

INSTITUTO SUPERIOR DE AGRONOMIA  
ESTATÍSTICA E DELINEAMENTO

14 de Janeiro, 2015 1ª CHAMADA de EXAME 2014-15 Uma resolução possível

I

É apresentada uma tabela de contingência (tabela de contagens de dupla entrada), com contagens relativas ao cruzamento de dois factores: o factor proveniência (factor A, com  $a = 6$  categorias) e o factor adaptabilidade (factor B, com  $b = 2$  categorias).

1. Nesta tabela foram pré-fixados os totais marginais de linha ( $N_i = 150, \forall i$ ), sendo perguntado se a distribuição destas 150 contagens pelas duas categorias do outro factor se pode considerar homogénea. Tendo em conta que apenas existem duas categorias no factor adaptabilidade, as hipóteses deste teste de homogeneidade podem ser escritas apenas em termos da probabilidade de morte em cada linha. Designe-se por  $\pi_i$  a probabilidade duma observação corresponder a uma árvore da proveniência  $i$  morrer antes dos dois anos. Tem-se então:

**Hipóteses:**  $H_0 : \pi_i = \pi, \forall i$  vs.  $H_1 : \exists i, j$  tais que  $\pi_i \neq \pi_j$   
[Homogeneidade] [Heterogeneidade]

**Estatística do Teste:**  $\chi^2 = \sum_{i=1}^a \sum_{j=1}^b \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \sim \chi_{(a-1)(b-1)}^2$ , sob  $H_0$ ,

com  $a = 6, b = 2, O_{ij}$  o número de observações na célula  $(i, j)$  e  $\hat{E}_{ij}$  os valores esperados estimados correspondentes, obtidos admitindo a hipótese nula de homogeneidade. A distribuição é apenas assintótica, mas a sua validade pode ser admitida caso se verifiquem as condições de Cochran (nenhuma célula ter valor esperado  $\hat{E}_{ij}$  inferior a 1 e não mais de 20% das células terem valor esperado inferior a 5). Neste caso não é difícil verificar a validade das condições de Cochran. De facto, aplicando a fórmula de cálculo do número esperado estimado de observações num teste de homogeneidade ao nosso contexto, tem-se:  $\hat{E}_{ij} = \frac{N_i \times N_{.j}}{N} = \frac{150 \times N_{.j}}{900} = \frac{N_{.j}}{6}$ . Assim, o número esperado estimado de observações é igual para todas as células duma mesma coluna. O menor valor esperado tem de corresponder à coluna com menos observações, ou seja, à primeira coluna ( $j = 1$ ), correspondente ao número de pinheiros mortos ao fim de 2 anos. Tem-se, para qualquer proveniência  $i, \hat{E}_{i1} = 35.5$ . Assim, nenhum valor esperado estimado é inferior a 5, pelo que se verificam as condições de Cochran e é admissível a aproximação assintótica na distribuição da estatística de Pearson.

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

**Região Crítica:** (Unilateral direita) Rejeitar  $H_0$  se  $\chi_{\text{calc}}^2 > \chi_{\alpha(5 \times 1)}^2 = \chi_{0.05(5)}^2 = 11.070$ .

**Conclusão:** Tendo em conta o valor calculado da estatística do teste, que é dado no enunciado, tem-se  $\chi_{\text{calc}}^2 = 51.2762 \gg 11.070$ . Logo, rejeita-se  $H_0$ , ou seja, rejeita-se a hipótese da probabilidade de mortalidade ser igual para todas as proveniências.

2. Já se viu que o valor esperado estimado para qualquer célula na primeira coluna é  $\hat{E}_{i1} = 35.5$ . Analogamente, todas as células da segunda coluna (nível de adaptabilidade  $j = 2$ , indicando sobrevivência) têm igual valor esperado estimado:  $\hat{E}_{i2} = \frac{N_i \times N_{.2}}{N} = \frac{150 \times 687}{900} = \frac{687}{6} = 114.5$ .

