INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

28 de Janeiro, 2019 **EXAME - 2a. Chamada 2018-19** **Uma resolução possível**

I

É dada uma tabela de contingências com $a=2$ linhas e $b=5$ colunas, e contagens O_{ij} correspondentes a cada célula resultante da combinação duma linha i (presença ou ausência do vírus) com uma coluna j (região de origem do genótipo). Uma vez que apenas foi fixado o tamanho global da amostra, $N=664$, responde-se à pergunta através de um teste de *independência*. Designe-se por π_{ij} a probabilidade de uma observação recair na célula (i,j), ou seja, de um genótipo ter o resultado i e ser proveniente da região j. Designe-se por $\pi_{i.}$ a probabilidade marginal de uma observação ter o resultado i, e por $\pi_{.j}$ a probabilidade marginal de uma observação ser proveniente da região j. A independência corresponde a ter-se, sempre, $\pi_{ij} = \pi_{i.} \times \pi_{.j}$. As probabilidades marginais são desconhecidas, mas podem ser estimadas, usando as frequências absolutas de linha, $N_{i.}$, e de coluna, $N_{.j}$, respectivamente. Sabemos que, num teste de indepêndencia, os valores esperados estimados são dados por $\hat{E}_{ij}=\frac{N_{i.}\times N_{.j}}{N}$.

1. Eis o teste pedido:

   Hipóteses: $H_0 : \pi_{ij} = \pi_{i.} \times \pi_{.j}$, $\forall\, i,j \quad vs. \quad H_1 : \exists\, i,j,$ tal que $\pi_{ij} \neq \pi_{i.} \times \pi_{.j}$.

   Estatística do Teste: A estatística de Pearson, é dada por $X^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}}$. A distribuição assintótica desta estatística, caso seja verdade H_0, é $\chi^2_{(a-1)(b-1)}$, com $(a-1)(b-1)=4$.

   Nível de Significância Pode escolher-se $\alpha=0.05$.

   Região Crítica: (Unilateral direita) A regra de rejeição consiste em rejeitar H_0 se $\chi^2_{\text{calc}} > \chi^2_{0.05(4)}=9.48773$.

   Conclusões: No enunciado é dado o valor calculado da estatística de teste: $X^2_{calc} = 36.115$. Logo, rejeita-se H_0, podendo afirmar-se que existe evidência estatisticamente significativa (ao nível de significância $\alpha = 0.05$) de que a incidência do vírus na casta Aragonez não é independente da região.

2. As condições de Cochran visam legitimar a distribuição assintótica como distribuição aproximada da estatística de testes, sob H_0. Exigem que não haja valores esperados estimados inferiores a 1, e não mais de 20% sejam inferiores a 5. Ora, o menor valor esperado estimado correspondente ao problema sob estudo ocorre na célula em que se cruzam as menores frequências de linha e de coluna, que no nosso caso corresponde à presença do vírus ($i=2$) no Dão ($j=2$). Tem-se o valor $\hat{E}_{22}=\frac{N_{2.}\times N_{.2}}{N}=\frac{116\times 32}{664}=5.590361 > 5$. Logo, verificam-se as condições de Cochran e pode admitir-se válida a distribuição assintótica χ^2_4.

3. É pedido para calcular a parcela da estatística do teste correspondente à célula $(i,j)=(2,3)$. Neste caso, o valor observado é $O_{23}=54$. O correspondente valor esperado estimado é $\hat{E}_{23}=\frac{N_{2.}\times N_{.3}}{N}=\frac{116\times 185}{664}=32.31928$. Logo, a parcela correspondente é

$$\frac{(O_{23}-\hat{E}_{23})^2}{\hat{E}_{23}} = \frac{(54-32.31928)^2}{32.31928} = 14.54406\ .$$

Trata-se dum valor muito grande que, só por si, já conduziria à rejeição da Hipótese Nula de independência. Estamos perante uma associação positiva: o número observado de presenças do vírus na região do Douro ($O_{23}=54$) é maior do que seria de esperar ao abrigo da hipótese de independência ($\hat{E}_{23}=32.31928$). Assim, pode afirmar-se que o vírus tem maior incidência no Douro do que seria de esperar no caso de independência. **Nota:** Este facto resulta da preferência do vírus por ambientes húmidos, em detrimento de ambientes secos.

