interacção, isto é, o acréscimo na altura específico da combinação da proveniência i com o local j. Dadas as restrições $(\alpha\beta)_{ij} = 0$ se $i = 1$ e/ou $j = 1$), o modelo apenas prevê efeitos de interacção nas situações experimentais correspondentes a Tavira ($j = 2$) e para proveniências diferentes da Grécia ($i > 1$). Finalmente ϵ_{ijk} é o erro aleatório da observação Y_{ijk}.
- Admite-se que os erros aleatórios são todos Normais, de média zero e variâncias homogéneas: $\epsilon_{ijk} \cap \mathcal{N}(0, \sigma^2)$, para qualquer i, j, k.
- Admite-se que os erros aleatórios ϵ_{ijk} são independentes.

2. Tratando-se dum modelo ANOVA factorial, a dois factores com interacção, a tabela-resumo terá de ter quatro linhas, correspondentes aos três tipos de efeitos previstos (principal de cada factor e de interacção), bem como à variabilidade residual e, opcionalmente, uma quinta linha associada à variabilidade total. A tabela terá as habituais colunas de graus de liberdade, Somas de Quadrados, Quadrados Médios e valor das estatísticas F. Vejamos como se pode preencher esta tabela.

   Sabemos que os graus de liberdade associados a $QMRE$ são dados por $n - ab$, onde $n = 60$ é o número total de observações e $ab = 10$ é o número de parâmetros existentes no modelo. Assim, $g.l.(SQRE) = 50$. Sabemos ainda que, para os vários tipos de efeitos, os graus de liberdade são dados pelo número de parcelas de cada tipo de efeito, após a introdução das restrições, ou seja, associado a SQA há $a - 1 = 4$ g.l., associado a SQB há $b - 1 = 1$ g.l., e associado a $SQAB$ há $(a-1)(b-1) = 4$ graus de liberdade.

   No enunciado é dada a Soma de Quadrados associada ao que foi designado factor A, tendo-se $SQA = 280.61$, donde se conclui que $QMA = \frac{SQA}{a-1} = \frac{280.61}{4} = 70.1525$. No enunciado é também dado o Quadrado Médio Residual, tendo-se $QMRE = 16.59$, donde $SQRE = QMRE \times (n - ab) = 16.59 \times 50 = 829.50$. Ora, sabemos pelo formulário que:

$$\begin{aligned} SQB &= a\, n_c \sum_{j=1}^{2} (\overline{y}_{.j.} - \overline{y}_{...})^2 \\ &= 5 \times 6 \times [(28.14 - 31.76298)^2 + (35.38 - 31.76298)^2] = 786.2645 \ . \end{aligned}$$

   Donde $QMB = \frac{SQB}{b-1} = 786.2645$. O enunciado refere ainda a variância da totalidade das 60 observações, $s_y^2 = 34.49584$, donde se pode concluir que a Soma de Quadrados Total é $SQT = (n-1)\, s_y^2 = 59 \times 34.49584 = 2035.255$. Uma vez que sabemos que esta Soma de Quadrados Total se pode decompor como $SQT = SQA + SQB + SQAB + SQRE$, torna-se possível calcular $SQAB = SQT - (SQA + SQB + SQRE) = 2035.255 - (280.61 + 786.2645 + 829.50) = 138.8801$. Assim, o Quadrado Médio associado à interacção é dado por $QMAB = \frac{SQAB}{(a-1)(b-1)} = \frac{138.8801}{4} = 34.7200$.

   Finalmente, os valores das estatísticas F são dados, para os três tipos de efeitos, pela razão entre o Quadrado Médio do referido tipo de efeito e $QMRE$. A tabela completa fica assim:

	g.l.	Soma de Quadrados	Quadrado Médio	F
Proveniência	4	280.61	70.1525	4.229
Local	1	786.2645	786.2645	47.394
Interacção	4	138.8801	34.7200	2.093
Residual	50	829.50	16.59	–

3. Vai-se efectuar em pormenor o teste aos efeitos principais do Factor A (proveniência dos pinheiros), e descrever sinteticamente os testes aos efeitos principais do Factor B (local) e aos efeitos de interacção.

   Hipóteses: $H_0 : \alpha_i = 0 \ , \ \forall \ i$ vs. $H_1 : \exists\, i$ tal que $\alpha_i \neq 0$.