Logo, a contribuição de qualquer proveniência  $i = 1, \dots, 6$  para o valor calculado da estatística de Pearson é dado pela soma de duas parcelas do tipo:

$$\frac{(O_{i1} - \hat{E}_{i1})^2}{\hat{E}_{i1}} + \frac{(O_{i2} - \hat{E}_{i2})^2}{\hat{E}_{i2}} = \frac{(O_{i1} - 35.5)^2}{35.5} + \frac{(O_{i2} - 114.5)^2}{114.5}.$$

A identificação de qual a proveniência  $i$  que dará o maior valor à soma destas duas parcelas é fácil. Trata-se de identificar a proveniência cujos valores observados sejam, simultaneamente, mais distantes de 35.5 na coluna de pinheiros mortos, e de 114.5 na coluna de pinheiros sobreviventes (a mesma localidade tem de satisfazer estes dois critérios já que a soma, quer dos pinheiros observados, quer esperados, em cada linha tem de ser sempre igual a  $N_{i.} = 150$ ). Uma inspeção visual da tabela confirma tratar-se da Turquia, cujos pinheiros registaram uma mortalidade muito mais elevada do que as restantes (quase o dobro do valor esperado estimado) e para a qual o valor da soma destas duas parcelas é:

$$\frac{(67 - 35.5)^2}{35.5} + \frac{(83 - 114.5)^2}{114.5} = 37.01594 .$$

Este valor corresponde a mais de 72% do valor calculado da estatística ( $X_{calc}^2 = 51.2762$ ), o que confirma tratar-se da proveniência que (de longe) mais contribui para o valor de  $X_{calc}^2$  e, por conseguinte, para a rejeição da hipótese nula de homogeneidade.

## II

1. O valor  $R^2 = 0.5281$  significa que esta regressão linear múltipla explica quase 53% da variabilidade dos teores de hexanal observados. Não é um valor muito bom, tratando-se duma regressão de qualidade modesta.
2. A qualidade modesta da regressão não significa necessariamente que o modelo não seja significativamente diferente do modelo nulo, e é isso que se pede para verificar nesta alínea. Tem-se:

**Hipóteses:**  $H_0 : \mathcal{R}^2 = 0$  vs.  $H_1 : \mathcal{R}^2 > 0$ .

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

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

**Região Crítica:** (Unilateral direita) Rejeitar  $H_0$  se  $F_{calc} > f_{0.05[8,30]} = 2.27$ .

**Conclusões:** Pelo enunciado, sabemos que  $F_{calc} = 4.197 > 2.27$ . Logo rejeita-se  $H_0$ , i.e., apesar do valor não muito elevado do coeficiente de determinação, a regressão ajustada é (ao nível  $\alpha = 0.05$ ) significativamente diferente do modelo nulo.

3. Pede-se um teste de hipóteses com as seguintes hipóteses (uma vez que o ónus da prova recai sobre a afirmação do enunciado):

**Hipóteses:**  $H_0 : \beta_7 \geq 0.005$  vs.  $H_1 : \beta_7 < 0.005$ .

**Estatística do Teste:**  $T = \frac{\hat{\beta}_7 - \beta_{7|H_0}}{\hat{\sigma}_{\hat{\beta}_7}} \cap t_{n-(p+1)}$ , sob  $H_0$ .

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

**Região Crítica:** (unilateral esquerda) Rejeitar  $H_0$  se  $T_{calc} < -t_{0.05(30)} = -1.69726$ .

**Conclusões:** Tem-se  $T_{calc} = \frac{0.0038168-0.005}{0.0327531} = -0.036125$ . Este valor não pertence à região crítica, logo não se rejeita a hipótese nula  $\beta_7 \geq 0.005$ , não sendo possível concluir pela afirmação do enunciado.

