INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

11 de Janeiro, 2018 **PRIMEIRO EXAME 2017-18** **Uma resolução possível**

I

É pedido para efectuar um teste χ^2 para averiguar se a variável aleatória X, que conta o número de moscas por armadilha, tem distribuição de Poisson.

1. Não sendo especificado nenhum valor do único parâmetro (λ) da distribuição de Poisson, será necessário estimá-lo. A sua estimação parte do facto de que numa distribuição de Poisson, o parâmetro é o valor esperado da v.a.: $E[X]=\lambda$. Uma vez que o valor esperado é estimado pela média amostral, a melhor estimativa de λ corresponde a tomar $\hat{\lambda}=\overline{x}=\frac{0\times8+1\times6+2\times2+3\times6+4\times1+5\times0+6\times0+7\times1}{24}=\frac{39}{24}=1.625$.

2. A distribuição χ^2 da estatística de Pearson é apenas assintótica, ou seja, aproximada para grandes amostras. O Critério de Cochran é uma regra que permite considerar que a amostra é suficientemente grande para se aceitar a validade dessa distribuição assintótica. Ao abrigo do critério de Cochran, nenhum valor esperado (estimado) deve ser inferior a 1, e não mais do que 20% devem ser inferiores a 5. Ora, os valores esperados estimados são dados por $\hat{E}_i=N\times\hat{\pi}_i$, onde $N=24$ e $\hat{\pi}_i=P[X=i]$ é a probabilidade (estimada) de o número de moscas na armadilha ser i ($i=0,1,2,...$), caso seja verdade que $X\cap Pois(\hat{\lambda}=1.625)$. Sabemos que $\hat{\pi}_i=P[X=i]=\mathrm{e}^{-\hat{\lambda}}\frac{\hat{\lambda}^i}{i!}$. Embora o critério de Cochran diga respeito aos valores *esperados* e não aos valores *observados*, a tabela dos valores observados O_i sugere que essas probabilidades sejam menores para os valores grandes de X, pelo que vamos calcular o valor esperado para um desses valores. Escolhendo $i=6$, tem-se que $\hat{E}_6=N\times\hat{\pi}_6=24\times\mathrm{e}^{-1.625}\frac{1.625^6}{6!}=0.1208566<1$. Havendo uma contagem esperada inferior a 1, tem-se já uma violação do critério de Cochran. Será necessário agregar classes de contagem, a fim de se poder aplicar o teste de χ^2 a estes dados.

   Nota: A escolha da última classe obrigaria a trabalhar, não com a probabilidade $P[X=7]$, mas sim com a probabilidade acumulada $P[X\geq 7]$, já que a soma das probabilidades de todas as classes deve ser 1 e, numa v.a. com distribuição de Poisson, são possíveis todos os valores em $\mathbb{N}_0$, pelo que a classe terminal terá de ser sempre uma classe da form $X\geq c$.

3. Tendo sido efectuada a agregação de classes referida no enunciado, a nova tabela de contagens à qual se aplicará o teste de χ^2 é:

No. moscas	0	1	2	≥ 3
No. armadilhas	8	6	2	8

   Eis o teste à validade da distribuição de Poisson:

   Hipóteses: $X\cap Pois(\hat{\lambda}=1.625)$ *vs.* $X\not\cap Pois(\hat{\lambda}=1.625)$.

   Estatística do Teste: É a estatística de Pearson, na forma de contagens unidimensionais: $X^2=\sum_{i=1}^{k}\frac{(O_i-\hat{E}_i)^2}{\hat{E}_i}$, sendo $k=4$. A distribuição assintótica desta estatística, caso seja verdade H_0, é χ^2_{k-1-r}, onde $r=1$ indica o parâmetro que foi necessário estimar.

   Nível de significância: $\alpha=$ P[Erro do tipo I] = P[Rej. H_0 | H_0 verdade] $=0.01$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $\chi^2_{calc} > \chi^2_{\alpha(k-1-r)} = \chi^2_{0.01(2)} = 9.210$.

Conclusões: Como $\chi^2_{calc} = 6.8229$, não se rejeita a hipótese nula, ou seja, admite-se (para o nível $\alpha = 0.01$) que a distribuição de X seja Poisson, com o valor do parâmetro 1.625.

II

1. Estuda-se a regressão linear simples de `log(PPB)` (Y) sobre o índice `NDWI` (x), cuja equação do modelo é $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, e que foi ajustada com $n = 91$ pares de observações.

   (a) Pede-se um teste à hipótese $\beta_1 = 4$, já que sabemos que o declive da recta corresponde à variação esperada em Y associada a aumentar o preditor x em uma unidade. Assim,

   Hipóteses: $H_0 : \beta_1 = 4 \quad vs. \quad H_1 : \beta_1 \neq 4$.

   Estatística do Teste: $T = \frac{\hat{\beta}_1 - \beta_{1|H_0}}{\hat{\sigma}_{\hat{\beta}_1}} \cap t_{n-2}$, 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: (Bilateral) Rejeitar H_0 se $|T_{calc}| > t_{0.025(89)} \approx 1.99$.

