INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

10 de Janeiro, 2019 **EXAME - 1a. Chamada 2018-19 Uma resolução possível**

I

Há $N=96$ contagens do número de não pegamentos em $m=20$ provas. Assim, caso essas contagens sejam governadas por uma lei Binomial, terá de ser uma Binomial de parâmetro $m=20$. O valor de p é desconhecido.

1. Para estimar p, recorde-se que o valor esperado dum variável aleatória $X \frown B(m,p)$ é $E[X]=mp$. Esse valor esperado pode ser estimado pela média amostral, $\overline{x}$, pelo que a estimativa $\hat{p}$ terá de verificar a equação $\overline{x}=m\hat{p}$. No nosso exemplo tem-se $\overline{x}=\frac{(0\times45)+(1\times30)+(2\times13)+(3\times8)}{96}=0.8333333$. Logo, $\hat{p}=\frac{\overline{x}}{m}=\frac{0.8333333}{20}=0.04166667$ é o valor estimado de p que melhor corresponde aos dados observados.

2. A distribuição χ^2 da estatística do teste de Pearson é apenas assintótica. As condições de Cochran visam garantir que a amostra é de dimensão suficiente para se admitir aproximadamente válida essa distribuição assintótica. As condições de Cochran incidem sobre as contagens esperadas (E_i), e exigem que em nenhuma contagem o valor de E_i seja inferior a 1, e que não haja valores $E_i<5$ em mais do que um quinto das contagens. No nosso caso, os valores esperados em cada uma das $k=4$ categorias de contagem são estimados, pois foi necessário estimar o parâmetro p.

   Os valores esperados estimados são dados por $\hat{E}_i = N \times \hat{\pi}_i$, onde $\hat{\pi}_i$ indica as probabilidades correspondentes a cada categoria de contagem, numa distribuição $B(m,\hat{p})$. Usando o valor estimado $\hat{p}=0.05$ (conforme indicado no enunciado), será necessário calcular as probabilidades dos valores 0, 1, 2, e dos valores maiores ou iguais a 3, numa Binomial $B(20,0.05)$ (recorde-se que a soma de probabilidades em todas as categorias de contagem tem de ser igual a 1, pelo que a última categoria deve ser considerada a classe composta por valores $X\geq 3$). Essas probabilidades podem ser calculadas directamente pela expressão das probabilidades duma Binomial: $P[X=x]=\binom{m}{x}\hat{p}^x\,(1-\hat{p})^{m-x}$ (alternativamente, poderiam ser usadas as tabelas da Binomial). Temos então os seguintes valores esperados estimados:

   - $\hat{E}_0 = N\times\hat{\pi}_0 = 96\cdot\binom{20}{0}\cdot 0.05^0\cdot 0.95^{20} = 96\cdot 0.95^{20} = 96\cdot 0.3584859 = 34.41465$;
   - $\hat{E}_1 = N\times\hat{\pi}_1 = 96\cdot\binom{20}{1}\cdot 0.05^1\cdot 0.95^{19} = 96\cdot 20\cdot 0.05\cdot 0.3773536 = 96\cdot 0.3773536 = 36.22595$;
   - $\hat{E}_2 = N\times\hat{\pi}_2 = 96\cdot\binom{20}{2}\cdot 0.05^2\cdot 0.95^{18} = 96\cdot 190\cdot 0.0025\cdot 0.3972143 = 96\cdot 0.1886768 = 18.11297$;
   - $\hat{E}_{\geq 3} = N\times\widehat{P[X\geq 3]} = 96\cdot[1-(\hat{\pi}_0+\hat{\pi}_1+\hat{\pi}_2)] = 96\cdot[1-(0.3584859+0.3773536+0.1886768)] = 96\cdot 0.0754837 = 7.246435$.

   Assim, nenhum valor esperado estimado é sequer inferior a 5, pelo que se pode admitir a validade da distribuição assintótica χ^2.

3. Eis o teste pedido:

   Hipóteses: $H_0 : X \frown B(20,\hat{p}=0.05)$ *vs.* $H_1 : X \not\frown B(20,\hat{p}=0.05)$.

Estatística do Teste: A estatística de Pearson, é dada por $X^2 = \sum_i \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i}$. Havendo $r = 1$ parâmetro estimado, a distribuição assintótica desta estatística, caso seja verdade H_0, é χ^2_{k-r-1}, com $k-r-1 = 4-1-1 = 2$.

Nível de Significância Não sendo explicitado no enunciado, pode-se escolher $\alpha = 0.05$.

Região Crítica: (Unilateral direita) A regra de rejeição consiste em rejeitar H_0 se $\chi^2_{\text{calc}} > \chi^2_{0.05(2)} = 5.991$.