II

1. (a) A recta de regressão que melhor explica a variabilidade da variável `altura` é a que tiver, como preditor, a variável mais fortemente correlacionada com `altura`. Pelo enunciado, verifica-se que esse preditor é a variável `areabasal`, com um coeficiente de correlação linear $r=0.8808362$, resultando num Coeficiente de Determinação de $R^2=(0.8808362)^2=0.7758724$. Assim, essa recta de regressão explicará mais de $77,5\%$ da variância das alturas observadas. A equação da recta ajustada é da forma $y=b_0+b_1\,x$. A fórmula do declive é $b_1=\frac{cov_{xy}}{s_x^2}=r_{xy}\,\frac{s_y}{s_x}$, sendo s_x e s_y os desvios padrão do preditor `areabasal` e da variável resposta `altura`, respectivamente. Estes valores estão disponíveis no enunciado, tendo-se $b_1=0.8808362\times\frac{6.097297}{8.287654}=0.6480386$. Para calcular a ordenada na origem usa-se a fórmula $b_0=\overline{y}-b_1\,\overline{x}$, onde $\overline{y}=17.529341$ e $\overline{x}=14.685629$ são as médias da `altura` e da `areabasal`, respectivamente. Logo, $b_0=8.012486$. A equação da recta é assim $y=8.012486+0.6480386\,x$.

   (b) É pedido um intervalo de confiança para $\mu_{Y|x}$, quando $x=\overline{x}$, sendo y a variável `altura` e x a variável `areabasal`. A forma desse intervalo, a $(1-\alpha)\times 100\%$ de confiança, é muito semelhante ao intervalo de predição que consta do formulário, sendo a única diferença o facto de, debaixo da raíz quadrada, não constar a parcela "1+". Assim, é da forma:

$$\left]\ (b_0+b_1x)-t_{\frac{\alpha}{2};n-2}\sqrt{QMRE\left[\frac{1}{n}+\frac{(x-\overline{x})^2}{_{(n-1)}\,s_x^2}\right]}\ ,\ (b_0+b_1x)+t_{\frac{\alpha}{2};n-2}\sqrt{QMRE\left[\frac{1}{n}+\frac{(x-\overline{x})^2}{_{(n-1)}\,s_x^2}\right]}\ \right[\ .$$

   Os cálculos deste IC simplificam bastante pela escolha de $x=\overline{x}$, que anula a segunda parcela debaixo da raíz quadrada. O erro padrão associado será apenas da forma $\sqrt{\frac{QMRE}{n}}$. Sabemos que $n=167$. A fim de calcular o Quadrado Médio Residual, sabemos que $SQRE=SQT-SQR=SQT\,(1-R^2)=(n-1)\,s_y^2\,(1-R^2)=166\times(6.097297)^2\times(1-0.7758724)=1383.178$. Assim, $QMRE=\frac{1383.178}{165}=8.382898$, e o erro padrão associado ao intervalo de confiança é $\sqrt{\frac{QMRE}{n}}=0.2240469$. O ponto central do intervalo é $b_0+b_1\,x=8.012486+0.6480386\times 14.685629=17.52934$. Finalmente, e pelas tabelas, $t_{0.025(165)}\approx 1.97$. Logo, o intervalo pedido é] 17.08797 , 17.97071 [, um intervalo bastante preciso para a altura média nas parcelas com área basal $\overline{x}=14.685629$.

