INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

11 de Janeiro, 2018 **SEGUNDO TESTE 2017-18** **Uma resolução possível**

I

1. Estuda-se a regressão linear múltipla de `log(PPB)` sobre $p=10$ preditores, e que foi ajustada com $n=91$ observações. A qualidade de ajustamento do modelo é medida através do coeficiente de determinação R^2 e testada através dum teste F de ajustamento global. O Coeficiente de Determinação obtido nesta regressão é $R^2=0.7257$, e corresponde a afirmar que o modelo ajustado explica cerca de $72,57\%$ da variabilidade observada nos valores da variável resposta, `log(PPB)`, um valor razoavelmente bom. Esse valor é muito significativamente diferente de zero (o valor correspondente ao Modelo Nulo), como se pode verificar no teste F de ajustamento global:

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

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

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

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05[10,80]} \approx 1.95$ (entre os valores tabelados 1.91 e 1.99)

   Conclusões: Tem-se no enunciado que $F_{calc}=21.16$. Logo, rejeita-se H_0, i.e., considera-se que a qualidade de ajustamento do modelo completo é significativamente melhor (ao nível $\alpha=0.05$) que a do Modelo Nulo (que corresponde à Hipótese Nula). A rejeição é mesmo muito enfática, como se pode verificar por um valor de prova (*p-value*) indistinguível da precisão de máquina ($p<2.2\times10^{-16}$).

2. O valor de R^2 modificado pode ser calculado, ou com base na sua expressão dada no formulário, $R^2_{mod}=1-\frac{QMRE}{QMT}$, ou com base na expressão alternativa, dada nas aulas, $R^2_{mod}=1-(1-R^2)\frac{n-1}{n-(p+1)}$. Esta última expressão é mais imediata, uma vez que se conhecem os valores $R^2=0.7257$, $n=91$ e $p=10$, pelo que substituindo tem-se $R^2_{mod}=0.6914$. Trata-se dum valor necessariamente mais baixo do que $R^2=0.7257$, e cuja diferença já é visível, embora não seja enorme. Essa diferença resulta de a variabilidade não explicada pelo modelo ($1-R^2=0.2743$) ser 'castigada' pelo R^2 modificado, através duma penalização, dada pelo factor $\frac{n-1}{n-(p+1)}=\frac{90}{80}=1.125$. Assim, a variabilidade não explicada foi aumentada em cerca de $12,5\%$ pela definição de R^2_{mod}.

   Alternativamente, e partindo da expressão constante do formulário o Quadrado Médio Residual é o quadrado do valor disponível no enunciado com a designação `Residual standard error`, ou seja, $QMRE=0.2367^2=0.05602689$. O Quadrado Médio Total referido no formulário é $QMT=\frac{SQT}{n-1}$. É possível obter o valor de SQT a partir da definição (e do valor) do Coeficiente de Determinação e da fórmula fundamental da regressão: $R^2=\frac{SQR}{SQT}=\frac{SQT-SQRE}{SQT}=1-\frac{SQRE}{SQT}$. Isolando SQT, tem-se $SQT=\frac{SQRE}{1-R^2}=\frac{QMRE\times(n-(p+1))}{1-R^2}=\frac{0.05602689\times80}{1-0.7257}=16.34033$ e $QMT=\frac{16.34033}{90}=0.1815592$. Usando este valor obtém-se $R^2_{mod}=1-\frac{0.05602689}{0.1815592}=0.6914126$.

3. Trata-se dum gráfico de resíduos estandardizados (R_i), no eixo vertical, contra valores do efeito alavanca (h_{ii}) no eixo horizontal. São ainda visíveis, nos cantos superior e inferior direito, curvas de igual valor das distâncias de Cook, que medem a influência de cada observação no ajustamento do modelo. Nenhuma observação tem um resíduo estandardizado invulgar, havendo uma única