Conclusões Vamos calcular o valor da estatística para os nossos dados. Tem-se:

$$X^2_{calc} = \frac{(45 - 34.41465)^2}{34.41465} + \frac{(30 - 36.22595)^2}{36.22595} + \frac{(13 - 18.11297)^2}{18.11297} + \frac{(8 - 7.246435)^2}{7.246435} = 5.847554 \, .$$

Logo, embora por pouco, não se rejeita H_0, podendo admitir-se a hipótese de as contagens seguirem a distribuição Binomial indicada, ao nível de significância $\alpha = 0.05$.

II

1. (a) 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 r_{xy} o coeficiente de correlação entre o preditor x (`nfrColh`) e a variável resposta y (`producao`), e s_x e s_y os respectivos desvios padrão. Estes três valores estão disponíveis no enunciado, tendo-se $b_1 = 0.9853879 \times \frac{0.2975462}{2.3597379} = 0.1242504$. Para calcular a ordenada na origem usa-se a fórmula $b_0 = \overline{y} - b_1\, \overline{x}$, onde $\overline{y} = 0.5221$ e $\overline{x} = 4.025$ são as médias da produção e do número de frutos em Setembro, respectivamente, e onde b_1 é o declive atrás calculado. Logo, $b_0 = 0.5221 - 0.1242504 \times 4.025 = 0.02199214$. A recta ajustada tem, pois, equação $y = 0.02199214 + 0.1242504\, x$. O Coeficiente de Determinação numa Regressão Linear Simples é o quadrado do coeficiente de correlação entre o preditor e a variável resposta, pelo que $R^2 = (0.9853879)^2 = 0.9709893$. Assim, a recta explica cerca de $97,1\%$ da variabilidade das produções observadas na amostra.

   (b) É pedido para calcular o Quadrado Médio Residual que, em qualquer Modelo Linear, é usado para estimar a variância dos erros aleatórios do Modelo, ou seja, σ^2. Conhecemos o valor do Coeficiente de Determinação R^2, e sabemos que, por definição, $R^2 = \frac{SQR}{SQT}$. Ora, $SQT = (n-1)\, s_y^2 = 148 \times (0.2975462)^2 = 13.10299$. Assim, $SQR = R^2 \times SQT = 0.9709893 \times 13.10299 = 12.72287$. Além disso, $SQRE = SQT - SQR = 13.10299 - 12.72287 = 0.3801237$. Logo, $QMRE = \frac{SQRE}{n-2} = \frac{0.3801237}{147} = 0.002586$.

   (c) É pedido um intervalo a 95% de condiança para o declive da recta populacional, β_1. A expressão geral deste tipo de intervalos a $(1-\alpha) \times 100\%$ de confiança é

$$\left] \, b_1 - t_{\frac{\alpha}{2};n-2} \cdot \hat{\sigma}_{\hat{\beta}_1} \, , \; b_1 + t_{\frac{\alpha}{2};n-2} \cdot \hat{\sigma}_{\hat{\beta}_1} \, \right[\, .$$

   O ponto central do IC é o declive da recta amostral, já acima calculado: $b_1 = 0.1242504$. O erro padrão associado ao declive é dado por $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{QMRE}{(n-1)\, s_x^2}}$ (a expressão da verdadeira variância do estimador $\hat{\beta}_1$ é dada no formulário, sendo $\frac{\sigma^2}{(n-1)\, s_x^2}$; o erro padrão de $\hat{\beta}_1$ é a raíz quadrada desta expressão, substituindo o valor desconhecido σ^2 pela sua estimativa $QMRE$). Tem-se $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{QMRE}{(n-1)\, s_x^2}} = \sqrt{\frac{0.002586}{148 \times (2.3597379)^2}} = 0.00177$. Finalmente, $t_{0.025(147)} \approx 1.97$. Assim, o intervalo a 95% de confiança é o intervalo $]\, 0.1208 \, , \, 0.1277 \, [$. Trata-se dum intervalo com boa precisão, que apenas introduz incerteza na terceira casa decimal do valor de β_1.

(d) É dada uma observação com $x_i = 9.9$ e $y_i = 0.932$. Os valores permitem identificá-la no gráfico como a observação mais à direita (o valor $x_i = 9.9$ é o valor máximo do preditor `nfrColh`), deslocada para baixo em relação à tendência geral dos pontos observados. É de prever que esta observação tenha um resíduo negativo, relativamente importante. O seu efeito alavanca será o maior de todos, tendo em conta que o valor dos efeitos alavanca numa regressão linear simples cresce com o afastamento do valor x_i em relação à média $\overline{x}$, pelo que o maior efeito alavanca corresponderá sempre, ou à observação com o menor valor de x, ou à observação com o maior valor de x, consoante o extremo mais afastado de $\overline{x}$. Neste caso, e uma vez que $\overline{x} = 4.025$, o maior efeito alavanca corresponderá à observação do enunciado. Passemos aos cálculos pedidos.