   (c) O gráfico da esquerda é um gráfico de resíduos usuais (no eixo vertical) sobre valores ajustados $\hat{y}$ (no eixo horizontal). A serem verdade os pressupostos do modelo de regressão linear, os pontos deveriam estar dispostos numa banda horizontal, sem quaisquer padrões especiais. No entanto, é visível uma curvatura na disposição dos pontos, com resíduos negativos nos dois extremos e maior frequência de valores positivos na parte central. Essa curvatura indica que a linearidade da equação do modelo não corresponde à tendência de fundo na relação entre `altura` e `areabasal`, já que a recta ajustada sobrestima as alturas em parcelas com área basal média extrema, e tende a sobrestimar nos restantes casos. Por

outro lado, é ainda visível uma tendência para aquilo que se designa um efeito em forma de funil, com dispersões de resíduos maiores à medida que se avança da esquerda para a direita no eixo horizontal. Esse padrão sugere a violação do pressuposto do modelo que exige variâncias homogéneas nos erros aleatórios.

O gráfico da direita é um *qq-plot*, de quantis empíricos dos resíduos estandardizados (no eixo vertical) contra quantis teóricos duma distribuição $\mathcal{N}(0,1)$ (no eixo horizontal). O pressuposto de Normalidade dos erros aleatórios, exigido pelo Modelo Linear, deve traduzir-se numa linearidade dos pontos neste gráfico, facto que se observa. Assim, não há razões para questionar a Normalidade dos ϵ_i.

(d) i. Uma relação linear entre $\ln(y)$ e $\ln(x)$ corresponde a uma relação potência entre y e x:

$$\ln(y) = b_0 + b_1 \ln(x) \quad \Leftrightarrow \quad y = \mathrm{e}^{b_0 + b_1 \ln(x)} \quad \Leftrightarrow \quad y = \mathrm{e}^{b_0}\mathrm{e}^{\ln(x^{b_1})} \quad \Leftrightarrow \quad y = \mathrm{e}^{b_0}\, x^{b_1} \ .$$

Assim, a curva ajustada entre `altura` (y) e `areabasal` (x) é $y = \mathrm{e}^{1.51442}\, x^{0.51361} \Leftrightarrow y = 4.546783\, x^{0.51361}$. Em termos muito aproximados, a altura média das árvores dominantes é proporcional à raíz quadrada da área basal.

ii. O valor do Coeficiente de Determinação neste modelo refere-se à proporção explicada na variabilidade das *log-alturas*, e não na variabilidade das alturas. Neste sentido, este Coeficiente de Determinação não é directamente comparável ao que se obteve no modelo linear sobre as variáveis não transformadas.

2. (a) O valor $R^2 = 0.8933$ indica que este modelo, com $p = 4$ preditores e $n = 167$ observações, explica quase 90% da variabilidade observada na variável resposta `altura`. Trata-se dum valor muito elevado, sendo de esperar que um teste de ajustamento global conduza à rejeição da Hipótese Nula. Tem-se:

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

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

Nível de significância: $\alpha = 0.05$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05[4,162]}$ que, pelas tabelas, é um valor entre os valores tabelados 2.37 e 2.45.

Conclusões: O valor calculado da estatística é $F_{calc} = 339.2$, sendo a rejeição de H_0 muito clara. Assim, o modelo ajustado é muito significativamente diferente do Modelo Nulo, como seria de esperar dado o valor muito elevado de R^2. O valor de prova (*p-value*) no enunciado é inferior à precisão de máquina, ou seja, é indistinguível de zero, confirmando a rejeição muito enfática da H_0.

