INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

9 de Janeiro, 2017 **PRIMEIRO EXAME 2016-17** **Uma resolução possível**

I

A tabela tem contagens de borboletas (de segunda espécie) ingeridas pelos pássaros, em três diferentes categorias de aparência externa das borboletas (três níveis dum factor). As probabilidades associadas a cada classe, caso os pássaros não distingam as diferentes colorações das borboletas, são igualmente indicadas no enunciado. Essas probabilidades constituem a hipótese nula do teste, não sendo necessária qualquer estimação aquando da formulação de H_0. Assim,

1. **Hipóteses:** Represente-se por π_i a probabilidade do resultado $i=1,2,3$, ou seja, a probabilidade de uma borboleta comida pelos pássaros ter a aparência correspondente à categoria i. A hipótese nula é a hipótese de que as probabilidades π_i correspondem às frequências relativas da respectiva aparência, entre as borboletas da segunda espécie. Ou seja,

   Hipótese Nula (H_0): $\pi_1 = 0.33$, $\pi_2 = 0.54$, $\pi_3 = 0.13$.

   Hipótese Alternativa (H_1): pelo menos uma das igualdades em H_0 não se verifica.

   Estatística do Teste: É a estatística de Pearson, na forma de contagens unidimensionais: $X^2 = \sum_{i=1}^{k} \frac{(O_i-E_i)^2}{E_i}$, sendo $k=3$. A distribuição assintótica desta estatística, caso seja verdade H_0, é χ^2_{k-1}, uma vez que não foi necessário estimar qualquer parâmetro.

   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 $\chi^2_{calc} > \chi^2_{\alpha(k-1)} = \chi^2_{0.05(2)} = 5.991$.

2. Tem-se $O_1=7$, $O_2=30$ e $O_3=8$; e $E_1 = N\times\pi_1 = 45\times0.33 = 14.85$, $E_2 = N\times\pi_2 = 45\times0.54 = 24.30$, $E_3 = N\times\pi_3 = 45\times0.13 = 5.85$. Logo:

$$\begin{aligned} X^2_{calc} &= \frac{(7-14.85)^2}{14.85} + \frac{(30-24.30)^2}{24.30} + \frac{(8-5.85)^2}{5.85} \\ &= 4.149663 + 1.337037 + 0.7901709 \quad = \quad 6.276871 \end{aligned}$$

   É legítimo admitir a validade da distribuição assintótica da estatística do teste, uma vez que o menor dos valores esperados é 5.85, acima do limiar de 5 (e por maioria de razão do limiar 1) referido nas condições de Cochran.

   Como $\chi^2_{calc} = 6.276871 > 5.991$, rejeita-se a hipótese nula, ao nível de significância $\alpha = 0.05$ e conclui-se pela rejeição da hipótese de as borboletas da segunda espécie serem comidas na mesma proporção com que se encontram na população.

3. A ser verdade a hipótese da investigadora, o número de borboletas comidas pelos pássaros deveria ser menor que o esperado ao abrigo da hipótese nula, para as borboletas de segunda espécie mais parecidas com a primeira espécie, ou seja, na primeira categoria. Ora, já se viu que, $E_1 = 14.85$, que é bastante superior ao que se observou ($O_1 = 7$). Assim, a hipótese da investigadora parece plausível. No entanto, convém ter cuidado na formulação desta conclusão: como se referiu nas aulas, um teste estatístico não permite concluir que a mimetização é um facto, mas apenas que essa hipótese é coerente com os dados observados.

II

1. Considere-se a regressão linear múltipla com $p=5$ preditores.

   (a) Sabemos que a variância σ^2 é estimada, num modelo linear, pelo Quadrado Médio Residual. O valor de $\sqrt{QMRE}$ (que costuma estar disponível nas listagens produzidas pelo `R`, sob a designação `Residual standard error`) não se encontra disponível. No entanto, conhecemos o valor de $R^2 = \frac{SQR}{SQT} = 0.9925$, bem como o valor de $SQT = (n-1)\, s_y^2 = 35 \times 55.9640^2 = 109\,618.92536$. Logo, $SQRE = SQT - SQR = SQT\,(1-R^2) = 109\,618.92536 \times (1-0.9925) = 822.1419402$. Finalmente, $QMRE = \frac{SQRE}{n-(p+1)} = \frac{822.1419402}{36-6} = 27.4047$.