4. (a) O coeficiente de determinação de um dado modelo tem de ser sempre maior ou igual ao valor de  $R^2$  de qualquer seu submodelo, pelo que tem de verificar-se  $R_s^2 \leq 0.5281$  (indicando por  $R_s^2$  o valor de coeficiente de determinação do submodelo com dois preditores referido no enunciado). Por outro lado, e pela mesma razão,  $R_s^2$  terá de ser maior ou igual ao valor do coeficiente de determinação das regressões lineares *simples*, quer de **hexanal** sobre **nerol**, quer de **hexanal** sobre **2feniletanol**. Tendo em conta que nas regressões lineares simples, os valores dos coeficientes de determinação são o quadrado dos coeficientes de correlação entre o preditor e a variável resposta, os valores disponíveis na matriz de correlações do enunciado permitem concluir que  $R_s^2 \geq (-0.165)^2 = 0.027225$ , mas também que  $R_s^2 \geq (0.692)^2 = 0.478864$ . Assim, e como afirma o enunciado, tem de ter-se  $0.4788 < R_s^2 \leq 0.5281$ .
- (b) Sabemos que no submodelo com dois preditores se tem  $AIC = -222.23$ . Pela definição do  $AIC$ , tem-se:

$$\begin{aligned} AIC &= n \ln \left( \frac{SQRE_k}{n} \right) + 2(k+1) \\ \Leftrightarrow -222.23 &= 39 \ln \left( \frac{SQRE_2}{39} \right) + 2 \times 3 \\ \Leftrightarrow SQRE_2 &= 39 e^{\frac{-222.23-6}{39}} = 0.1120859 . \end{aligned}$$

Ora,  $R_s^2 = \frac{SQR_s}{SQT} = 1 - \frac{SQRE_2}{SQT}$ . Sabemos também que  $SQT = (n-1) s_y^2 = 38 \times 0.0061367 = 0.2331946$ . Logo,  $R_s^2 = 1 - \frac{0.1120859}{0.2331946} = 0.5193461$ .

- (c) Sabemos (e tendo em conta que neste submodelo há  $k = 2$  preditores) que:  $R_{mod}^2 = 1 - \frac{QMRE}{QMT} = 1 - \frac{SQRE/(n-(k+1))}{SQT/(n-1)} = 1 - \frac{0.1120859/36}{0.0061367} = 0.49264$ . Assim, a comparação pedida envolve quatro valores: os do modelo (original) completo, dados no enunciado,  $R^2 = 0.5281$  e  $R_{mod}^2 = 0.4023$ ; e os do submodelo de dois preditores entretanto calculados:  $R_s^2 = 0.5193461$  e  $R_{s_{mod}}^2 = 0.49264$ . Além do comentário sobre o valor modesto de todos estes indicadores, é também possível sublinhar a diferença considerável das duas medidas associadas ao modelo completo. O valor bastante menor do  $R^2$  modificado resulta da “penalização” que a respectiva definição impõe a modelos com muitos parâmetros (neste caso,  $p+1 = 9$ ), quando o número de observações não é muito superior (neste caso,  $n = 39$ ), sobretudo quando o coeficiente de determinação usual não é muito elevado (como é o caso). Esta penalização torna-se mais visível na expressão dada nas aulas para o  $R^2$  modificado,  $R_{mod}^2 = 1 - (1 - R^2) \frac{n-1}{n-(p+1)}$ , em que a parte da variabilidade não explicada (no modelo completo é  $1 - R^2 = 0.4719$ ) é multiplicada por um factor sempre maior que 1 (no modelo completo é  $\frac{n-1}{n-(p+1)} = \frac{38}{30} = 1.2667$ ), antes de ser de novo subtraída à unidade. Assim, no modelo completo a parte correspondente ao não ajustamento é agravada em cerca de 27% (elevando-a para cerca de 0.60, pelo que  $R_{mod}^2$  baixa para quase 0.40). No submodelo, e apesar dum valor do coeficiente de determinação usual bastante próximo do do modelo completo, existem apenas  $k = 2$  preditores, pelo que a penalização imposta pelo  $R_{mod}^2$  não é tão grande ( $\frac{n-1}{n-(k+1)} = \frac{38}{36} = 1.0556$ ). Assim, o facto de  $R_s^2 = 0.51935$  e  $R_{s_{mod}}^2 = 0.49264$  serem bastante mais próximos reflecte o facto de o submodelo ser bastante mais parcimonioso.