observação (em 91) com um resíduo $R_i \approx 3$, sendo todos os restantes valores absolutos inferiores a 2 (ou valores muito próximos de 2). Quanto ao efeito alavanca, sabemos ter de estar compreendido entre $\frac{1}{n} = \frac{1}{91} = 0.010989$ e 1, e de ter valor médio igual a $\overline{h} = \frac{p+1}{n} = \frac{11}{91} = 0.1208791$. Verifica-se que algumas observações têm valor alavanca relativamente elevado, com destaque para a observação que surge mais à direita no gráfico, com efeito alavanca um pouco acima de 0.5. Nenhuma observação tem uma distância de Cook acima do limiar de guarda (0.5), já que nenhum ponto se encontra para além das isolinhas de Cook para esse valor, que são visíveis nos cantos à direita no gráfico. Tendo em conta a fórmula que relaciona as distâncias de Cook D_i com os resíduos estandardizados e os efeitos alavanca (disponível no formulário), nomeadamente $D_i = R_i^2 \left(\frac{h_{ii}}{1-h_{ii}}\right) \frac{1}{p+1}$, é possível calcular um valor aproximado para a distância de Cook associada à observação mais à direita no gráfico, para a qual $h_{ii} \approx 0.5$ e $R_i \approx 1$. Assim, tem-se $D_i \approx \frac{1}{p+1} = \frac{1}{11} = 0.0909$. Trata-se dum valor relativamente pequeno, ilustrando que os conceitos de distância de Cook (influência) e valor do efeito alavanca, estando embora relacionados, não são equivalentes.

4. É agora dado um submodelo com apenas $k=6$ preditores.

   (a) Pede-se um teste F parcial para comparar o submodelo com o modelo completo original, de $p=10$ preditores. Tem-se:

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

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

   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 $F_{calc} > f_{0.05[4,80]} \approx 2.5$.

   Conclusões: Tem-se $F_{calc} = \frac{80}{4} \frac{0.7257-0.7007}{1-0.7257} = 1.822822$. Logo, não se rejeita H_0, i.e., considera-se que a qualidade de ajustamento do modelo completo não difere significativamente (ao nível $\alpha = 0.05$) da do submodelo. Nesse caso, será justificável trabalhar com o submodelo mais parcimonioso, com pouco mais de metade dos preditores.

   (b) O critério de Informação de Akaike (AIC), numa regressão linear múltipla com k preditores, é dado no formulário: $AIC = n \ln\left(\frac{SQRE_k}{n}\right) + 2(k+1)$. Temos $n=91$; $k=6$; e um valor de $SQRE$ que pode ser calculado com base no valor (dado no enunciado) de $\sqrt{QMRE}=0.2413$ e na relação $QMRE = \frac{SQRE}{n-(k+1)} \Leftrightarrow SQRE = [n-(k+1)] \times (\sqrt{QMRE})^2 = 84 \times 0.2413^2 = 4.890958$. Logo, tem-se $AIC = 91 \times \ln\left(\frac{4.890958}{91}\right) + 14 = -252.0359$. Este valor é directamente comparável com o valor obtido no modelo completo (dado no enunciado), já que o AIC pode ser usado para comparar modelos de regressão linear diferentes, desde que tenham a mesma variável resposta e sejam ajustados com o mesmo conjunto de dados, o que é o caso. Nessa comparação, o modelo com o menor AIC é considerado o melhor modelo. Tendo em conta que o valor $AIC = -251.98$ obtido no modelo completo é maior do que o valor agora obtido, o Critério de Akaike sugere que o submodelo, apesar do seu R^2 inferior, é preferível.

   (c) É pedido um teste à igualdade dos β_j correspondentes aos preditores `B6s20` e `B12s20`. Os valores estimados relativamente próximos destes dois parâmetros (27.2611 e 24.9354) reforçam o interesse no teste. A hipótese da igualdade destes β_js corresponderia a afirmar que a variação esperada na variável resposta (`log(PPB)`), associada a aumentar em uma unidade as reflectâncias numa dessas duas bandas é igual. Vamos construir um intervalo de confiança, adaptando a notação normalmente usada para definir os índices dos βs e que, para facilitar, serão nesta alínea referidos por β_6 e β_{12}, fazendo o j corresponder ao número

de cada banda. Sabemos que os ICs a $(1-\alpha)\times 100\%$ para uma diferença de dois parâmetros, neste caso $\beta_6 - \beta_{12}$, são centrados na estimativa amostral, $b_6 - b_{12}$. A sua semi-amplitude é dada (como noutros casos) pelo produto do erro padrão estimado (neste caso, $\hat{\sigma}_{\hat{\beta}_6-\hat{\beta}_{12}}$), vezes o valor da distribuição t-Student ($t_{\frac{\alpha}{2}(n-(p+1))}$). Assim, o intervalo é da forma:

$$\left] (b_6 - b_{12}) - t_{\frac{\alpha}{2}(n-(p+1))} \cdot \hat{\sigma}_{\hat{\beta}_6-\hat{\beta}_{12}} \ , \ (b_6 - b_{12}) + t_{\frac{\alpha}{2}(n-(p+1))} \cdot \hat{\sigma}_{\hat{\beta}_6-\hat{\beta}_{12}} \right[$$

Ora, tem-se $b_6 - b_{12} = 27.2611 - 24.9354 = 2.3257$. Por outro lado, e tendo em conta que temos $p=6$ preditores neste modelo, o valor da distribuição t para um intervalo a 95% de confiança é $t_{0.025(84)} \approx 1.99$. Finalmente, calcula-se o erro padrão como a raíz quadrada da estimativa da variância da diferença de duas variáveis aleatórias (usando a propriedade $V[X-Y] = V[X] + V[Y] - 2\,Cov[X,Y]$). Assim, adaptando ao nosso contexto (e tendo em conta que apenas conhecemos as estimativas das variâncias e covariância necessárias), tem-se:

$$\hat{\sigma}_{\hat{\beta}_6-\hat{\beta}_{12}} \ = \ \sqrt{\hat{V}[\hat{\beta}_6 - \hat{\beta_{12}}]} \ = \ \sqrt{\hat{V}[\hat{\beta}_6] + \hat{V}[\hat{\beta}_{12}] - 2\,Cov[\hat{\beta}_6, \hat{\beta}_{12}]}$$

Usando os valores disponíveis na matriz de (co-)variâncias estimadas dos $\hat{\beta}_j$ (disponível no enunciado e sem a qual não seria possível obter a última parcela sob a raíz quadrada), tem-se: $\hat{\sigma}_{\hat{\beta}_6-\hat{\beta}_{12}} = \sqrt{113.4074 + 87.5207 - 2 \times 80.1549} = 6.373249$. Assim, o intervalo a 95% de confiança para $\beta_6 - \beta_{12}$ é o intervalo $]-10.35707 \ , \ 15.00847\ [$. Este intervalo (algo comprido) contém o valor zero. Assim, é admissível que $\beta_6 - \beta_{12} = 0$, ou seja, é admissível a igualdade destes coeficientes, como era admitido no enunciado.

(d) Sugere-se a exclusão apenas do preditor `B8s10`, o único correspondente à resolução espacial dos 10 metros. Para calcular o valor do Coeficiente de Determinação do submodelo com 5 preditores, resultante da sua exclusão, podemos utilizar a estatística do teste F parcial comparando o submodelo inicial deste ponto, com $p=6$ preditores e o submodelo de apenas $k=5$ preditores resultante da exclusão de `B8s10`, ou seja, a estatística $F = \frac{n-(p+1)}{p-k} \, \frac{R_c^2 - R_s^2}{1-R_c^2}$. Uma vez que os dois modelos em comparação diferem numa única variável (`B8s10`), sabemos que o valor desta estatística F parcial será o quadrado do valor da estatistica t usada para testar (no modelo inicial) a hipótese de que o coeficiente β_j da variável excluída seja nulo, e que é dada no enunciado: $t_{calc} = 2.009$. Assim, obtemos a seguinte igualdade, em que a única incógnita é o valor do R^2 procurado:

$$F = \frac{n-(p+1)}{p-k} \cdot \frac{R_c^2 - R_s^2}{1-R_c^2} \quad \Leftrightarrow \quad (2.009)^2 = \frac{91-7}{1} \cdot \frac{0.7007 - R_s^2}{1-0.7007}$$

Isolando R_s^2, obtém-se:

$$\frac{0.2009^2 \times 0.2993}{84} = 0.7007 - R_s^2 \quad \Leftrightarrow \quad R_s^2 = 0.7007 - 0.01438094 = 0.6863191 \ .$$

Assim, a exclusão do preditor `B8s10` diminui a proporção da variabilidade observada nos log-PPB para cerca de $68,63\%$.