(b) É pedido para iniciar um algoritmo de exclusão sequencial, baseado em testes t ao nível $\alpha = 0.05$, sobre o modelo completo original. Os testes dizem respeito às Hipóteses Nulas da forma $H_0 : \beta_j = 0$ e alternativas da forma $H_1 : \beta_j \neq 0$, e as variáveis x_j candidatas a exclusão são aquelas em que *não* se rejeita H_0 (para o β_j correspondente). A estatística do teste é, em cada caso, $T = \frac{\hat{\beta}_j - \beta_{j|H_0}}{\hat{\sigma}_{\hat{\beta}_j}} = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$, sabendo-se que a sua distribuição, caso seja verdade H_0, é $t_{n-(p+1)}$. Assim, as fronteiras da Região Crítica (bilateral) são dadas por $\pm t_{0.025(162)} \approx \pm 1.97$. Os valores das estatísticas de teste são dadas no enunciado, na coluna de nome '`t value`'. Constata-se que existem duas variáveis para as quais não se rejeita a Hipótese Nula, concretamente os preditores `idade` e `vivas`. Esta conclusão é corroborada pelos valores de prova (*p-values*), dados na coluna final, tendo-se, respectivamente, $p = 0.493$ e $p = 0.428$. Qualquer destas variáveis preditoras pode, individualmente, ser excluída do modelo sem prejuízo significativo na qualidade de ajustamento. Uma vez que apenas se

pode excluir um preditor em cada iteração do algoritmo, opta-se por excluir o preditor com o maior *p-value* (t_{calc} mais próximo de zero), que é aquele cuja exclusão menos prejudica a qualidade de ajustamento. Trata-se do preditor `idade`, ficando um submodelo com os três preditores `diam`, `vivas` e `areabasal`.

(c) É pedido um teste F parcial, para comparar o modelo de $p=4$ preditores e o submodelo de $k=1$ preditor, ajustado na alínea 1a). 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 $F_{calc} > f_{0.05[3,162]} \approx 2.65$.

Conclusões: Tem-se $F_{calc} = \frac{162}{3} \times \frac{0.8933-0.7758724}{1-0.8933} = 59.42915$. Logo, rejeita-se claramente H_0, isto é, considera-se que a qualidade de ajustamento do modelo completo difere significativamente (ao nível $\alpha=0.05$) da do submodelo. Deste modo, o modelo completo é preferível ao submodelo, resultado coerente com o que seria de esperar dados os valores dos Coeficientes de Determinação.

(d) Pede-se o valor do AIC que, pelo formulário, é dado por $AIC = n \ln\left(\frac{SQRE_k}{n}\right) + 2(k+1)$. No que respeita ao modelo ajustado neste ponto, o único valor que não está imediatamente disponível é $SQRE$, mas pode ser obtido a partir do valor de $\sqrt{QMRE} = 2.016$, dado no enunciado (com a designação `Std.Error`). Assim, $SQRE_4 = QMRE \times [n-(p+1)] = (2.016)^2 \times 162 = 658.4095$, pelo que (e tendo em conta que é o modelo completo, logo $k = p = 4$) $AIC_4 = 239.0961$. Quanto ao modelo de Regressão Linear Simples de `altura` sobre `areabasal`, tem-se $k=1$. Como se viu na alínea 1b), $SQRE_1 = 1383.178$, pelo que $AIC_1 = 357.0623$. Uma vez que menores valores do AIC correspondem a melhores modelos, também por este critério a opção será pelo modelo completo, de quatro preditores.

3. A informação suplementar, dada no enunciado, significa que é de duvidosa validade o pressuposto de estarmos perante $n=167$ observações com erros aleatórios *independentes*. De facto, é natural que as observações efectuadas numa mesma parcela, ao longo do tempo, estejam correlacionadas entre si. Por outras palavras, duas observações da variável resposta `altura`, efectuadas numa mesma parcela em dois momentos consecutivos, tendem a estar mais próximas entre si do que duas medições efectuadas em parcelas diferentes, mesmo após levar em conta os valores dos preditores usados no modelo. Assim, estamos perante uma situação em que observações numa mesma parcela são pseudo-repetições, e não repetições realmente independentes. Esse facto deve ser incorporado no modelo, mas de formas que ultrapassam o âmbito desta disciplina. Em todo o caso, deveria evitar-se ajustar os modelos como foi feita acima.