i. O resíduo usual é dado por:

$$e_i = y_i - \hat{y}_i = 0.932 - (b_0 + b_1 \cdot 9.9) = 0.932 - (0.02199214 + 0.1242504 \times 9.9) = -0.3200711 \ .$$

ii. Pelo formulário, sabemos que o valor do efeito alavanca é dado por:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{(n-1)\, s_x^2} = \frac{1}{149} + \frac{(9.9 - 4.025)^2}{148 \times (2.3597379)^2} = 0.04859332 \ .$$

Não é um valor em si mesmo muito elevado, uma vez que a gama de variação possível dos valores de h_{ii} oscila entre $\frac{1}{n} = 0.006711409$ e 1, e que o valor médio dos efeitos alavanca nesta Regressão Linear Simples é $\frac{2}{n} = 0.01342282$. Mas como já se viu, é o maior efeito alavanca de qualquer das 149 observações.

iii. Pelo formulário, sabemos que o resíduo internamente estandardizado é dado por:

$$R_i = \frac{e_i}{\sqrt{QMRE\,(1 - h_{ii})}} = \frac{-0.3200711}{\sqrt{0.002586 \times (1 - 0.04859332)}} = -6.452813 \ .$$

Sabendo-se que os resíduos estandardizados estão geralmente compreendidos num intervalo do tipo $]-3, 3[$, é evidente que se trata dum resíduo muito importante, como era de esperar tendo em conta o afastamento relativo deste ponto em relação à tendência geral da nuvem de pontos.

iv. Pelo formulário, sabemos que a distância de Cook pode ser calculada da seguinte forma:

$$D_i = R_i^2 \left(\frac{h_{ii}}{1 - h_{ii}} \right) \frac{1}{p+1} = (-6.452813)^2 \times \frac{0.04859332}{1 - 0.04859332} \times \frac{1}{2} = 1.0634 \ .$$

Trata-se dum valor muito grande de distância de Cook, muito acima do limiar 0.5. Esta observação é muito influente, ou seja, a sua exclusão provocaria alterações importantes na recta ajustada. Inspeccionando o gráfico, é possível concluir que a sua presença atrai para baixo a recta ajustada, na parte direita da nuvem de pontos, ou seja,é uma observação que contribui para diminuir o declive da recta ajustada. O valor elevado de D_i é de certa forma previsível, uma vez que sendo a observação com o maior efeito alavanca, é também uma observação com um grande resíduo estandardizado. A observação é assinalável a vários títulos, e importaria confirmá-la, a fim de verificar se corresponde a uma observação legítima.

2. (a) O valor $R^2 = 0.9308$ indica que este modelo com $p = 4$ preditores e $n = 149$ observações, explica $93,08\%$ da variabilidade observada na variável resposta `producao`, um valor muito elevado. É 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 $F_{calc} > f_{\alpha[4,144]}$ que, pelas tabelas, é um valor entre os valores tabelados 2.37 e 2.45.

Conclusões: No enunciado está omisso o valor calculado da estatística F, mas esse valor pode ser obtido pela fórmula, já que é conhecido R^2. Tem-se $F_{calc} = 484.2312$. A rejeição de H_0 é muito clara, pelo que o modelo ajustado é muito significativamente diferente do Modelo Nulo, como seria de esperar dado o valor muito elevado de R^2.

(b) Sabemos que, numa Regressão Linear, o AIC (cuja expressão é dada no formulário), é constituido por duas parcelas, a primeira das quais mede a qualidade do ajustamento do modelo (através do valor de $SQRE$) e a segunda mede a complexidade do modelo (através do número de parâmetros do modelo, $k+1$). Em ambos os casos, menores valores da parcela indicam um melhor modelo: mais bem ajustado, isto é com menor Soma de Quadrados Residual, na primeira parcela; e mais parcimonioso, no caso da segunda parcela. Os AICs de modelos diferentes são comparáveis, mesmo que não se trate (como é o caso) de modelos encaixados, ou seja, de um modelo e submodelo. Apenas é necessário que a variável resposta seja igual e os dados com que se ajustaram os modelos sejam os mesmos (como é o caso). Ora, na regressão linear simples há um modelo mais parcimonioso (um único preditor) e que tem uma melhor qualidade de ajustamento, já que tem um R^2 superior: $R^2 = 0.9709893$. Assim, mesmo sem calcular o valor dos dois AICs, é possível assegurar que ambas as parcelas do modelo de regressão linear simples são mais pequenas, pelo que o respectivo AIC é também menor. Assim, o modelo de regressão linear simples de `producao` sobre `nfrColh` é preferível ao modelo de quatro preditores, ao abrigo do critério AIC.