II

1. Trata-se dum delineamento factorial a dois factores, sendo a variável resposta Y o teor de amido na matéria fresca de abóboras; o primeiro factor (A) a data de colheita, com $a=4$ níveis e o segundo factor (B) o tratamento usado (também com $b=4$ níveis). O delineamento é equilibrado, uma vez que em cada uma das $ab=16$ células (situações experimentais) existem $n_c=3$ repetições (parcelas). Havendo repetições nas células, é possível (e desejável) estudar a existência de eventuais efeitos de interacção, e foi esse o modelo ANOVA ajustado:

   - $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, para qualquer $i=1,2,3,4$, $j=1,2,3,4$ e $k=1,2,3$, sendo μ_{11} o teor esperado de amido na primeira data de colheita (que, por ordem alfabética será o nível `Nov`, ou seja, Novembro), e com o primeiro tratamento (`A`); α_i o efeito principal (acréscimo ao teor médio populacional de amido nessa primeira célula) associado à data de colheita i (com a restrição $\alpha_1 = 0$); β_j o efeito principal (acréscimo ao teor médio de amido da primeira célula) associado ao tratamento j (com a restrição $\beta_1 = 0$); $(\alpha\beta)_{ij}$ o efeito de interacção associado ao cruzamento da data i de colheita com o tratamento j (com as restrições $(\alpha\beta)_{ij} = 0$ se i e/ou j forem iguais a 1). Finalmente ϵ_{ijk} é o erro aleatório associado à observação Y_{ijk}.
   - Admite-se que os erros aleatórios são 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. Sabemos que os graus de liberdade associados a $QMRE$ são dados por $n-ab$, onde n é o número total de observações, $n=n_cab=3\times16=48$, e $ab=16$ é o número total de parâmetros existentes no modelo. Assim, $g.l.(SQRE)=32$. 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 $a-1=3$ g.l., associado a SQB há igualmente $b-1=3$ g.l., e associado a $SQAB$ há $(a-1)(b-1)=9$ graus de liberdade. Os Quadrados Médios são dados pelas Somas de Quadrados a dividir pelos respectivos graus de liberdade, pelo que $QMB=\frac{1.208}{3}=0.403$. O Quadrado Médio associado ao factor A (para o qual não se dispõe ainda da Soma de Quadrados) pode ser calculado a partir da definição da respectiva estatística $F_A = \frac{QMA}{QMRE}$, uma vez que se sabe que $F_A=33.015$ e $QMRE=0.254$. Logo, $QMA=33.015\times0.254=8.38581$. Assim, a Soma de Quadrados associada ao mesmo Factor A será dada por $SQA=QMA\times(a-1)=8.38581\times3=25.15743$. Finalmente, o valor da estatística F_{AB} associada ao teste aos efeitos de interacção é $F_{AB}=\frac{QMAB}{QMRE}=\frac{0.472}{0.254}=1.858268$.

   Nota: Alguns destes valores sofrem erros de arredondamento nos cálculos.

   Logo, a tabela-resumo completa obtida é:

```
                Df   Sum Sq   Mean Sq  F value
data             3   25.157     8.386   33.015
tratamento       3    1.208     0.403    1.586
data:tratamento  9    4.250     0.472    1.858
Residuals       32    8.122     0.254
```

3. A afirmação do utilizador é que apenas serão significativos os efeitos principais do factor A (`data`), não se rejeitando as hipóteses nulas dos testes aos efeitos principais do factor B (`tratamento`) e de interacção. Vai-se efectuar em pormenor o teste aos efeitos de interacção, e descrever sinteticamente os testes aos efeitos principais do cada factor.

Hipóteses: $H_0 : (\alpha\beta)_{ij} = 0 \ , \ \forall \ i,j$ vs. $H_1 : \exists\, i,j$ tal que $(\alpha\beta)_{ij} \neq 0$.

Estatística do Teste: $F_{AB} = \frac{QMAB}{QMRE} \cap F_{[(a-1)(b-1),n-ab]}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(9,32)} \approx 2.20$ (entre os valores tabelados 2.12 e 2.21).