Deve também sublinhar-se que, ao contrário do que acontece com o  $R^2$  usual, que não pode aumentar num submodelo, o  $R^2$  modificado pode crescer quando se passa dum modelo

completo para um submodelo (dado o papel do número de preditores na definição de  $R_{mod}^2$ ). É o que se verifica neste exemplo, em que o valor de  $R_{mod}^2$  é bastante maior no submodelo do que no modelo completo. Mais uma vez, este crescimento do  $R^2$  modificado reflecte a maior parcimónia do submodelo, na obtenção dum coeficiente de determinação próximo do do modelo completo.

5. Nesta alínea, o preditor (teor de 2-feniletanol) é indicado por  $x$  e a variável resposta (teor de hexanal) é indicada por  $y$ .

- (a) Tem-se  $y = b_0 + b_1 x$ , com  $b_1 = \frac{cov_{xy}}{s_x^2} = r_{xy} \frac{s_y}{s_x} = 0.692 \times \sqrt{\frac{0.0061367}{0.0048344}} = 0.7796546$  e  $b_0 = \bar{y} - b_1 \bar{x} = 0.117162 - 0.7796546 \times 0.124846 = 0.01982524$ . Assim, a equação da recta de regressão pedida é  $y = 0.019825 + 0.77965 x$ .
- (b) No eixo horizontal tem-se os valores do efeito alavanca  $h_{ii}$  (*leverage*) de cada observação. Sabemos que estes valores têm de estar compreendidos entre 0 e 1, sendo o seu valor médio dado por  $\bar{h} = \frac{p+1}{n} = \frac{2}{39} = 0.051282$ . No nosso caso, existem apenas três observações acima da média, duas quais (as observações  $i = 4$  e  $i = 18$ ) com efeitos alavanca muito acima da média (próximos de 0.4). Sabemos que, uma vez que a variância dos resíduos (usuais) é dada por  $V[E_i] = \sigma^2(1 - h_{ii})$ , quanto maior o efeito alavanca, mais a observação correspondente tenderá a atrair a recta ajustada. As mesmas duas observações têm valores muito elevados dos resíduos (internamente) estandardizados ( $R_{18} > 4$  e  $R_4 \approx -3$ ), pelo que, apesar de estas observações atraírem para si a recta ajustada, ficam distantes dessa recta. Esta aparente contradição sugere a possibilidade de se tratar de duas observações distantes das restantes, logo com um impacto grande na definição da posição da recta ajustada, mas também distantes entre si e não colineares com o grupo das restantes, logo obrigando a recta a passar pelo meio delas. Tal como o sinal dos resíduos indica, as observações estão em posições diferentes face à recta: a observação 18, de resíduo positivo, está acima da recta e a observação 4, de resíduo negativo, está abaixo da recta. Outro aspecto a salientar no gráfico é o valor elevadíssimo das distâncias de Cook associadas a estas duas observações: claramente maiores do que 1. Tal facto indica que a exclusão de qualquer destas observações do conjunto de dados conduziria a grandes alterações na recta ajustada. Do conjunto dos aspectos referidos, ressalta que as observações  $i = 4$  e  $i = 18$  têm um impacto muito grande na regressão ajustada, facto que causa algum desconforto com o ajustamento do modelo e sugere uma exploração da natureza e possíveis causas dos valores associados a estas observações.
- (c) A exclusão das observações 4 e 18 revela que a relação linear nas restantes 37 observações é quase inexistente, tão fraca que uma regressão linear simples de **hexanal** sobre **2feniletanol**, baseada nessas 37 observações restantes tem um coeficiente de determinação  $R^2 = (0.21908)^2 = 0.048$ , ou seja, explicaria menos de 5% da variabilidade observada nos teores de **hexanal**. Seria útil acompanhar esta conclusão pela inspecção da nuvem de pontos relacionando as variáveis do modelo. Recorde-se o Exercício 9 de Regressão Linear Simples, discutido nas aulas práticas, com a relação entre log-peso do corpo e log-peso do cérebro em numerosas espécies animais. Nesse exercício, a exclusão das espécies de dinossauros tornava mais clara uma forte relação linear existente nas espécies contemporâneas. Aqui passa-se algo de sentido contrário: a relação linear simples agora ajustada tem um forte impacto de apenas duas observações, e está quase ausente nas restantes observações. Deve acrescentar-se que a natureza dessas duas observações (observações erradas ou legítimas? observações associadas a algum outro factor ou variável não controlada?) não pode ser conhecida com base na informação disponibilizada no enunciado. Mas o seu enorme impacto suscita fortes dúvidas, não apenas sobre a validade da regressão linear simples, como