   (b) Sabemos que as duas variantes do algoritmo de exclusão sequencial (baseado no AIC e baseado nos testes-t bilaterais às hipóteses $H_0 : \beta_j = 0$), se excluirem um preditor num mesmo passo do algoritmo, excluem o mesmo preditor. Ora, pela inspecção dos *p-values* na tabela apresentada no enunciado, conclui-se rapidamente que (para os habituais níveis de significância α) há um único preditor candidato a sair no primeiro passo do algoritmo de exclusão sequencial, o preditor `AsA` (para o qual o valor de prova no teste referido é $p = 0.31399$). Pode calcular-se o QMRE do submodelo sem este preditor, uma vez que se conhece o valor do AIC correspondente ao submodelo. Assim, e com base na fórmula do AIC (disponível no formulário), tem-se, para o submodelo com $k = p-1$ preditores:

$$\begin{aligned}
AIC_{p-1} &= n \ln\left(\frac{SQRE_{p-1}}{n}\right) + 2 \times p \\
123.64 &= 36 \times \ln\left(\frac{SQRE_{p-1}}{36}\right) + 10 \\
\mathrm{e}^{113.64/36} &= \frac{SQRE_{p-1}}{36} = QMRE_{p-1} \times \frac{31}{36} \\
QMRE_{p-1} &= 23.49216 \times \frac{36}{31} = 27.28122\ .
\end{aligned}$$

   Trata-se de um valor muito próximo do obtido no $QMRE$ do modelo completo (de 5 preditores), o que reflecte o grau bastante semelhante de ajustamento dos dois modelos.

   (c) 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 inusual, embora a observação 24 tem um valor já bastante grande, em módulo, com $R_i \approx -3$. Quanto ao efeito alavanca, que sabemos ter de oscilar entre $\frac{1}{n} = \frac{1}{36} = 0.02778$ e 1, e de ter valor médio igual a $\overline{h} = \frac{p+1}{n} = \frac{6}{36} = 0.16667$, verifica-se que algumas observações têm valor alavanca relativamente elevado, com destaque para a observação 24 com efeito alavanca acima de 0.5. 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 $D_i = R_i^2 \left(\frac{h_{ii}}{1-h_{ii}}\right) \frac{1}{p+1}$, os valores relativamente grandes de $|R_i|$ e h_{ii} obrigam a que também a distância de Cook seja bastante elevada, verificando-se no gráfico que ela é, não apenas superior ao nível de guarda 0.5, mas também superior a 1. Assim, esta observação tem uma grande influência no modelo ajustado. O seu valor elevado de efeito alavanca sugere que se trata duma observação com valores afastados dos valores médios das variáveis preditoras. Os valores da observação 24, indicados no enunciado, mostram que lhe corresponde o *menor*

valor observado da variável `GSSG`, e os *maiores* valores observados das variáveis `GSH` e `AsA`. Sendo uma observação extrema, é-o de forma atípica, uma vez que as correlações entre `GSSG` e as outras duas variáveis referidas são positivas (embora pequenas). Assim, não seria de esperar uma observação onde coincidissem o menor valor de `GSSG` e os maiores valores de `GSH` e `AsA`. A atipicidade da observação 24 contribui para explicar a sua grande influência no ajustamento.

2. É dado um submodelo com apenas $k=2$ preditores. Pede-se um teste F parcial para comparar o submodelo com o modelo completo original, de $p=5$ 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 $F_{calc} > f_{0.05[3,30]} = 2.92$.

   Conclusões: Tem-se $F_{calc} = \frac{30}{3} \frac{0.9925-0.9850}{1-0.9925} = 10$. Logo, rejeita-se H_0, i.e., considera-se que a qualidade de ajustamento do modelo completo é significativamente melhor (ao nível $\alpha=0.05$) do que a do submodelo. Este resultado pode ser surpreendente, dada a pequena perda no coeficiente de determinação (da ordem de 0.0075) associada a um modelo com menos 3 preditores. Neste caso, o valor muito elevado do R^2 inicial está a ter um papel importante nesta rejeição. Repare-se que o denominador da razão (na estatística do teste) onde surgem os coeficientes de determinação é de apenas $1 - 0.9925 = 0.0075$, o que inflaciona o valor da estatística do teste, levando o valor calculado dessa estatística para a região crítica.