Conclusões: Como $F_{calc} = 1.858 < 2.20$, não se rejeita H_0, concluindo-se que não existem efeitos significativos de interacção (ao nível $\alpha = 0.05$). A conclusão apenas seria diferente (e mesmo assim, por pouco) caso se optasse por um nível de significância relativamente elevado ($\alpha = 0.10$), tendo em conta o *p-value* disponível no enunciado ($p = 0.0951$).

No teste aos efeitos principais do factor A (`data`), com hipóteses $H_0 : \alpha_i = 0$, para todo o i, contra H_1 : existe pelo menos uma data $i > 1$ onde $\alpha_i \neq 0$, o valor calculado da estatística de teste é muito elevado ($F_{calc} = 33.105$) dispondo-se no enunciado do valor de prova $p = 6.48 \times 10^{-6}$. Este valor muito inferior a qualquer dos níveis habituais de significância α conduz à rejeição da H_0, ou seja, conclui-se pela existência de efeitos principais associados às diferentes datas de colheita.

Finalmente, no teste aos efeitos principais do factor B (`tratamento`), as hipóteses do teste são $H_0 : \beta_j = 0$, para todo o j, contra H_1 : existe pelo menos um tratamento $j > 1$ tal que $\beta_j \neq 0$. O valor calculado da estatística de teste é baixo ($F_{calc} = 1.586$) e o valor de prova no enunciado $p = 0.2120$ é superior a qualquer dos habituais níveis de significância. Assim, conclui-se pela inexistência de efeitos principais de tratamento, nos teores médios de amido na abóbora.

A afirmação do utilizador é assim, correcta, para níveis de significância como $\alpha = 0.05$ ou $\alpha = 0.01$.

4. As conclusões da alínea anterior sugerem que apenas poderá haver diferenças associadas a diferentes datas, pelo que a especificação de tratamentos feita no enunciado seria despropositada. No entanto, responder-se-á à pergunta através de comparações entre pares de médias *de célula*, não apenas por ser o que se pede no enunciado, mas também porque as comparações múltiplas de Tukey podem dar resultados nem sempre coerentes com os dos testes F.

   Pretende-se comparar (com $\alpha = 0.05$) os pares de médias que se podem formar a partir das dezasseis médias de célula. Sabemos que duas médias de célula populacionais, μ_{ij} e $\mu_{i'j'}$, devem ser consideradas diferentes se as respectivas médias amostrais diferirem, em módulo, em mais do que a diferença significativa de Tukey, $q_{\alpha(ab,n-ab)} \sqrt{\frac{QMRE}{n_c}}$ (disponível no formulário), ou seja, se $|\overline{y}_{ij.} - \overline{y}_{i'j'.}| > q_{0.05(16,32)} \sqrt{\frac{0.254}{3}}$. O valor da distribuição de Tukey pode ser obtido nas tabelas desta distribuição e é 5.24. Logo, o termo de comparação é $5.24 \times 0.2909754 = 1.524711$. Olhando para as médias das dezasseis células disponíveis no enunciado, imediatamente se verifica que a média amostral da célula $(3,4)$ (correspondente à primeira data de Setembro, tratamento `D`), que é 2.554, apenas pode ser considerada significativamente diferente de qualquer outra média de célula $\overline{y}_{ij}$ para a qual se verifique $|2.554 - \overline{y}_{ij}| > 1.524711$, ou seja (e tendo em conta que a média da célula $(3,4)$ é a maior das médias amostrais de célula), para as quais $\overline{y}_{ij} < 2.554 - 1.524711 = 1.029289$. Ora, esta condição não é verificada por qualquer das médias amostrais de célula correspondentes às duas datas de Setembro, nem pelas médias amostrais das células correspondentes à data de Outubro com os dois primeiros tratamentos. Assim, apenas as datas de Novembro, bem como as de Outubro com os tratamentos `C` e `D` podem ser consideradas significativamente diferentes da maior média amostral de célula. A afirmação do enunciado é falsa.