também sobre a validade das próprias regressões múltiplas ajustadas nas alíneas anteriores (mais uma vez, sem que seja possível confirmar esta observação com base na informação disponível). Seria aconselhável estudar os gráficos de resíduos da regressão múltipla inicial. Pode tratar-se dum caso de relação espúria entre variáveis, eventualmente resultante de uma ou outra observação com valores errados.

### III

1. Trata-se dum delineamento a dois factores, o factor **variedade** (factor A, com  $a = 6$  níveis), e o factor **árvore** (factor B, com 2 níveis em todas as variedades). O objectivo do estudo é avaliar os eventuais efeitos destes factores sobre a variável resposta (largura das folhas). Pela própria natureza dos factores em questão, o delineamento deve ser considerado *hierarquizado*, com árvores subordinadas a variedades. Não faz sentido considerar o delineamento factorial pois cada árvore apenas pode pertencer a uma variedade. Assim, temos o factor subordinado árvores, com  $b_1 = \dots = b_6 = 2$  árvores para cada variedade. Ao todo há  $\sum_{i=1}^6 b_i = 12$  situações experimentais (as árvores individuais), e  $n_c = 10$  repetições em cada uma dessas situações experimentais, num total de  $n = 120$  folhas observadas. O modelo mais adequado será o modelo hierarquizado:

- $Y_{ijk} = \mu_{11} + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$ ,  $\forall i, j, k$ , onde  $Y_{ijk}$  indica a largura de cada folha observada ( $k = 1, 2, \dots, 10$ ) na árvore  $j$  ( $j = 1, 2$ ) da variedade  $i$  ( $i = 1, \dots, 6$ ). Impõem-se as restrições  $\alpha_1 = 0$ ,  $\beta_{1(i)} = 0$  para  $i = 1, \dots, 6$ . Com estas restrições, o parâmetro  $\mu_{11}$  é a largura média populacional das folhas da primeira árvore da primeira variedade (*azeiteira*);  $\alpha_2$  é o efeito nas larguras das folhas da segunda variedade (*blanqueta*); e por aí fora. Quanto aos efeitos de árvore, de que apenas sobra um em cada variedade,  $\beta_{2(i)}$ , trata-se do efeito nas larguras das folhas associado a passar da primeira para a segunda árvore estudada em cada variedade. O erro aleatório  $\epsilon_{ijk}$  associado à observação  $Y_{ijk}$  corresponde à variabilidade não explicada pelos efeitos previstos no modelo.
- $\epsilon_{ijk} \cap \mathcal{N}(0, \sigma^2)$ , para qualquer  $i, j, k$ .
- Os erros aleatórios  $\epsilon_{ijk}$  são independentes.