Nota: Contraste-se a situação desta alínea com a tradicional utilização do AIC para comparar um modelo com um seu submodelo: nesse caso, o submodelo é sempre mais parcimonioso, mas tem necessariamente um ajustamento igual ou pior (um $SQRE$ igual ou mais elevado). Não é possível saber antecipadamente se o valor do AIC no submodelo será menor que no modelo completo: isso dependerá da relação entre a perda na primeira parcela do AIC e o ganho na segunda parcela. Apenas efectuando as contas será possível sabê-lo.

(c) É pedido um teste t ao valor de β_4, e mais concretamente o seguinte teste.

Hipóteses: $H_0 : \beta_4 = 0 \quad vs. \quad H_1 : \beta_4 \neq 0.$

Estatística do Teste: $T = \frac{\hat{\beta}_4 - \beta_{4|H_0}}{\hat{\sigma}_{\hat{\beta}_4}} \cap t_{n-(p+1)}$, 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: (Bilateral) Rejeitar H_0 se $|T_{calc}| > t_{0.025(144)} \approx 1.98$.

Conclusões: Tem-se $T_{calc} = \frac{b_4 - 0}{\hat{\sigma}_{\hat{\beta}_4}} = \frac{0.116200}{0.015066} = 7.713$. Logo, rejeita-se claramente H_0, pelo que a exclusão do preditor `nfrSet` piora de forma significativa (ao nível $\alpha = 0.05$) a qualidade do ajustamento.

(d) É pedido um teste F parcial, para comparar o modelo de $p = 4$ preditores inicial e o submodelo de $k = 2$ preditores desta alínea. Tem-se:

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

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

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

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05[2,144]} \approx 3.05$.

Conclusões: Tem-se $F_{calc} = \frac{144}{2} \times \frac{0.9308-0.8954}{1-0.9308} = 36.83237$. Logo, rejeita-se 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, apesar da proximidade dos respectivos coeficientes de determinação.

(e) É pedido para considerar 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).

Os valores das estatísticas de teste são dadas no enunciado, na coluna de nome '`t value`' e, com uma excepção, os respectivos valores de prova (*p-values*) surgem ao lado, na coluna final. Existe pelo menos uma variável candidata à exclusão, já que no teste associado ao preditor `nfrJun`, o *p-value* é $p = 0.0860 > 0.05$, não se rejeitando a Hipótese Nula $\beta_3 = 0$. Embora esteja omisso o *p-value*, já se efectuou na alínea 2c) o teste associado ao último preditor, concluindo-se que não era dispensável. Há, pois, uma única variável preditora candidata a sair, `nfrJun`, sendo o submodelo resultante deste primeiro passo o que inclui os três restantes preditores. Este resultado pode, à primeira vista, parecer surpreendente, dado que o coeficiente de correlação entre `nfrJun` e `producao` é muito elevado ($r = 0.9444425$), pelo que `nfrJun` é um bom preditor de `producao`. A sua exclusão logo no primeiro passo do algoritmo resulta do facto de se tratar dum preditor altamente correlacionado com outro preditor, `nfrSet`, que por sua vez está ainda mais fortemente correlacionado com `producao`. Assim, desde que este último preditor permaneça no modelo, a contribuição *adicional* do preditor `nfrJun` para a previsão da produção é marginal, e esse preditor é descartável. Dito de outra forma: o conhecimento do número de frutos em Junho é uma boa maneira de prever a produção final. Mas o número de frutos em Setembro (mais perto da data da colheita) é um ainda melhor preditor, que dispensa o conhecimento do número de frutos na data anterior.

III

É evidente que se está num contexto ANOVA, com a variável resposta dada pelo *rendimento*. Trata-se duma questão muito semelhante à do Exercício ANOVA 13 das aulas práticas.

1. Existem dois factores para explicar o rendimento: o *local* (com dois níveis, Régua e Tabuaço), e o *ano*. O mero facto de os anos em cada local serem diferentes permite concluir que *não* estamos perante um delineamento factorial (em cujo caso todos os anos teriam de surgir combinados com ambos os locais). Estamos perante um delineamento hierarquizado, em que o factor *ano* está subordinado ao factor *local*, tendo a experiência sido feita na Régua em $b_1 = 5$ anos diferentes e no Tabuaço em $b_2 = 2$ diferentes anos. Em cada uma das $b_1 + b_2 = 7$ situações experimentais há o mesmo número de observações: $n_c = 8$. Assim, estamos perante um delineamento equilibrado, com um total de $7 \times 8 = 56$ observações. Cada uma dessas observações é identificada por uma tripla indexação: Y_{ijk} onde i indica o nível do factor dominante (*local*, logo $i = 1, 2$); j indica o *ano* (podendo, na Régua, ter-se $j = 1, 2, 3, 4, 5$, e no Tabuaço $j = 1, 2$). O modelo ANOVA para este delineamento hierarquizado é o seguinte:

   - A equação do modelo é $Y_{ijk} = \mu_{11} + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$, sendo μ_{11} o rendimento esperado na primeira localidade (Régua) no primeiro ano aí observado (1999); α_i o efeito principal

(aumento esperado no rendimento) associado à localidade i (com a restrição $\alpha_1 = 0$, pelo que apenas sobra o parâmetro α_2, do efeito principal associado a Tabuaço); $\beta_{j(i)}$ o efeito associado ao ano j da localidade i (com a restrição $\beta_{1(i)} = 0$, para qualquer localidade $i = 1, 2$); e sendo ϵ_{ijk} 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 variáveis aleatórias independentes.