5. O parâmetro β_2 deste modelo, como indicado na primeira alínea, onde se especificou o modelo, é o efeito principal associado ao segundo nível do factor B, ou seja o efeito principal associado a

usar-se o tratamento B. O seu valor estimado é dado no formulário: $\hat{\beta}_2 = \overline{y}_{12.} - \overline{y}_{11.}$. Assim, temos no nosso caso, a estimativa $\hat{\beta}_2 = 0.3839 - 0.2751 = 0.1088$. Recordando que, num modelo ANOVA a dois factores com interacção, e com as restrições acima indicadas, os efeitos são estimados pelas quantidades amostrais correspondentes à definição do parâmetro, facilmente se deduz que o parâmetro β_2 é dado por $\mu_{12} - \mu_{11}$. Assim, o valor estimado 0.1088 corresponde à diferença estimada entre o teor médio de amido em Novembro, com o tratamento B, e o teor médio de amido, na mesma data, com o tratamento A. No entanto, quer o teste F aos efeitos principais do factor B, quer o teste de Tukey realizado na alínea anterior, dizem-nos que este valor não difere significativamente de zero, pelo que se deve admitir que $\mu_{12} = \mu_{11}$.

6. O resíduo duma observação é sempre a diferença entre o valor observado e o valor correspondente, ajustado pelo modelo. Num modelo ANOVA para um delineamento factorial a dois factores, com efeitos de interacção, o valor ajustado de qualquer observação é a média amostral da mesma célula onde foi feita a observação, ou seja (ver o formulário) $e_{ijk} = y_{ijk} - \overline{y}_{ij.}$. Tratando-se da terceira observação ($k = 3$) efectuada em Novembro ($i = 1$), com o tratamento D ($j = 4$), tem-se $e_{143} = y_{143} - \overline{y}_{14.} = 0.11194 - 0.3419 = -0.22996$.

III

1. No contexto da regressão linear múltipla indicada no enunciado, tem-se:

   (a) a matriz do modelo, $\mathbf{X}$ tem n linhas (tantas quantas as observações com base nas quais se ajusta o modelo) e $p+1$ colunas (tantas quantos os parâmetros do modelo). A primeira coluna é uma coluna de n uns, $\vec{\mathbf{1}}_n$ (que fica associada à constante β_0 na equação do modelo) e cada uma das p restantes colunas é composta pelas n observações de uma das variáveis preditoras (ou seja, é o vector $\vec{\mathbf{x}}_j$ das observações do j-ésimo preditor, associado à constante β_j na equação do modelo). Assim, a matriz do modelo $\mathbf{X}$ tem este aspecto:

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

   Por definição, o espaço $\mathcal{C}(\mathbf{X})$ das colunas da matriz $\mathbf{X}$ é o subespaço de $\mathbb{R}^n$ (o espaço onde residem as colunas de $\mathbf{X}$) gerado por todas as possíveis combinações lineares das colunas de $\mathbf{X}$, ou seja, o espaço dos vectores da forma $a_0\vec{\mathbf{1}}_n + a_1\vec{\mathbf{x}}_1 + a_2\vec{\mathbf{x}}_2 + ... + a_p\vec{\mathbf{x}}_p$, para qualquer conjunto de coeficientes $a_0, a_1, ..., a_p$.

   (b) Por definição (ver formulário), a matriz de projecção ortogonal no subespaço $\mathcal{C}(\mathbf{X})$ é dada por $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$. Uma matriz quadrada $\mathbf{H}$ diz-se simétrica se $\mathbf{H}^t = \mathbf{H}$. Ora, tendo em conta as propriedades de produtos matriciais, também constantes do formulário, nomeadamente $(\mathbf{AB})^t = \mathbf{B}^t\mathbf{A}^t$; $(\mathbf{A}^t)^t = \mathbf{A}$; e $(\mathbf{A}^{-1})^t = (\mathbf{A}^t)^{-1}$, tem-se:

   $$\begin{aligned} \mathbf{H}^t &= [\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t]^t = (\mathbf{X}^t)^t[(\mathbf{X}^t\mathbf{X})^{-1}]^t\mathbf{X}^t = \mathbf{X}[(\mathbf{X}^t\mathbf{X})^t]^{-1}\mathbf{X}^t \\ &= \mathbf{X}[\mathbf{X}^t(\mathbf{X}^t)^t]^{-1}\mathbf{X}^t = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{H}\ , \end{aligned}$$

   como se queria mostrar. Por outro lado, $\mathbf{H}$ diz-se idempotente se $\mathbf{HH} = \mathbf{H}$. Ora,

   $$\mathbf{HH} = \mathbf{X}\underbrace{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t \cdot \mathbf{X}}_{=\mathbf{I}_{p+1}}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{H}\ .$$