3. Agora serão consideradas apenas regressões lineares simples.

   (a) O preditor mais fortemente correlacionado com a variável resposta `Carot` será o preditor a que corresponderá a regressão linear simples com mais elevado coeficiente de determinação, e portanto que melhor explicará a variabilidade observada na variável resposta. No enunciado geral do grupo verifica-se que a variável mais correlacionada com `Carot` é o teor de clorofila *a*, `Chla`, verificando-se $r_{xy} = 0.989$. Esta regressão linear simples explicará uma proporção $R^2 = r_{xy}^2 = 0.989^2 = 0.978121$, cerca de 97.8%, da variabilidade observada no teor de carotenóides.

   (b) A forma mais simples de responder é constatar que, no modelo com dois preditores do ponto anterior, um teste t bilateral à hipótese nula $H_0 : \beta_{GSH} = 0$ tem um valor de prova (*p-value*) correspondente $p = 0.000345$, pelo que, para os níveis de significância usuais, rejeita-se esta hipótese nula, em favor da hipótese alternativa $H_1 : \beta_{GSH} \neq 0$. Assim, o preditor `GSH` não é dispensável no submodelo a dois preditores, ou seja, o modelo de regressão linear simples de `Carot` sobre `Chla` explica significativamente menos do que o modelo com dois preditores. Tal como na passagem do modelo completo original (cinco preditores) para o submodelo a dois preditores, também aqui a diferença é significativa apesar dos valores bastante próximos dos coeficientes de determinação (0.985 e 0.978), o que de novo salienta que para valores muito elevados de R^2 no modelo completo, pequenas diferenças podem ser consideradas significativas.

   (c) Em qualquer modelo de regressão linear, tem-se $SQR = (n-1)\, s_{\hat{y}}^2$. Já vimos que para esta regressão linear simples, $R^2 = \frac{SQR}{SQT} = 0.978121$. Por outro lado, sabemos que $SQT = 109\,618.92536$ (valor calculado na alínea 1a e que permanece igual, já que não depende do submodelo ajustado). Logo, $SQR = 0.978121 \times 109\,618.92536 = 107220.6$ e $s_{\hat{y}}^2 = \frac{SQR}{n-1} = \frac{107220.6}{35} = 3063.445$.

(d) Os intervalos a $(1-\alpha)\times 100\%$ de confiança para β_1, numa regressão linear simples, são da forma

$$]\ b_1 - t_{\alpha/2(n-2)}\,\hat{\sigma}_{\hat{\beta}_1}\ ,\ b_1 + t_{\alpha/2(n-2)}\,\hat{\sigma}_{\hat{\beta}_1}\ [\ .$$

Ora, $b_1 = \frac{cov_{xy}}{s_x^2} = r_{xy}\,\frac{s_y}{s_x} = 0.989\,\frac{55.9640}{115.0790} = 0.48096$. Por outro lado, $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{QMRE}{(n-1)\,s_x^2}} = \frac{\sqrt{\frac{SQT-SQR}{n-2}}}{\sqrt{35}\times 115.0790} = 0.012336$, tendo em conta os valores de SQT e SQR acima calculados. Finalmente, $t_{0.025(34)} \approx 2.03$. O intervalo a 95% de confiança é assim o intervalo $]0.4559, 0.5060[$. Assim, a 95% de confiança, a variação esperada no teor de carotenoides, associada a um aumento de um μmol ml^{-1} está, a 95% de confiança, entre 0.4559 e 0.5060 μmol ml^{-1}.

III

1. Trata-se dum delineamento factorial a dois factores, sendo a variável resposta Y o teor final de zinco no grão; o primeiro factor (A) a quantidade de adubação, com $a = 3$ níveis e o segundo factor (B) o tipo de grão (com $b = 2$ níveis). O delineamento é equilibrado, uma vez que em cada uma das $ab=6$ células (situações experimentais) existem $n_c=8$ observações. Havendo repetições nas células, é possível (e desejável) estudar a existência de eventuais efeitos de interacção.