Com base neste modelo, tem-se que o rendimento esperado no Tabuaço ($i = 2$) em 1999 ($j = 1$ no Tabuaço) é dado (tendo em conta as propriedades dos valores esperados) por:

$$\mu_{21} = E[Y_{21k}] = E[\mu_{11} + \alpha_2 + \underbrace{\beta_{1(2)}}_{=0} + \epsilon_{21k}] = \mu_{11} + \alpha_2 + \underbrace{E[\epsilon_{21k}]}_{=0} = \mu_{11} + \alpha_2 \ .$$

Já para o Tabuaço ($i = 2$) em 2003 ($j = 2$), tem-se:

$$\mu_{22} = E[Y_{22k}] = E[\mu_{11} + \alpha_2 + \beta_{2(2)} + \epsilon_{22k}] = \mu_{11} + \alpha_2 + \beta_{2(2)} + \underbrace{E[\epsilon_{22k}]}_{=0} = \mu_{11} + \alpha_2 + \beta_{2(2)} \ .$$

Assim, o parâmetro $\beta_{2(2)}$ corresponde à diferença no rendimento médio populacional no Tabuaço, nos dois anos em que o estudo abrangeu essa localidade.

2. O quadro de síntese desta ANOVA tem uma linha associada a cada tipo de efeito previsto no modelo (Factor dominante A, *local*; e Factor subordinado B, *ano*), e ainda uma linha correspondente à variabilidade Residual. Para obter as quantidades correspondentes à tabela, podem usar-se os valores disponíveis no enunciado, as fórmulas disponíveis no formulário, bem como a conhecida relação de que as três Somas de Quadrados totalizam SQT. Assim, tem-se:

   - $g.l.(SQA) = a - 1 = 1$;
   - $g.l.(SQB(A)) = (b_1 - 1) + (b_2 - 1) = 4 + 1 = 5$;
   - $g.l.(SQRE) = n - (b_1 + b_2) = 56 - 7 = 49$;
   - $SQA = 0.6402$ (enunciado);
   - $SQRE = 8.311$ (enunciado);
   - $SQB(A) = SQT - (SQA + SQRE) = (n-1)s_y^2 - (0.6402 + 8.311) = 55 \times 0.329922 - 8.9512 = 9.19451$.

   Os Quadrados Médios obtêm-se dividindo cada Soma de Quadrados pelos respectivos graus de liberdade, e o valor das duas estatística F obtem-se dividindo o Quadrado Médio de cada tipo de efeito pelo Quadrado Médio Residual. Assim, a tabela completa é a seguinte:

Variação	g.l.	Soma de Quadrados	Quadrado Médio	F_{calc}
Local	$a-1=1$	$SQA=0.6402$	$QMA=0.6402$	$F_A=3.7745$
Anos	$(b_1-1)+(b_2-1)=5$	$SQB(A)=9.19451$	$QMB(A)=1.838902$	$F_{B(A)}=10.8418$
Residual	$n-(b_1+b_2)=49$	$SQRE=8.311$	$QMRE=0.1696122$	–

3. É pedido para indicar se os efeitos de ano ($\beta_{j(i)}$) são significativos. O teste F a esses efeitos permite responder à pergunta:

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

Estatística do Teste: $F_{B(A)} = \frac{QMB(A)}{QMRE} \cap F_{\left[\sum_{i=1}^{a}(b_i-1)\,,\,n-\sum_{i=1}^{a} b_i\right]}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(5,49)} \approx 2.40$.

Conclusões: Como $F_{calc} = 10.8418 > 2.40$, rejeita-se H_0, concluindo-se pela existência de efeitos significativos de ano (ao nível $\alpha = 0.05$). Assim, a variabilidade de ano para ano é importante, e caso tivesse sido ignorada (tratando anos diferentes como meras repetições, ou apenas realizando a experiência num único ano), estar-se-ia a ignorar uma importante fonte de variabilidade dos rendimentos, o que poderia mascarar a existência de efeitos de localidade, mesmo quando estes estejam presentes.

4. É pedido para efectuar um teste F aos efeitos principais do factor dominante *local*. Como já se viu, existem apenas $a = 2$ níveis, pelo que após a restrição $\alpha_1 = 0$, apenas existe um parâmetro desse tipo de efeitos: α_2. Eis o teste pedido:

   Hipóteses: $H_0 : \alpha_2 = 0$ vs. $H_1 : \alpha_2 \neq 0$.