2. Pede-se um teste  $F$  aos efeitos do factor dominante (variedade), cuja hipótese nula corresponde à inexistência desse tipo de efeitos. Eis o teste detalhado:

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

**Estatística do Teste:**  $F_A = \frac{QMA}{QMRE} \cap F_{[(a-1), n-(b_1+\dots+b_6)]}$ , sob  $H_0$ .

**Nível de significância:**  $\alpha = 0.05$ .

**Região Crítica:** (Unilateral direita) Rejeitar  $H_0$  se  $F_{calc} > f_{0.01(5,108)} \approx 2.3$  (entre os valores tabelados 2.29 e 2.37).

**Conclusões:** Como  $F_{calc} = 14.494 > 2.37$ , rejeita-se  $H_0$ , o que corresponde a admitir a existência de efeitos de variedade (ao nível  $\alpha = 0.05$ ).

Assim, foi importante prever este tipo de efeitos no modelo. Pode concluir-se que a largura média das folhas sofre variações entre as seis variedades consideradas no estudo.

3. É pedido um teste de Tukey para comparar médias de algumas das situações experimentais (árvores), mais concretamente, para comparar a largura média das folhas entre as duas árvores

estudadas em cada variedade. Deverá considerar-se que as larguras médias populacionais de duas árvores quaisquer,  $\mu_{ij}$  e  $\mu_{i'j'}$ , são diferentes, caso as respectivas médias amostrais verifiquem a desigualdade:

$$|\bar{y}_{ij} - \bar{y}_{i'j'}| > q_{\alpha(\sum_{i=1}^6 b_i, n - \sum_{i=1}^6 b_i)} \sqrt{\frac{QMRE}{n_c}} = q_{0.05(12, 108)} \sqrt{\frac{2.91}{10}} \approx 4.72 \times 0.53944 = 2.546 .$$

Uma inspecção visual rápida revela que a única variedade onde talvez se verifique esta condição é a variedade *azeiteira*. No entanto,  $|\bar{y}_{11} - \bar{y}_{12}| = |12.865 - 10.721| = 2.144 < 2.546$ . Assim, e embora por pouco, nenhuma variedade revela diferenças significativas entre as larguras médias das folhas de cada árvore (ao nível  $\alpha = 0.05$ ). Esta conclusão, resultante da utilização do teste de Tukey, é coerente com o resultado dum teste  $F$  à existência de efeitos de árvore (factor subordinado). De facto, a hipótese nula dum tal teste  $F$  ( $H_0 : \beta_{2(i)} = 0, \forall i$ ), corresponde a afirmar que, para uma mesma variedade  $i$ , a largura média populacional das folhas das duas árvores seria sempre igual (já que  $\mu_{i1} = E[Y_{i1k}] = \alpha_i$  e  $\mu_{i2} = E[Y_{i2k}] = \alpha_i + \beta_{2(i)}$ ). A existência dum efeito de árvore não nulo (isto é,  $\beta_{2(i)} \neq 0$  para algum  $i$ ) é a hipótese alternativa do teste  $F$  e corresponde a dizer que, pelo menos numa variedade  $i$ ,  $\mu_{i1} \neq \mu_{i2}$ . O valor de prova ( $p$ -value) do teste  $F$  entre estas duas hipóteses é dado no enunciado  $p = 0.114$  e, sendo claramente superior aos níveis usuais de significância, conduz à não rejeição de  $H_0$ , ou seja, a uma conclusão análoga à do teste de Tukey: inexistência de efeitos de árvore dentro de cada variedade.