   Estatística do Teste: $F_A = \frac{QMA}{QMRE} \cap F_{[a-1,n-ab]}$, sob H_0.

   Nível de significância: $\alpha = 0.05$.

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(4,50)} \approx 2.57$ (entre os valores tabelados 2.53 e 2.61).

   Conclusões: Como $F_{calc} = \frac{QMA}{QMRE} = 4.229 > 2.57$, rejeita-se H_0, sendo possível concluir pela existência de efeitos principais de proveniência (ao nível $\alpha = 0.05$).

   No teste aos efeitos principais do factor local do estudo, as hipóteses do teste podem ser escritas apenas como $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$, uma vez que após a imposição da restrição $\beta_1 = 0$, apenas sobra um efeito deste tipo, o efeito β_2 associado a Tavira. O valor calculado da estatística de teste é muito grande ($F_{calc} = 47.394$) deixando antever a rejeição de H_0, facto que é confirmado determinando nas tabelas o limiar da região crítica unilateral direita: $f_{0.05(1,50)} \approx 4.04$ (entre os valores tabelados 4.00 e 4.08). Assim, conclui-se claramente pela existência de efeitos principais de localidade, o que neste caso significa que existe um efeito associado à passagem do local de plantação de Sines para Tavira. Uma rápida inspecção das médias de local sugere que se trata dum maior crescimento dos pinheiros em Tavira, pelo que se deduz que β_2 terá um valor positivo.

   No teste aos efeitos de interacção, com hipóteses $H_0 : (\alpha\beta)_{ij} = 0$, para todo o i e j, contra a hipótese alternativa de que existe pelo menos uma célula (i, j) onde $(\alpha\beta)_{ij} \neq 0$, o valor calculado da estatística de teste é $F_{calc} = 2.093$, inferior ao limiar da região crítica, que é (por coincidência) igual ao do teste aos efeitos do factor A $f_{0.05(4,50)} \approx 2.57$. Logo, não se rejeita H_0 (para $\alpha = 0.05$), e conclui-se pela inexistência de efeitos significativos de interacção.

4. Nesta alínea é pedido para verificar se o facto da maior altura média amostral de Sines (31.16, para pinheiros provenientes de Marrocos) ser menor que a mais baixa altura média amostral em Tavira (33.56, para pinheiros da segunda proveniência italiana) é uma relação que se possa estender à população. Vamos responder efectuando, como solicitado no enunciado, um teste de Tukey, e usando $\alpha = 0.05$. Ora, o termo de comparação é (como indicado no formulário e usando as tabelas da distribuição de Tukey):

$$q_{\alpha(ab,n-ab)} \sqrt{\frac{QMRE}{n_c}} = q_{0.05(10,50)} \sqrt{\frac{16.59}{6}} = 4.68 \times 1.662829 = 7.782039 \ .$$

   Ora, a diferença entre as médias amostrais das duas células referidas acima é apenas $|31.16 - 33.56| = 2.40$, logo inferior ao termo de comparação, pelo que não é uma diferença significativa (ao nível $\alpha = 0.05$). Assim, não é possível afirmar que as médias populacionais em Tavira sejam sempre maiores às de Sines, independentemente das proveniências. Alguns pares de médias populacionais podem ser consideradas diferentes (por exemplo, o crescimento médio dos pinheiros gregos em Sines e em Tavira), mas será preciso levar em conta as proveniências, e não apenas o local da realização do estudo.

IV

1. (a) A matriz $\mathbf{X}$ do modelo tem n linhas, correspondentes às n observações, e $p+1$ colunas, sendo a primeira uma coluna de uns e as restantes, as colunas das n observações em cada uma das p variáveis preditoras. Assim, tem-se:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1_{(1)}} & x_{2_{(1)}} & \cdots & x_{p_{(1)}} \\ 1 & x_{1_{(2)}} & x_{2_{(2)}} & \cdots & x_{p_{(2)}} \\ 1 & x_{1_{(3)}} & x_{2_{(3)}} & \cdots & x_{p_{(3)}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1_{(n)}} & x_{2_{(n)}} & \cdots & x_{p_{(n)}} \end{bmatrix}$$

Por definição, o espaço das colunas da matriz $\mathbf{X}$, $\mathcal{C}(\mathbf{X})$ é constituído por todos os vectores (necessariamente de $\mathbb{R}^n$) que se podem obter como combinações lineares das colunas da matriz $\mathbf{X}$.