III

Estamos num contexto ANOVA, com a variável resposta dada pelo comprimento médio dos ramos (`comp`) e com dois factores que podem contribuir para explicar diferenças nos valores dessa variável: os sistemas de condução (`conducao`) e os terrenos (`bloco`).

1. O factor A, `conducao`, tem $a = 4$ níveis, enquanto o factor B, `bloco`, tem $b = 2$ níveis. O delineamento usado é de tipo *factorial*, uma vez que os quatro sistemas de condução foram

cruzados com os dois terrenos. Em cada uma das $ab = 8$ situações experimentais resultantes efectuou-se o mesmo número de observações: $n_c = 10$. Assim, estamos perante um delineamento *equilibrado*, e um total de $n = abn_c = 80$ observações.

Cada observação é identificada por uma tripla indexação: Y_{ijk} onde i indica o sistema de condução (factor A, `conducao`, podendo ter-se $i = 1, 2, 3, 4$); j indica o terreno (factor B, `bloco`, com $j = 1, 2$); e k indica a repetição no seio da célula (i, j) (com $k = 1, 2, ..., 10$). Uma vez que existem repetições nas células, pode usar-se um modelo ANOVA a dois factores, com efeitos de interacção:

- A equação do modelo é $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, sendo μ_{11} o rendimento esperado com o primeiro sistema de condução (Eixo, por ordem alfabética) no primeiro terreno; α_i o efeito principal (aumento esperado no comprimento) associado ao sistema de condução i; β_j o efeito principal associado ao terreno j; $(\alpha\beta)_{ij}$ o efeito de interacção entre o sistema de condução i e o terreno j; e sendo ϵ_{ijk} o erro aleatório associado à observação Y_{ijk}. Como em qualquer modelo ANOVA, é necessário impôr restrições que, neste caso, são dadas por $\alpha_1 = 0$; $\beta_1 = 0$; e $(\alpha\beta)_{ij} = 0$ se $i = 1$ e/ou $j = 1$.
- 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 variáveis aleatórias independentes.

Com base neste modelo, tem-se que o rendimento esperado com o sistema Solaxe ($i = 3$), no primeiro terreno ($j = 1$) é:

$$\mu_{31} = E[Y_{31k}] = E[\mu_{11} + \alpha_3 + \underbrace{\beta_1}_{=0} + \underbrace{(\alpha\beta)_{31}}_{=0} + \epsilon_{31k}] = \mu_{11} + \alpha_3 + \underbrace{E[\epsilon_{31k}]}_{=0} = \mu_{11} + \alpha_3 \ .$$

Assim, o parâmetro α_3 é a diferença das médias populacionais nas célula $(3, 1)$ e $(1, 1)$: $\alpha_3 = \mu_{31} - \mu_{11}$.

2. É pedido para indicar se os efeitos principais de sistema de condução são significativos. O teste F aos efeitos dos sistemas (α_i) permite responder à pergunta:

   Hipóteses: $H_0 : \alpha_i = 0 \ , \ \forall \ i = 2, 3, 4$ vs. $H_1 : \exists \, i = 2, 3, 4$ tal que $\alpha_i \neq 0$.

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

   Nível de significância: $\alpha = 0.05$.

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(3,72)} \approx 2.75$.

   Conclusões: É necessário calcular o valor da estatística de teste, omitido no enunciado. Sabe-se que $F_A = \frac{QMA}{QMRE}$, e que $QME = \frac{SQA}{a-1}$. Logo, $QMA = \frac{2572}{3} = 857.3333$, e $F_{calc} = \frac{857.3333}{176.1} = 4.86845 > 2.75$. Assim, rejeita-se H_0 (ao nível $\alpha = 0.05$), concluindo-se pela existência de efeitos principais significativos de sistema de condução. Por outras palavras, é possível afirmar que, pelo menos um sistema de condução, vai estar associado a um diferente comprimento médio dos ramos.