   Estatística do Teste: $F_A = \frac{QMA}{QMRE} \cap F_{\left[a-1\,,\,n-\sum_{i=1}^{a} b_i\right]}$, sob H_0.

   Nível de significância: $\alpha = 0.05$.

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(1,49)} \approx 4.04$.

   Conclusões: Como $F_{calc} = 3.7745 < 4.04$, não se rejeita H_0, pelo que não se pode concluir pela existência de efeitos significativos de local (ao nível $\alpha = 0.05$ ou inferior). Por outras palavras, a mera transição de um local para outro não permite afirmar que o rendimento populacional difere.

   Nota 1: Um olhar para as médias em cada uma das 7 situações experimentais permite compreender o porquê desta conclusão: havendo uma enorme variabilidade nos rendimentos entre anos diferentes na Régua, o rendimento médio verificado nos cinco anos na Régua foi 0.992 kg/planta, não havendo sustentação para a conclusão de que seja significativamente diferente do rendimento médio observado no Tabuaço: 0.7553.

   Nota 2: Uma vez que as Hipóteses neste teste envolvem um único parâmetro (α_2), e uma vez que os modelos ANOVA são Modelos Lineares, seria possível igualmente efectuar um teste t às mesmas hipóteses. Os resultados desse teste alternativo seriam equivalentes.

5. Nesta alínea é pedido para utilizar a teoria de Tukey para comparar a média populacional da situação experimental Régua ($i = 1$) em 1999 ($j = 1$), ou seja, μ_{11}, com as restantes. Nessa situação experimental tem-se a menor média amostral: $\overline{y}_{11.} = 0.291$. Ao nível global de significância $\alpha = 0.05$, o termo de comparação de Tukey é dado por:

$$q_{\alpha(b_1+b_2\,,\,n-(b_1+b_2))}\sqrt{\frac{QMRE}{n_c}} = q_{0.05(7,49)}\sqrt{\frac{0.1696122}{8}} \approx 4.34 \times 0.1456074 = 0.6319363\ .$$

   Sempre que $|\overline{y}_{ij.} - \overline{y}_{11.}| > 0.6319363$, deve concluir-se que $\mu_{ij} \neq \mu_{11}$ (com nível global de significância 0.05). Assim, qualquer rendimento médio amostral superior a $0.291 + 0.6319363 = 0.9229$ corresponde a uma média populacional que é significativamente diferente de μ_{11}. Tal situação apenas ocorre com dois outros anos na Régua: 2002 (para o qual a média amostral é $\overline{y}_{13.} = 1.327$) e 2003 (para o qual a média amostral é $\overline{y}_{14.} = 1.682$). Assim, não é possível afirmar que o menor dos rendimentos amostrais médios seja significativamente diferente de *todos* os outros rendimentos amostrais médios nas situações experimentais consideradas.

IV

1. (a) Seja $\vec{\mathbf{Y}}$ o vector aleatório com as n observações da variável resposta, e $\vec{\boldsymbol{\epsilon}}$ o vector aleatório dos correspondentes erros aleatórios. Seja $\mathbf{X}_{n\times(p+1)}$ a matriz (não aleatória) do modelo, cuja primeira coluna é constituída por n uns, e cujas colunas seguintes contêm as n observações de cada uma das p variáveis preditoras. Seja $\vec{\boldsymbol{\beta}}$ o vector (não aleatório) constituído pelos $p+1$ parâmetros do modelo: $\vec{\boldsymbol{\beta}} = (\beta_0, \beta_1, ..., \beta_p)^t$. O Modelo de Regressão Linear Múltipla admite os seguintes pressupostos:

   - Equação do Modelo: $\vec{\mathbf{Y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}$;
   - Pressupostos sobre os erros aleatórios: $\vec{\boldsymbol{\epsilon}} \sim \mathcal{N}_n(\vec{\mathbf{0}}\,,\, \sigma^2\,\mathbf{I}_n)$.

   A equação do Modelo corresponde à relação linear de fundo entre os preditores e a variável resposta. Os erros aleatórios representam a variabilidade em torno dessa relação linear, admitindo-se a Multinormalidade, independência e variâncias homogéneas no segundo pressuposto do Modelo.