   O modelo ajustado é o modelo ANOVA a dois factores, com efeitos de interacção:

   - $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, para qualquer $i=1,2,3$, $j=1,2$ e $k=1,2,...,8$, sendo μ_{11} o teor esperado de zinco para o grão polido, sem adubação; α_i o efeito principal (acréscimo ao teor de zinco) associado ao nível de adubação i (com a restrição $\alpha_1 = 0$); β_j o efeito principal (acréscimo ao teor de zinco) associado ao tipo de grão $j = 2$ (dada a restrição $\beta_1 = 0$); $(\alpha\beta)_{ij}$ o efeito de interacção, isto é, o acréscimo ao teor de zinco associado à combinação do nível i de adubação com o tipo de grão j. Dadas as restrições $(\alpha\beta)_{ij}=0$ se $i=1$ e/ou $j=1$), o modelo apenas prevê efeitos de interacção nas situações experimentais $(i,j)=(2,2)$ (adubação de 300 mg/kg de solo, para o grão integral) $(i,j)=(3,2)$ (adubação de 600 mg/kg de solo, para o grão integral). Finalmente ϵ_{ijk} é o erro aleatório da observação Y_{ijk}.
   - Admite-se que os erros aleatórios são todos 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. (a) Sabemos que os graus de liberdade associados a $QMRE$ são dados por $n-ab$, onde n é o número total de observações, $n = n_c ab = 8\times 6 = 48$, e $ab = 6$ é o número total de parâmetros existentes no modelo. Assim, $g.l.(SQRE) = 42$. 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 $a-1=2$ g.l., associado a SQB há $b-1=1$ g.l., e associado a $SQAB$ há $(a-1)(b-1)=2$ graus de liberdade. Os Quadrados Médios são dados pelas Somas de Quadrados a dividir pelos respectivos graus de liberdade, pelo que $QMB = \frac{467.44}{1} = 467.44$. A Soma de Quadrados Residual pode ser obtida a partir de $SQRE = QMRE \times (n-ab) = 2.35\times 42 = 98.70$. Finalmente, os valores das duas estatísticas de teste em falta resultam do facto de ambas serem dadas pelo respectivo Quadrado Médio a dividir pelo Quadrado Médio Residual. Assim, $F^B_{calc} = \frac{QMB}{QMRE} = \frac{467.44}{2.35} = 198.911$ e $F^{AB}_{calc} = \frac{QMAB}{QMRE} = \frac{1.57}{2.35} = 0.668$.
   Nota: Alguns destes valores sofrem erros de arredondamento nos cálculos.
   Logo, a tabela-resumo completa é:

```
                Df  Sum Sq  Mean Sq  F value
Tratamento       2   12.78     6.39   2.715
Tipo             1  467.44   467.44 198.911
Tratamento:Tipo  2    3.14     1.57   0.668
Residuals       42   98.70     2.35
```

(b) Vai-se efectuar em pormenor o teste aos efeitos principais do Factor A (níveis de adubação), e descrever sinteticamente os testes aos efeitos principais do Factor B (tipos de grão) e aos efeitos de interacção.

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

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

Nível de significância: $\alpha = 0.10$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.10(2,42)} \approx 2.44$ (entre os valores tabelados 2.39 e 2.44).

Conclusões: Como $F_{calc} = \frac{QMA}{QMRE} = 2.715 > 2.44$, rejeita-se (por pouco) H_0, sendo possível concluir pela existência de efeitos principais de níveis de adubação (ao nível $\alpha = 0.10$).

No teste aos efeitos de interacção, com hipóteses $H_0 : (\alpha\beta)_{ij} = 0$, para todo o i e j, contra H_1 : existe pelo menos uma célula (i, j) onde $(\alpha\beta)_{ij} \neq 0$, o valor calculado da estatística de teste é muito baixo ($F_{calc} = 0.668$), e inferior ao limiar da região crítica, que é (por coincidência) igual ao do teste anterior $f_{0.10(2,42)} \approx 2.44$. Logo, não se rejeita H_0 (para $\alpha = 0.10$), pelo que não há efeitos significativos de interacção.

Finalmente, no teste aos efeitos principais do factor tipo de grão, as hipóteses do teste podem ser escritas apenas como $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$, uma vez que após a imposição da restrição $\beta_1 = 0$, apenas sobra um efeito deste tipo, o efeito β_2 associado à passagem do grão polido para o grão integral. O valor calculado da estatística de teste é enorme ($F_{calc} = 198.911$) deixando antever a rejeição de H_0, facto que é confirmado determinando a partir das tabelas o limiar da região crítica unilateral direita: $f_{0.10(1,42)} \approx 2.84$ (entre os valores tabelados 2.84 e 2.79). Assim, conclui-se claramente pela existência de efeitos principais de tipo de grão nos teores médios de zinco no grão.