4. O teste de Bartlett visa estudar a hipótese de homogeneidade das variâncias dos erros aleatórios em cada situação experimental (cada uma das 12 árvores estudadas). Designando por  $\sigma_{ij}^2$  a variância populacional dos erros aleatórios associados à árvore  $j$  da variedade  $i$ , a hipótese nula será  $H_0 : \sigma_{ij}^2 = \sigma^2, \forall i, j$  (hipótese compatível com a homogeneidade de variâncias admitida nas ANOVAs), enquanto que a hipótese alternativa é  $H_1 : \exists i, j, i', j'$  tais que  $\sigma_{ij}^2 \neq \sigma_{i'j'}^2$  (hipótese que viola o pressuposto de homogeneidade das variâncias). A estatística deste teste tem uma expressão complicada (constante do formulário), mas o fundamental é saber que a sua distribuição assintótica ao abrigo da hipótese nula (válida porque  $n_c \geq 5$ ), é  $\chi_{12-1}^2$  e que a região de rejeição é unilateral direita. Desta forma, deve-se rejeitar  $H_0$  caso  $K_{calc}^2 > \chi_{0.05(11)}^2 = 19.675$ . No nosso caso,  $K_{calc}^2 = 18.1765$ , pelo que (embora por pouco) não se rejeita  $H_0$  ao nível  $\alpha = 0.05$ . Esta conclusão permite admitir o pressuposto de homogeneidade de variâncias do modelo ANOVA.

## IV

Este exercício é uma versão, ligeiramente adaptada, do Exercício 19 de Regressão Linear Múltipla.

1. A forma geral da matriz  $\mathbf{H}$  de projecção ortogonal sobre  $\mathcal{C}(\mathbf{X})$ , o espaço das colunas da matriz  $\mathbf{X}$  do modelo, é  $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ . Neste exercício, tem-se  $\mathbf{X} = \mathbf{1}_n$  (vector de  $n$  uns), pelo que a matriz de projecção ortogonal será  $\mathbf{H} = \mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}_n^t$ . Como sempre, trata-se duma matriz  $n \times n$ . Mas o facto de  $\mathbf{X} = \mathbf{1}_n$  ser um vector-coluna (uma matriz  $n \times 1$ ) de tipo especial permite calcular  $\mathbf{H}$ . O produto central  $(\mathbf{1}_n^t\mathbf{1}_n)$  é de dimensão  $1 \times 1$  (ou seja, é um escalar), e de cálculo fácil:

$$\mathbf{1}_n^t\mathbf{1}_n = [1 \quad 1 \quad \cdots \quad 1] \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = n .$$

A inversa desta “matriz de dimensão  $1 \times 1$ ” é o inverso multiplicativo do escalar (já que a matriz identidade  $1 \times 1$  é a matriz apenas com o elemento 1, logo se  $\mathbf{A} = n$ ,  $\mathbf{A}^{-1}$  será o número que multiplicado por  $n$  dá 1, ou seja, o inverso multiplicativo  $\frac{1}{n}$ ). Uma vez que num produto matricial uma constante se pode escrever em qualquer posição, temos  $\mathbf{H} = \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^t$ . Mas  $\mathbf{1}_n \mathbf{1}_n^t$  é uma matriz  $n \times n$  em que o elemento genérico  $(i, j)$  resulta do produto interno da linha  $i$  da matriz  $\mathbf{1}_n$  (ou seja, o escalar 1) pela coluna  $j$  da matriz  $\mathbf{1}_n^t$  (de novo, o escalar 1). Logo, todos os elementos de  $\mathbf{1}_n \mathbf{1}_n^t = \mathbf{J}$  são de valor 1, completando a justificação. (**Atenção:**  $\mathbf{1}_n^t \mathbf{1}_n \neq \mathbf{1}_n \mathbf{1}_n^t$ ! Nem sequer são matrizes da mesma dimensão).