   (b) O vector dos estimadores é dado pela fórmula que consta do formulário: $\hat{\vec{\boldsymbol{\beta}}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}}$. Usando a equação do Modelo, tem-se:

$$\hat{\vec{\boldsymbol{\beta}}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}) = \cancel{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}}\,\vec{\boldsymbol{\beta}} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}} = \vec{\boldsymbol{\beta}} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}\,.$$

   Pelas propriedades da distribuição Multinormal sabe-se que a Multinormalidade dum vector aleatório (como é $\vec{\boldsymbol{\epsilon}}$) não é destruída, nem pela pré-multiplicação por uma matriz não aleatória (como $(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$), nem pela soma dum vector não aleatório (como $\vec{\boldsymbol{\beta}}$). Logo, o vector aleatório $\hat{\vec{\boldsymbol{\beta}}}$ tem distribuição Multinormal. Falta apenas identificar os seus dois parâmetros, que sabemos ser o vector esperado e a matriz de (co-)variâncias respectivos.
   Usando as propriedades operatórias dos vectores esperados e das matrizes de (co-)variâncias, bem como as propriedades de matrizes (estudadas nas aulas), tem-se:

$$E[\hat{\vec{\boldsymbol{\beta}}}] \;=\; E[\vec{\boldsymbol{\beta}} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}] \;=\; \vec{\boldsymbol{\beta}} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\underbrace{E[\vec{\boldsymbol{\epsilon}}]}_{=\vec{\mathbf{0}}} \;=\; \vec{\boldsymbol{\beta}}.$$

   e

$$\begin{aligned} V[\hat{\vec{\boldsymbol{\beta}}}] &= V[\vec{\boldsymbol{\beta}} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}] \;=\; V[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\,\vec{\boldsymbol{\epsilon}}] \;=\; (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\,V[\vec{\boldsymbol{\epsilon}}]\,[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t]^t \\ &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\,\sigma^2\mathbf{I}_n\,[\mathbf{X}^t]^t[(\mathbf{X}^t\mathbf{X})^{-1}]^t \;=\; \sigma^2\,(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}[(\mathbf{X}^t\mathbf{X})^t]^{-1} \\ &= \sigma^2\,\cancel{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}}\,[\mathbf{X}^t(\mathbf{X}^t)^t]^{-1} \;=\; \sigma^2\,(\mathbf{X}^t\mathbf{X})^{-1}\,. \end{aligned}$$

   Logo, tem-se a distribuição indicada no enunciado.

   (c) i. A partir da expressão para R^2_{mod} dada no formulário, e das definições de $QMRE$, QMT e R^2, tem-se:

$$\begin{aligned} R^2_{mod} &= 1 - \frac{QMRE}{QMT} = 1 - \frac{SQRE/[n-(p+1)]}{SQT/(n-1)} \;=\; 1 - \frac{n-1}{n-(p+1)}\,\frac{SQRE}{SQT} \\ &= 1 - \frac{n-1}{n-(p+1)}\,\frac{SQT - SQR}{SQT} \;=\; 1 - \frac{n-1}{n-(p+1)}\,(1-R^2)\,. \end{aligned}$$

   ii. A expressão para R^2_{mod} do ponto anterior significa que R^2_{mod} é uma função crescente em R^2, ou seja, a maiores valores de R^2, maiores valores de R^2_{mod}. Logo, basta substituir na expressão anterior o maior (1) e menor (0) valores possíveis de R^2 para se ter a

gama de possíveis valores de R^2_{mod}. É imediato que, quando $R^2=1$, também $R^2_{mod}=1$. Quando $R^2=0$, tem-se:

$$R^2_{mod} = 1 - \frac{n-1}{n-(p+1)} = \frac{[n-(p+1)]-(n-1)}{n-(p+1)} = \frac{-p}{n-(p+1)} \ ,$$

como se pedia para provar.

iii. Trata-se apenas de interpretar o significado de $R^2_{mod}=1-\frac{QMRE}{QMT}$ quando este indicador toma valores negativos. Nesse caso, tem-se $QMRE > QMT$. Ora, em qualquer Modelo Linear $QMRE$ é o estimador de σ^2, ou seja, da variância da variável resposta Y em torno da hipersuperfície de regressão (que é o significado de σ^2). Por outro lado, $QMT=\frac{SQT}{n-1}=\frac{\cancel{(n-1)}\,s^2_y}{\cancel{n-1}}=s^2_y$, que é a variância amostral das observações de Y, ou seja, é o estimador da variância de Y, na ausência da relação linear com os preditores.

2. A equação do modelo M_{A+B} é $Y_{ijk}=\mu_{11}+\alpha_i+\beta_j+\epsilon_{ijk}$. A equação do modelo M_{A*B} tem ainda os parâmetros de interacção: $Y_{ijk}=\mu_{11}+\alpha_i+\beta_j+(\alpha\beta)_{ij}+\epsilon_{ijk}$.