(c) A Soma de Quadrados dos Resíduos é, por definição, $SQRE = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$. Esta soma de quadrados é a norma ao quadrado do vector dos resíduos, isto é, do vector cujo elemento genérico é dado por $Y_i - \hat{Y}_i$. Trata-se do vector $\vec{\mathbf{Y}} - \vec{\hat{\mathbf{Y}}}$. Sabemos que o segundo destes vectores é dado por $\vec{\hat{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$. Assim, $SQRE = \|\vec{\mathbf{Y}} - \mathbf{H}\vec{\mathbf{Y}}\|^2$. Pondo em evidência (à direita) o vector $\vec{\mathbf{Y}}$, tem-se $SQRE = \|(\mathbf{I} - \mathbf{H})\vec{\mathbf{Y}}\|^2$. Ora, pela definição de norma para qualquer vector $\vec{\mathbf{x}}$, tem-se $\|\vec{\mathbf{x}}\|^2 = \vec{\mathbf{x}}^t\vec{\mathbf{x}}$. Logo, $SQRE = [(\mathbf{I} - \mathbf{H})\vec{\mathbf{Y}}]^t[(\mathbf{I} - \mathbf{H})\vec{\mathbf{Y}}]$. Usando as propriedades de produtos e transpostas de matrizes, bem como a simetria e idempotência da matriz de projecção $\mathbf{H}$ (ver a alínea anterior) e da matriz identidade, tem-se $SQRE = \vec{\mathbf{Y}}^t(\mathbf{I} - \mathbf{H})^t(\mathbf{I} - \mathbf{H})\vec{\mathbf{Y}} = \vec{\mathbf{Y}}^t(\mathbf{I}^t - \mathbf{H}^t)(\mathbf{I} - \mathbf{H})\vec{\mathbf{Y}} = \vec{\mathbf{Y}}^t(\mathbf{I}^t\mathbf{I} - \mathbf{I}^t\mathbf{H} - \mathbf{H}^t\mathbf{I} + \mathbf{H}^t\mathbf{H})\vec{\mathbf{Y}} = \vec{\mathbf{Y}}^t(\mathbf{I} - \mathbf{H} - \mathbf{H} + \underbrace{\mathbf{H}\mathbf{H}}_{=\mathbf{H}})\vec{\mathbf{Y}} = \vec{\mathbf{Y}}^t(\mathbf{I} - \mathbf{H})\vec{\mathbf{Y}}$, como se queria mostrar.

2. O enunciado refere um modelo ANOVA para um delineamento hierarquizado a dois factores.

(a) Por definição, os Quadrados Médios são dados pelas Somas de Quadrados a dividir pelos respectivos graus de liberdade. Assim, no modelo a um único factor A (com a níveis), tem-se por definição, $QMRE_A = \frac{SQRE_A}{n-a}$. No delineamento hierarquizado a dois factores, tem-se (ver também o formulário): $QMRE_{A/B} = \frac{SQRE_{A/B}}{n - \sum_{i=1}^{a} b_i}$. Sabemos ainda que a definição de Soma de Quadrados associada ao Factor subordinado B, no delineamento hierarquizado, é $SQB(A) = SQRE_A - SQRE_{A/B}$ e que $QMB(A) = \frac{SQB(A)}{\sum_{i=1}^{a}(b_i - 1)} = \frac{SQB(A)}{(\sum_{i=1}^{a} b_i) - a}$. Logo, o Quadrado Médio Residual no modelo apenas com o Factor A pode ser escrito como:

$$QMRE_A = \frac{SQRE_A}{n-a} = \frac{SQRE_{A/B} + SQB(A)}{n-a} = \frac{QMRE_{A/B}(n - \sum_i b_i) + QMB(A)(\sum_i b_i - a)}{n-a}.$$

Dividindo tudo por $QMRE_{A/B}$, e uma vez que $F_{B(A)} = \frac{QMB(A)}{QMRE_{A/B}}$, tem-se: $\frac{QMRE_A}{QMRE_{A/B}} = \frac{(n - \sum_i b_i) + F_{B(A)}(\sum_i b_i - a)}{n-a}$. Logo,

$$\begin{aligned} QMRE_A < QMRE_{A/B} &\Leftrightarrow (\cancel{n} - \sum_i b_i) + F_{B(A)}(\sum_i b_i - a) < \cancel{n} - a \\ &\Leftrightarrow F_{B(A)}(\sum_i b_i - a) < \sum_i b_i - a \quad \Leftrightarrow \quad F_{B(A)} < 1 \end{aligned}$$

como se queria mostrar. **Aviso:** Na resolução da Primeira Chamada de Exame (realizada na mesma data e onde também constava esta pergunta) encontra-se uma resolução alternativa.