   Conclusões: Tem-se $T_{calc} = \frac{b_1 - 4}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{3.83488 - 4}{0.30432} = -0.5425868$. Logo, não se rejeita H_0, sendo de admitir que um aumento de uma unidade no índice `NDWI` provoca, em média um aumento de 4 unidades na log-Produtividade Primária Básica (para o nível $\alpha = 0.05$).

   (b) A expressão do intervalo a $(1-\alpha) \times 100\%$ de confiança para β_0 é:

   $$\left] \quad b_0 - t_{\frac{\alpha}{2}\,(n-2)}\, \hat{\sigma}_{\hat{\beta}_0} \quad , \quad b_0 + t_{\frac{\alpha}{2}\,(n-2)}\, \hat{\sigma}_{\hat{\beta}_0} \quad \right[\,,$$

   sendo $\hat{\sigma}_{\hat{\beta}_0} = \sqrt{\text{QMRE}\left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right]}$ (a expressão debaixo da raíz quadrada estima a verdadeira variância do estimador $\hat{\beta}_0$, cuja expressão é dada no formulário; a estimação resulta de substituir o desconhecido valor de σ^2 por $QMRE$). O enunciado dá-nos os valores de $b_0 = 2.77400$ e $\hat{\sigma}_{\hat{\beta}_0} = 0.02872$. O valor da distribuição t pode ser obtido (aproximadamente) nas tabelas, e é $t_{0.025(89)} \approx 1.99$. Logo, o intervalo de confiança pedido é:] 2.7168 , 2.8312 [. Trata-se dum intervalo a 95% de confiança para a log-Produtividade Primária Bruta correspondente ao valor 0 do índice `NDWI` (o valor intermédio da escala em que este índice é medido). Um intervalo correspondente para o valor de PPB nas suas unidades de medida (como pedido no enunciado) correspondente a `NDWI`$=0$, é obtido exponenciando os extremos do IC acima:] 15.12 , 16.97 [.

   (c) A expressão dum intervalo de predição para uma observação individual de Y, dado $X = x$, é dada no formulário. A partir dessa expressão facilmente se constata que:

   i. o ponto central do intervalo é $\hat{\mu}_{Y|X=x} = b_0 + b_1\, x$, que corresponde à estimativa do valor esperado de log-PPB para $X = x$. Quando $x = 0.1$, tem-se $\hat{\mu}_{Y|X=0.1} = 2.77400 + 3.83488\,(0.1) = 3.15749$.

   ii. O extremo direito do intervalo obtém-se somando ao ponto central agora calculado, a distância entre esse mesmo ponto central e o extremo esquerdo do intervalo, ou seja, o intervalo termina no valor $3.15749 + (3.15749 - 2.64287) = 3.67211$. Assim, o intervalo de predição é o intervalo] 2.64287 , 3.67211 [.

iii. O erro padrão associado ao estimador $\hat{\mu}_{Y|X=0.1}$ é dado por $\hat{\sigma}_{\hat{\mu}_{Y|X=x}} = \sqrt{QMRE\left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}\right]}$ (que corresponde à expressão do erro padrão associada a uma observação individual, constante do formulário, mas sem a parcela "1+"). Conhecem-se todas as quantidades envolvidas: $\sqrt{QMRE} = 0.2568$; $n = 91$; $x = 0.1$; $\overline{x} = 0.03286$; e $s_x^2 = 0.007910756$. Substituindo, obtém-se: $\hat{\sigma}_{\hat{\mu}_{Y|X=0.1}} = 0.03379672$.

(d) Uma relação linear da forma $ln(PPB) = \beta_0 + \beta_1 NDWI$ equivale a uma relação exponencial entre `PPB` (y) e `NDWI` (x). De facto, exponenciando a equação anterior, obtém-se $y = e^{\beta_0+\beta_1 x} = e^{\beta_0} e^{\beta_1 x}$. No caso em consideração, e tendo em conta os valores estimados b_0 e b_1, tem-se a equação ajustada da seguinte exponencial: $y = 16.0226\, e^{3.83488\, x}$. Sabemos que uma tal relação exponencial entre duas variáveis, neste caso, `PPB` e `NDWI`, corresponde a admitir que a taxa de variação relativa de `PPB` (considerada função de `NDWI`) é constante, sendo essa constante dada pelo parâmetro β_1, ou seja, corresponde a admitir $\frac{y'(x)}{y(x)} = \beta_1$. Assim, o valor estimado dessa taxa de variação relativa de `PPB` em relação a `NDWI` é, no nosso caso, $b_1 = 3.83488$.

(e) Um modelo potência entre duas variáveis y e x corresponde a uma regressão linear simples entre $log(y)$ e $log(x)$. No entanto, neste exemplo a variável preditora x (`NDWI`) toma valores negativos, pelo que a sua logaritmização não é possível. Assim, a relação potência sugerida no enunciado não é uma opção neste caso.