3. É pedido para efectuar um teste F aos efeitos principais de terrenos, ou seja do factor B, `bloco`. Como já se viu, existem apenas $b = 2$ níveis, pelo que após a restrição $\beta_1 = 0$, apenas existe um parâmetro desse tipo de efeitos: β_2. Eis o teste pedido:

   Hipóteses: $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$.

   Estatística do Teste: $F_B = \frac{QMB}{QMRE} \ \frown \ F_{[b-1 \, , \, n-ab]}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(1,72)} \approx 3.98$.

Conclusões: Como $F_{calc} = 11.570 > 3.98$, rejeita-se H_0. Pode concluir-se que $\beta_2 \neq 0$, ou seja, que a transição do primeiro para o segundo terreno irá introduzir um efeito significativo (ao nível $\alpha = 0.05$) na variável resposta `comp`.

Nota: O valor estimado desse efeito é dado (veja-se o formulário) por $\hat{\beta}_2 = \overline{Y}_{12.} - \overline{Y}_{11.} = 92.45 - 98.63 = -6.18$, pelo que se trata duma diminuição do comprimento dos ramos, no segundo terreno. Esta estimativa depende das restrições impostas no modelo.

4. Pede-se para estudar o gráfico de interacção dado no enunciado. Por cima dos marcadores de cada sistema de condução (no eixo horizontal) são dados dois pontos, correspondentes aos dois terrenos, e cujas ordenadas no eixo vertical correspondem ao valor médio da célula que combina esse sistema de condução e terreno. Os quatro pontos que correspondem a cada terreno são unidos por segmentos de recta. No gráfico é visível a conclusão a que se chegou na alínea anterior: globalmente, os comprimentos dos ramos descem quando passamos do primeiro para o segundo terreno (descida essa que já vimos ser significativa, ao nível $\alpha = 0.05$). No entanto, o gráfico saliente outra característica interessante: essa descida não parece ser uniforme (sendo, por exemplo, muito maior para o sistema Palmeta do que para o sistema Eixo), tendo-se mesmo, para o sistema Solaxe, um aumento no comprimento médio dos ramos no segundo terreno. Esta ausência de 'paralelismo' indicia a existência de efeitos de interacção. Para determinar se esses efeitos são, ou não, significativos, deve efectuar-se um teste F aos efeitos de interacção. Tem-se:

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

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

   Nível de significância: $\alpha = 0.05$.

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(3,72)} \approx 2.75$.

   Conclusões: Como $F_{calc} = 4.639 > 2.75$, rejeita-se H_0, ao nível $\alpha = 0.05$. Pode concluir-se que existem efeitos de interacção significativos, como sugerido pelo gráfico de interacção.

5. Nesta alínea é pedido para utilizar a teoria de Tukey para comparar a média populacional da situação experimental resultante de usar o sistema de condução Solaxe ($i = 3$) no segundo terreno ($j = 2$), ou seja, μ_{32}, com as restantes. A inspecção do gráfico de interacção da alínea anterior sugere que diferenças significativas devam surgir apenas nas comparações com outras médias populacionais do segundo terreno. Mas é necessário verificar.

   Ao nível global de significância $\alpha = 0.05$, o termo de comparação de Tukey é dado por:

$$q_{\alpha(ab\,,\,n-ab)} \sqrt{\frac{QMRE}{n_c}} = q_{0.05(8,72)} \sqrt{\frac{176.1}{10}} \approx 4.42 \times 4.196427 = 18.54821 \ .$$

   Ora, na situação experimental $(3, 2)$ tem-se a segunda maior média amostral: $\overline{y}_{32.} = 103.65$. Sempre que $|\overline{y}_{ij.} - \overline{y}_{32.}| > 18.54821$, deve concluir-se que $\mu_{ij} \neq \mu_{32}$ (com nível global de significância 0.05). É evidente que essa desigualdade não se verifica para a célula $(4, 1)$, que tem a maior média amostral ($\overline{y}_{41.} = 105.35$). Para as restantes, sempre que um rendimento médio amostral seja inferior a $103.65 - 18.54821 = 85.102$ conclui-se estar associado a uma média populacional diferente de μ_{32}. Tal situação apenas ocorre numa situação experimental, a que cruza o sistema Palmeta com o segundo terreno (para o qual a média amostral é $\overline{y}_{22.} = 73.48$).