(b) i. Sabemos que a matriz de projecção sobre o subespaço gerado pelo vector $\vec{\mathbf{1}}_n$ é dada por $\mathbf{P} = \vec{\mathbf{1}}_n(\vec{\mathbf{1}}_n^t\vec{\mathbf{1}}_n)^{-1}\vec{\mathbf{1}}_n^t = \frac{1}{n}\vec{\mathbf{1}}_n\vec{\mathbf{1}}_n^t$ (veja-se o Exercício 17 da Regressão Linear Múltipla). Ora, $(\mathbf{I}-\mathbf{P})\vec{\mathbf{y}} = \vec{\mathbf{y}} - \mathbf{P}\vec{\mathbf{y}}$, sendo a segunda parcela dada (ver ainda o Exercício 17) por $\mathbf{P}\vec{\mathbf{y}} = \frac{1}{n}\vec{\mathbf{1}}_n\vec{\mathbf{1}}_n^t\vec{\mathbf{y}} = \overline{y}\vec{\mathbf{1}}_n$, ou seja, o vector que repete n vezes a média das n observações de Y. Logo, tem-se:

$$(\mathbf{I}-\mathbf{P})\vec{\mathbf{y}} \;=\; \vec{\mathbf{y}} - \mathbf{P}\vec{\mathbf{y}} \;=\; \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \overline{y} \\ \overline{y} \\ \overline{y} \\ \vdots \\ \overline{y} \end{bmatrix} = \begin{bmatrix} y_1 - \overline{y} \\ y_2 - \overline{y} \\ y_3 - \overline{y} \\ \vdots \\ y_n - \overline{y} \end{bmatrix} \;=\; \vec{\mathbf{y}}^c$$

ii. A matriz de projecção ortogonal sobre $\mathcal{C}(\mathbf{X})$ é dada por $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, pelo que a projecção ortogonal do vector $\vec{\mathbf{y}}^c$ será:

$$\mathbf{H}\vec{\mathbf{y}}^c \;=\; \mathbf{H}(\mathbf{I}-\mathbf{P})\vec{\mathbf{y}} \;=\; (\mathbf{H}-\mathbf{HP})\vec{\mathbf{y}} \;=\; \mathbf{H}\vec{\mathbf{y}} - \mathbf{HP}\vec{\mathbf{y}}$$

Ora $\mathbf{H}\vec{\mathbf{y}}$ é o vector dos valores ajustados da variável resposta, ou seja, o vector $\hat{\vec{\mathbf{y}}}$ (razão pela qual a matriz $\mathbf{H}$ é conhecida em inglês pela "*hat matrix*", ou seja, a matriz que coloca o chapéu no vector $\vec{\mathbf{y}}$, e é representada pela primeira letra da palavra *hat*). Por outro lado, e como se viu nas aulas, o produto $\mathbf{HP} = \mathbf{H}\vec{\mathbf{1}}_n(\vec{\mathbf{1}}_n^t\vec{\mathbf{1}}_n)^{-1}\vec{\mathbf{1}}_n^t = \vec{\mathbf{1}}_n(\vec{\mathbf{1}}_n^t\vec{\mathbf{1}}_n)^{-1}\vec{\mathbf{1}}_n^t = \mathbf{P}$, uma vez que $\mathbf{H}\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n$, já que $\vec{\mathbf{1}}_n \in \mathcal{C}(\mathbf{X})$ e vectores que pertencem ao subespaço sobre o qual projecta $\mathbf{H}$ ficam invariantes na projecção (ver Exercício 4 da Regressão Linear Múltipla). Assim, a projecção ortogonal de $\vec{\mathbf{y}}^c$ sobre $\mathcal{C}(\mathbf{X})$ é um vector cujo elemento genérico é $\hat{y}_i - \overline{y}$. A norma ao quadrado desse vector é, por definição de SQR, a Soma de Quadrados associada à Regressão: $SQR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$.