2. Estudou-se uma regressão linear múltipla da mesma variável resposta, `log(PPB)`, sobre um conjunto de $p = 10$ preditores que *não* inclui o preditor `NDWI` usando na regressão linear simples do ponto anterior.

(a) 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 signficativamente 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.* $H_1 : \mathcal{R}^2 > 0$, onde $\mathcal{R}^2$ indica 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: $\alpha = \text{P[Erro do tipo I]} = \text{P[Rej. } H_0 \mid H_0 \text{ verdade]} = 0.05$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05[10,80]} \approx 1.95$ (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 ($p < 2.2 \times 10^{-16}$).

(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 $n = 91$; $k = p = 10$; e um valor de $SQRE$ que pode ser calculado com base no valor (dado no enunciado) de $\sqrt{QMRE} = 0.2367$ e na relação $QMRE = \frac{SQRE}{n-(p+1)} \Leftrightarrow SQRE = [n-(p+1)] \times (\sqrt{QMRE})^2 = 80 \times 0.2367^2 = 4.482151$. Logo, tem-se $AIC = 91 \times \ln\left(\frac{4.482151}{91}\right) + 22 = -251.9788$. Este valor é directamente comparável com o valor obtido na regressão linear simples (embora o

preditor da RLS não faça parte do conjunto de preditores nesta regressão múltipla), uma vez 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 valor menor de AIC é considerado o melhor modelo. Tendo em conta que o valor $AIC = -245.46$ obtido na regressão linear simples é maior do que o valor agora obtido, o Critério de Akaike sugere que o modelo da regressão linear múltipla, apesar da sua maior complexidade, é preferível.

(c) Um teste F parcial apenas pode ser usado para comparar um modelo e um respectivo submodelo (**Nota:** ou, mais em geral, modelos cujos conjuntos de preditores definam subespaços encaixados - veja a *Breve nota sobre o teste F parcial*, disponível na secção de material de apoio às aulas teóricas, na página *web* da disciplina). Não é o caso neste exemplo, uma vez que o preditor da RLS (o índice `NDWI`) não faz parte do conjunto de preditores usados na regressão linear múltipla. Assim, não é legítimo usar o teste F parcial nesta comparação. A comparação pode ser feita através do AIC (como na alínea anterior), ou mesmo usando os valores de R^2, já que nesse caso se estaria a comparar valores da proporção de variabilidade da *mesma variável resposta* explicada por cada modelo.

(d) 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 $R_i \approx 3$, tendo todos os restantes valores absolutos inferiores a 2 (ou valores muito pouco superiores a 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.

(e) É dado um submodelo com apenas $k = 6$ preditores. 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: $\alpha = \text{P}[\text{ Erro do tipo I }] = \text{P}[\text{ Rej. } H_0 \mid H_0 \text{ verdade }] = 0.05$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05[4,80]} \approx 2.5$.

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.

III

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 α=0.05). A conclusão apenas seria diferente (e mesmo assim, por pouco) caso se optasse por um nível de significância relativamente elevado (α=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 α=0.05 ou α=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 $\mu_{12} = \mu_{11}$.

IV

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 $\vec{\mathbf{Y}} - \hat{\vec{\mathbf{Y}}}$. Sabemos que o

segundo destes vectores é dado por $\vec{\mathbf{Y}} = \mathbf{H}\vec{\mathbf{Y}}$. 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,

$$
\begin{aligned}
QMRE_{A/B} \;>\; QMRE_A \quad &\Leftrightarrow \quad \frac{SQRE_{A/B}}{n - \sum_{i=1}^{a} b_i} \;>\; \frac{SQRE_A}{n-a} \;=\; \frac{SQRE_{A/B} + SQB(A)}{n-a} \\
&\Leftrightarrow \quad SQRE_{A/B}\left(\frac{1}{n - \sum_{i=1}^{a} b_i} - \frac{1}{n-a}\right) \;>\; \frac{SQB(A)}{n-a} \\
&\Leftrightarrow \quad SQRE_{A/B}\left(\frac{(\cancel{n} - a) - (\cancel{n} - \sum_{i=1}^{a} b_i)}{(n-a)(n - \sum_{i=1}^{a} b_i)}\right) \;>\; \frac{SQB(A)}{n-a} \\
&\Leftrightarrow \quad SQRE_{A/B}\left(\frac{\sum_{i=1}^{a} b_i - a}{\cancel{(n-a)}(n - \sum_{i=1}^{a} b_i)}\right) \;>\; \frac{SQB(A)}{\cancel{n-a}} \\
&\Leftrightarrow \quad QMRE_{A/B} \;>\; \frac{SQB(A)}{\sum_{i=1}^{a} b_i - a} \;=\; QMB(A) \\
&\Leftrightarrow \quad 1 \;>\; \frac{QMB(A)}{QMRE_{A/B}} \;=\; F_{B(A)}\ .
\end{aligned}
$$

   como se queria mostrar. **Aviso:** Na resolução do Segundo Teste (realizado 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.