IV

1. (a) O resíduo usual de qualquer observação i é a diferença entre o seu valor observado da variável resposta e o correspondente valor ajustado pelo modelo: $e_i = y_i - \hat{y}_i$. No nosso caso, os valores ajustados pelo modelo são da forma $\hat{y}_i = b\,x_i$. Logo, um resíduo é da forma $e_i = y_i - b\,x_i$.

   (b) A Soma de Quadrados dos Resíduos é assim dada por $SQRE = \sum_{i=1}^{n} (y_i - b\,x_i)^2$. A fim de minimizar esta Soma de Quadrados, vista como função do parâmetro desconhecido b, é necessário anular a sua derivada em ordem a b, ou seja, tomar:

$$[SQRE(b)]' = \sum_{i=1}^{n} 2\,(y_i - b\,x_i)(-x_i) = -\sum_{i=1}^{n} 2\,y_i\,x_i + \sum_{i=1}^{n} 2\,b\,x_i^2 = 0$$

$$\Leftrightarrow \quad b = \frac{\sum_{i=1}^{n} x_i\,y_i}{\sum_{i=1}^{n} x_i^2}\,.$$

   Nota 1: Repare-se como se trata duma fórmula análoga à do declive da recta de regressão usual, mas em que os momentos centrados (covariância e variância de x) são substituídos pelos correspondentes momentos não centrados.

   Nota 2: Que este ponto crítico é o valor de b que *minimiza* a Soma de Quadrados Residual é intuitivo no contexto, mas pode ser confirmado tomando a segunda derivada de $SQRE(b)$, que é $[SQRE(b)]'' = \sum_{i=1}^{n} 2\,x_i^2$, quantidade que é sempre positiva (esta segunda derivada só poderia ser nula se em todas as observações, $x_i = 0$, situação sem sentido no contexto estatístico em causa). Logo, para o valor de b acima obtido, $SQRE$ tem um mínimo.

   (c) A matriz $\mathbf{X}$ deste modelo sem constante aditiva apenas terá uma coluna: a coluna dos n valores x_i do preditor (recorde-se que a coluna dos uns estava associada à constante aditiva b_0, que não existe neste modelo forçado à origem). Assim $\mathbf{X} = (x_1, x_2, ..., x_n)^t$ e, usando a fórmula (constante do formulário), tem-se $\vec{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{y}}$. Aqui, $\vec{\boldsymbol{\beta}}$ tem um único elemento: o estimador de mínimos quadrados para b. O produto matricial $\mathbf{X}^t\mathbf{X}$ é igualmente um escalar, dado por $\sum_{i=1}^{n} x_i^2$. A sua inversa matricial é o seu recíproco. Finalmente o produto $\mathbf{X}^t\vec{\mathbf{y}}$ é dado por $\sum_{i=1}^{n} x_i\,y_i$. Assim, tem-se que a estimativa procurada é $b = \frac{\sum_{i=1}^{n} x_i\,y_i}{\sum_{i=1}^{n} x_i^2}$, como já se vira na alínea anterior.

   (d) A média dos valores ajustados $\hat{y}_i$ é dada por $\overline{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \frac{1}{n}\sum_{i=1}^{n} b\,x_i = b\,\overline{x} = \frac{\overline{x}\sum_{i=1}^{n} x_i\,y_i}{\sum_{i=1}^{n} x_i^2}$. Esta expressão não é algebricamente igual à média $\overline{y}$ dos y_i observados, ao contrário do que acontece com a recta de regressão usual.