(b) A estatística do teste F aos efeitos do factor A no modelo a um factor é dada por $F = \frac{QMA}{QMRE_A}$. No delineamento a dois factores hierarquizados, a estatística correspondente tem forma análoga, $F^* = \frac{QMA}{QMRE_{A/B}}$, sendo o numerador QMA definido exactamente da mesma forma nos dois modelos. Na alínea anterior viu-se que $F_{B(A)} < 1$ equivale a $QMRE_A < QMRE_{A/B}$. Como a menores denominadores (e iguais numeradores) correspondem maiores fracções, tem-se que a estatística F no modelo apenas com o factor A terá um valor maior do que a estatística do correspondente teste no modelo com o factor B subordinado ao factor A.

(c) Afirmar que $QMRE_A < QMRE_{A/B}$ parece paradoxal, uma vez que os Quadrados Médios Residuais nos modelos ANOVA estimam a variabilidade dos erros aleatórios (σ^2), ou seja, a variabilidade *não* explicada pelo modelo, e parece estranho que haja mais variabilidade inexplicada num modelo que, além de ter o Factor A, ainda tem mais um Factor capaz

de explicar variabilidade. Uma tal situação pode ocorrer, no entanto, quando a Soma de Quadrados explicada pelo factor adicional é muito pequena. De facto, o Quadrado Médio Residual do modelo hierarquizado, $QMRE_{A/B} = \frac{SQRE_{A/B}}{n - \sum_{i=1}^{a} b_i}$ tem um numerador que é necessariamente mais pequeno que o do Quadrado Médio Residual do modelo apenas com o Factor A, $QMRE_A = \frac{SQRE_A}{n-a}$, já que $SQRE_{A/B} = SQRE_A - SQB(A)$ e uma Soma de Quadrados nunca pode ser negativa, pelo que $SQRE_{A/B} \leq SQRE_A$. No entanto, o *denominador* de $QMRE_{A/B}$ também tem de ser menor que o denominador de $QMRE_A$, já que cada nível do factor A tem de ter pelo menos um nível do Factor B subordinado, ou seja, $b_i \geq 1$, o que implica que $\sum_{i=1}^{a} b_i \geq \sum_{i=1}^{a} 1 = a$, logo $n - \sum_{i=1}^{a} b_i \leq n - a$ (e como não faz sentido que em todos os níveis de A haja apenas um nível de B, a desigualdade é seguramente estrita). Assim, se $QMRE_{A/B}$ é, ou não, menor que $QMRE_A$ depende da relação entre o que o novo factor B consegue reduzir na Soma de Quadrados Residual, e o que obriga a reduzir nos graus de liberdade. As alíneas anteriores mostram que é possível que $QMRE_A < QMRE_{A/B}$, e que essa situação corresponde a ter $F_{B(A)} < 1$. Uma rápida consulta às tabelas da distribuição F (para qualquer dos níveis de significância α) mostra que se $F_{B(A)} < 1$ nunca se rejeita a hipótese de que os efeitos do Factor subordinado B sejam nulos. Assim, pode afirmar-se que $QMRE_A < QMRE_{A/B}$ corresponde a uma situação onde o segundo factor previsto no delineamento hierarquizado está longe de ter efeitos significativos e a perda de graus de liberdade é mais grave do que os ganhos na redução das Somas de Quadrados Residual. A lição geral desta discussão (que pode adaptar-se a outros delineamentos a dois factores) é que a introdução de novos factores no delineamento apenas é vantajosa se a esses novos factores correspondem na realidade efeitos significativos, ou seja, se a variabilidade que eles contribuem para explicar for relativamente grande.