2. Em geral, o vector de valores ajustados  $\hat{\mathbf{y}}$  é obtido pré-multiplicando o vector  $\mathbf{y}$  de observações da variável resposta, pela matriz de projecções ortogonais  $\mathbf{H}$ :  $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ . No nosso caso,  $\mathbf{H}$  é da forma indicada na alínea anterior, pelo que  $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} = \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^t \mathbf{y}$ . Dada a natureza especial do vector  $\mathbf{1}_n$ , o produto final é a soma dos valores observados  $y_i$ :

$$\mathbf{1}_n^t \mathbf{y} = [1 \ 1 \ \dots \ 1] \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^n y_i .$$

Substituindo na expressão para  $\hat{\mathbf{y}}$ , tem-se:  $\hat{\mathbf{y}} = \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^t \mathbf{y} = \bar{y} \mathbf{1}_n$ . Trata-se dum vector que repete  $n$  vezes a média das  $n$  observações de  $y$ , ou seja, tem-se  $\hat{y}_i = \bar{y}$ , para qualquer  $i = 1, \dots, n$ . Todos os valores ajustados de  $y$  no modelo nulo são dados pela média das  $n$  observações de  $y$ .

3. Em geral, o vector dos estimadores de mínimos quadrados dos  $p+1$  parâmetros do modelo é dado por  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y}$ . No nosso caso existe apenas um parâmetro do modelo ( $\beta_0$ ), pelo que o vector dos estimadores tem um único elemento (facto facilmente comprovável através das dimensões das matrizes envolvidas no produto que define  $\hat{\boldsymbol{\beta}}$ ). Tendo em conta a natureza especial da nossa matriz do modelo  $\mathbf{X} = \mathbf{1}_n$ , e as contas já efectuadas em alíneas anteriores, tem-se  $\hat{\beta}_0 = \hat{\boldsymbol{\beta}} = (\mathbf{1}_n^t \mathbf{1}_n)^{-1} \mathbf{1}_n^t \mathbf{Y} = \frac{1}{n} \mathbf{1}_n^t \mathbf{Y} = \frac{1}{n} \sum_{i=1}^n Y_i = \bar{Y}$ . Assim, o estimador de mínimos quadrados de  $\beta_0$  é a média amostral das observações de  $Y$ .
4. Por definição, a Soma de Quadrados da Regressão é  $SQR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ . Mas, como se viu, no contexto do modelo Nulo,  $\hat{y}_i = \bar{y}$ ,  $\forall i$ . Logo,  $SQR = 0$ . Trata-se dum resultado coerente: o modelo linear não explica nada da variabilidade de  $Y$ , dada a ausência de preditores explicativos dessa variabilidade. Por outro lado, a Soma de Quadrados Residual em qualquer modelo linear é  $SQRE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ . No nosso caso, tem-se  $SQRE = \sum_{i=1}^n (y_i - \bar{y})^2 = SQT$ , pela definição desta última soma de quadrados. Alternativamente, bastava verificar que, em geral  $SQRE = SQT - SQR$ , sendo no nosso caso a última parcela nula. Este resultado indica que no modelo Nulo, toda a variabilidade de  $Y$  é “residual”, ou seja, não explicada pelo modelo.
5. A estatística do teste  $F$  parcial indicado no enunciado é  $F = \frac{n-(p+1)}{p-k} \frac{SQRE_s - SQRE_c}{SQRE_c}$ , onde  $SQRE_s$  e  $SQRE_c$  indicam as somas de quadrados residuais, respectivamente, do submodelo e do modelo completo. Mas no caso do submodelo ser o modelo Nulo, tem-se  $k=0$  e (como se viu na alínea anterior)  $SQRE_s = SQT$ . Logo, a estatística do teste  $F$  parcial vem  $F = \frac{n-(p+1)}{p} \frac{SQT - SQRE_c}{SQRE_c} = \frac{n-(p+1)}{p} \frac{SQRE_c}{SQRE_c} = \frac{QMR_c}{QMR_c}$ , que é a estatística do teste  $F$  de ajustamento global do modelo completo, como se pedia para mostrar.