   Nota: Qualquer dúvida que estas expressões podem ser diferentes é eliminada por um simples contra-exemplo de $n = 2$ observações: $(x_1, y_1) = (1, 0)$ e $(x_2, y_2) = (0, 1)$. Aqui tem-se $\overline{y} = 0.5$, mas $\overline{\hat{y}} = \frac{\overline{x}\sum_{i=1}^{n} x_i\,y_i}{\sum_{i=1}^{n} x_i^2} = \frac{0.5\times 0}{1+0} = 0$.

2. (a) Sabemos que em qualquer modelo ANOVA, SQT é a soma das restantes Somas de Quadrados envolvidos na decomposição para esse modelo. Além disso, cada Soma de Quadrados pode ser escrita como o produto do respectivo Quadrado Médio e os seus graus de liberdade. Assim, e já que $QMT = \frac{SQT}{n-1} = s_y^2$, temos no modelo M_{A+B}:

$$\begin{aligned} SQT &= SQA + SQB + SQRE \\ QMT\,(n-1) &= QMA\,(a-1) + QMB\,(b-1) + QMRE\,[n-(a+b-1)] \\ s_y^2 &= \frac{(a-1)\,QMA + (b-1)\,QMB + [n-(a+b-1)]\,QMRE}{n-1}\,. \end{aligned}$$

Uma vez que o denominador $(n-1)$ é a soma dos graus de liberdade associadas às três parcelas do numerador (ou seja, $(a-1)+(b-1)+n-(a+b-1)=n-1$), a expressão final diz-nos que s_y^2 é uma média ponderada de QMA, QMB e $QMRE$, sendo os pesos de cada um desses Quadrados Médios dado pelos respectivos graus de liberdade.
O caso do modelo M_{A*B} faz-se de forma análoga:

$$\begin{aligned} SQT &= SQA + SQB + SQAB + SQRE \\ QMT\,(n-1) &= QMA\,(a-1) + QMB\,(b-1) + QMAB\,(a-1)(b-1) + QMRE\,[n-ab] \\ s_y^2 &= \frac{(a-1)\,QMA + (b-1)\,QMB + (a-1)(b-1)\,QMAB + [n-ab]\,QMRE}{n-1}\,. \end{aligned}$$

(b) Tem-se, a partir das definições respectivas, que:

$$\begin{aligned} QMRE_{A*B} > QMRE_{A+B} &\Leftrightarrow \frac{SQRE_{A*B}}{n-ab} > \frac{SQRE_{A+B}}{n-(a+b-1)} = \frac{SQAB + SQRE_{A*B}}{n-(a+b-1)} \\ &\Leftrightarrow \frac{SQRE_{A*B}}{n-ab} - \frac{SQRE_{A*B}}{n-(a+b-1)} > \frac{SQAB}{n-(a+b-1)} \\ &\Leftrightarrow SQRE_{A*B}\left[\frac{\cancel{n} - (a+b-1) - (\cancel{n}-ab)}{(n-ab)\cancel{[n-(a+b-1)]}}\right] > \frac{SQAB}{\cancel{n-(a+b-1)}} \\ &\Leftrightarrow \frac{SQRE_{A*B}}{n-ab} > \frac{SQAB}{ab-a-(b-1)} = \frac{SQAB}{(a-1)(b-1)} \\ &\Leftrightarrow QMRE_{A*B} > QMAB\,. \end{aligned}$$

Esta equivalência entre desigualdades significa que, sempre que no teste aos efeitos de interacção no modelo M_{A*B} a estatística de teste tomar valores $F_{AB} = \frac{QMAB}{QMRE_{A*B}} < 1$, o modelo com efeitos de interacção terá um Quadrado Médio Residual superior ao modelo sem efeitos de interacção. Assim, um modelo com efeitos de interacção pode ter uma qualidade de ajustamento pior do que um modelo sem efeitos de interacção.