iii. Já se viu que SQR é o quadrado do comprimento do vector projectado $\mathbf{H}\vec{\mathbf{y}}^c$. Pelo seu lado, $SQT = \sum_{i=1}^{n}(y_i - \overline{y})^2$ é o quadrado do comprimento do vector $\vec{\mathbf{y}}^c$, antes da projecção ortogonal sobre $\mathcal{C}(\mathbf{X})$. Essa projecção ortogonal cria um triângulo rectângulo, no qual $\vec{\mathbf{y}}^c$ é a hipotenusa e $\mathbf{H}\vec{\mathbf{y}}^c$ é um dos catetos. Ora o comprimento dum cateto é sempre menor do que o comprimento da hipotenusa. No limite, poderia ser igual se o

triângulo colapsasse num segmento de recta, o que corresponde a dizer que a hipotenusa pertenceria ao espaço $\mathcal{C}(\mathbf{X})$ e coincidiria com o cateto. Nesse caso, $\vec{\mathbf{y}}^c = \mathbf{H}\vec{\mathbf{y}}^c$ uma vez que $\vec{\mathbf{y}}^c$ ficaria invariante na projecção. Esta igualdade só é possível se $y_i = \hat{y}_i$, para todos os elementos do vector, ou seja, se os valores observados e ajustados coincidem. Nesse caso, todos os resíduos são nulos.

2. (a) Sabemos que um modelo (com a mesma variável resposta e ajustado com base nos mesmos dados) é melhor do que outro se o seu AIC fôr menor. Designando os AICs do modelo com k preditores e do submodelo com menos um preditor por, respectivamente, AIC_k e AIC_{k-1}, e tendo em conta a fórmula do AIC para modelos de regressão linear (disponível no formulário), tem-se que o submodelo é considerado preferível se:

$$\begin{aligned}
AIC_k > AIC_{k-1} &\Leftrightarrow n\,\ln\left(\frac{SQRE_k}{n}\right) + 2(k+1) > n\,\ln\left(\frac{SQRE_{k-1}}{n}\right) + 2k \\
&\Leftrightarrow 2 > n\left[\ln\left(\frac{SQRE_{k-1}}{n}\right) - \ln\left(\frac{SQRE_k}{n}\right)\right] \\
&\Leftrightarrow \frac{2}{n} > \ln\left(\frac{SQRE_{k-1}}{SQRE_k}\right) \\
&\Leftrightarrow \mathrm{e}^{\frac{2}{n}} > \frac{SQRE_{k-1}}{SQRE_k} \\
&\Leftrightarrow \mathrm{e}^{\frac{2}{n}}\,SQRE_k > SQRE_{k-1}
\end{aligned}$$

como se pedia para provar. Repare-se que, em limite, quando n tende para infinito, a condição de exclusão tende para uma condição impossível: $SQRE_k > SQRE_{k-1}$. Uma forma alternativa de pensar no problema é dizer que a Soma de Quadrados Residual do submodelo (que é necessariamente maior ou igual à do modelo) vai tendo uma margem cada vez menor para cumprir a relação $\mathrm{e}^{\frac{2}{n}}\,SQRE_k > SQRE_{k-1}$ à medida que n aumenta. Assim, para grandes amostras o algoritmo de exclusão sequencial tende a não simplificar os modelos.

(b) A partir dum modelo com k preditores, o algoritmo de exclusão sequencial baseado no AIC procura o submodelo com $k-1$ preditores de menor AIC. Uma vez que todos esses submodelos têm o mesmo número de preditores ($k-1$), a definição de AIC (ver formulário) desses vários submodelos apenas difere na primeira parcela, e o menor AIC corresponderá ao submodelo com o menor valor de Soma de Quadrados Residual. Por outro lado, o algoritmo de exclusão sequencial baseado nos testes t aos parâmetros β_j irá excluir o preditor para o qual o valor da estatística t é, em módulo, mais baixo (a que corresponderá o maior *p-value*). Sabemos também que os valores de t_{calc} nos testes t a $H_0 : \beta_j = 0$, quando elevados ao quadrado, são o valor da estatística do teste F parcial que compara o modelo de k preditores e cada um desses submodelos de $k-1$ preditores. Logo, ao menor valor de $|t_{calc}|$ corresponde o menor valor de F_{calc}. Mas, para os testes F parciais em causa, tem-se (ver formulário):

$$F_{calc} = \frac{n-(k+1)}{1}\,\frac{SQRE_{k-1} - SQRE_k}{SQRE_k}\,.$$

Os vários submodelos de $k-1$ preditores apenas diferem na primeira parcela do numerador. Logo, o menor valor destas estatísticas F tem de corresponder ao submodelo com o menor valor de $SQRE_{k-1}$, que é também o modelo escolhido no algoritmo de exclusão sequencial baseado no AIC.