   (a) Não se deve confundir os *tipos* de parâmetros de cada modelo (μ, os efeitos do Factor A, do Factor B e, eventualmente, de interacção) com o *número* desses parâmetros. É o *número* de parâmetros do modelo que define os graus de liberdade associados à Soma de Quadrados Residual, que são dados pelo número de observações, n, menos esse número total de parâmetros do modelo. Em ambos os modelos considerados existe um único parâmetro μ_{11}. Em relação aos parâmetros α_i, associados aos a níveis do factor A, haveria à partida a parâmetros, mas após a introdução da restrição $\alpha_1=0$ (comum a ambos os modelos) apenas sobram $a-1$ parâmetros desse tipo. De forma análoga, haveria (em ambos os modelos), à partida, b parâmetros β_j, um para cada nível do factor B, mas com a restrição $\beta_1=0$ (comum a ambos os modelos) sobram $b-1$. No modelo M_{A+B} não há mais parâmetros, tendo-se nesse modelo um total de $1+(a-1)+(b-1)=a+b-1$ parâmetros. No modelo M_{A*B} existem ainda os parâmetros $(\alpha\beta)_{ij}$, associados aos efeitos de interacção. As restrições $(\alpha\beta)_{ij}=0$ caso $i=1$ e/ou $j=1$ significam que haverá ao todo $(a-1)(b-1)$ parâmetros desse tipo. Logo, no modelo M_{A*B} o número total de parâmetros é dado por $a+(b-1)+(a-1)(b-1) = a+[\cancel{1}+(a-\cancel{1})](b-1) = a+a(b-1) = a[\cancel{1}+(b-\cancel{1})] = ab$ parâmetros.

   (b) A forma mais simples de verificar que não é possível estudar o modelo com efeitos de interacção caso não existam repetições nas ab células será o de constatar que com apenas $n_c=1$ observação em cada uma dessas situações experimentais, o número total de observações (n) será igual ao número total de parâmetros do modelo (ab). Logo, haverá $n-ab=0$ graus de liberdade associados à Soma de Quadrados Residual, pelo que nem será possível definir um Quadrado Médio Residual. Esta impossibilidade exprime o facto de não existir informação suficiente para ajustar o modelo com efeitos de interacção. Nesta situação de ausência de repetições nas situações experimentais dum delineamento factorial, a única possibilidade de estudar os dados passa por ajustar o modelo sem efeitos de interacção, ou seja, o modelo M_{A+B}.

   (c) A matriz do modelo M_{A+B}, ou seja, a matriz $\mathbf{X}_{A+B}$, é constituída por uma coluna de uns e por colunas indicatrizes de pertença a cada nível do factor A, excepto o primeiro, bem como colunas indicatrizes de pertença a cada nível do factor B, excepto o primeiro. A matriz do modelo M_{A*B}, $\mathbf{X}_{A*B}$, tem essas mesmas colunas e ainda as colunas indicatrizes de pertença a cada célula resultante do cruzamento de cada nível (excepto $i=1$) do factor A com cada nível (excepto $j=1$) do factor B.

Por definição, o espaço das colunas duma matriz é o conjunto de todas as possíveis combinações lineares das colunas dessa matriz. Ora todas as colunas da matriz $\mathbf{X}_{A+B}$ são também colunas da matriz $\mathbf{X}_{A*B}$, pelo que o espaço das colunas $\mathcal{C}(\mathbf{X}_{A+B})$ tem de estar contido no espaço das colunas $\mathcal{C}(\mathbf{X}_{A*B})$. No entanto, algumas combinações lineares das colunas de $\mathbf{X}_{A*B}$ (nomeadamente as que envolvam as indicatrizes de células) não podem ser criadas por combinações lineares das colunas de $\mathbf{X}_{A+B}$, pelo que o espaço das colunas de $\mathbf{X}_{A*B}$ é maior que o espaço das colunas de $\mathbf{X}_{A+B}$.

Como se viu nas aulas teóricas, a Soma de Quadrados Residual é a distância ao quadrado entre o vector das observações da variável resposta, $\vec{\mathbf{y}}$, e a sua projecção ortogonal sobre o espaço das colunas da matriz do modelo. Esse vector projectado é o vector do subespaço que está mais próximo de $\vec{\mathbf{y}}$. Assim, a projecção ortogonal de $\vec{\mathbf{y}}$ sobre $\mathcal{C}(\mathbf{X}_{A+B}) \subseteq \mathcal{C}(\mathbf{X}_{A*B})$ é o vector de $\mathcal{C}(\mathbf{X}_{A+B})$ que está mais próximo de $\vec{\mathbf{y}}$. Trata-se de um vector que também pertence a $\mathcal{C}(\mathbf{X}_{A*B})$. Logo, a menor distância entre o vector $\vec{\mathbf{y}}$ e um vector de $\mathcal{C}(\mathbf{X}_{A*B})$ nunca poderá ser maior que $SQRE_{A+B}$. Poderá ser igual, no caso de as duas projecções coincidirem (o que apenas acontecerá em situações excepcionais), ou poderá ser menor quando (como acontece em geral), a projecção de $\vec{\mathbf{y}}$ sobre $\mathcal{C}(\mathbf{X}_{A*B})$ produzir um vector diferente do obtido na primeira projecção.