(c) Pretende-se comparar (com $\alpha = 0.05$) os pares de médias que se podem formar a partir das seis 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 o termo de comparação $q_{\alpha(ab,n-ab)} \sqrt{\frac{QMRE}{n_c}}$ (disponível no formulário), ou seja, se $|\overline{y}_{ij.} - \overline{y}_{i'j'.}| > q_{0.05(6,42)} \sqrt{\frac{2.35}{8}}$. O valor da distribuição de Tukey pode ser obtido nas tabelas desta distribuição e é aproximadamente 4.22 (entre 4.19 e 4.23). Logo, o termo de comparação é (aproximadamente) $q_{0.05(6,42)} \sqrt{\frac{2.35}{8}} = 4.22 \times 0.5419871 = 2.2872$. Olhando para as médias das seis células disponíveis no enunciado, imediatamente se verifica que as seis médias de célula, quando lidas por linha, estão por ordem crescente. Nenhum par de médias correspondentes ao grão polido (primeira linha na tabela do enunciado) difere por mais de duas unidades, tal como nenhum par de médias correspondentes ao grão integral (segunda linha) difere em mais de duas unidades. É igualmente imediato verificar que a diferença entre a maior das médias da primeira linha e a menor das médias da segunda linha (a diferença $|9.934 - 5.131| = 4.803$) difere em mais do que o termo de comparação 2.2872. Logo, a conclusão do teste de Tukey (para $\alpha = 0.05$) é que qualquer das médias correspondentes a grão polido é significativamente diferente de qualquer das médias correspondente a grão integral, não havendo no entanto nenhuma diferença significativa

entre as médias de diferentes níveis de adubação, para um mesmo tipo de grão. Neste sentido, a conclusão do teste de Tukey, para $\alpha = 0.05$, não é inteiramente coerente com as conclusões dos testes F da alínea anterior. Coerente nos dois testes foi a conclusão de que os efeitos de tipo de grão são claramente significativos. Mas o teste F identificou a existência de efeitos de nível de adubação, que agora não são detectados. Uma das razões para essa discrepância é o nível de significância diferente usado nos dois testes: os testes F foram efectuados com $\alpha = 0.10$ e a rejeição de H_0 no teste aos efeitos de adubação foi por pouco. Aliás, registe-se que, caso o teste F aos efeitos de adubação tivesse também sido realizado com nível $\alpha = 0.05$, não teria havido rejeição de H_0 no teste aos efeitos do Factor A (o valor de $f_{0.05(2,42)}$ está entre 3.15 e 3.23). De qualquer forma, nem sempre os resultados dos testes F e de Tukey são inteiramente compatíveis, sobretudo quando as conclusões são marginais.

(d) A tabela-resumo da ANOVA correspondente ao modelo com o único factor, nível de adubação (ou seja, ao Modelo M_A, resultante de ignorar os diferentes tipos de grão), tem apenas 2 linhas: a da variabilidade associada ao factor e a correspondente à variabilidade residual. Por definição, a Soma de Quadrados, grau de liberdade e, por conseguinte, o Quadrado Médio associado ao factor adubação, são calculados como na tabela-resumo do modelo ANOVA a dois factores, com efeitos de interacção (o modelo M_{A*B}), pelo que os correspondentes valores são iguais nas duas tabelas. Uma vez que a soma de todas as Somas de Quadrados em cada tabela-resumo ANOVA tem de ser sempre igual a $SQT = (n-1)s_y^2$, e uma vez que os valores da variável resposta Y_i com que se ajusta os dois modelos são os mesmos, tem de verificar-se $SRQE_A = SQB + SQAB + SQRE_{A*B} = 569.38$. De forma análoga, os graus de liberdade das duas tabelas têm de somar n–1, pelo que $g.l.(SQRE_A) = 1+2+42 = 45$. Tem-se então (já que o número de níveis do factor A é $k = a$) que $QMRE = \frac{SQRE}{n-a} = \frac{569.38}{45} = 12.65289$ e $F_A = \frac{QMF}{QMRE} = \frac{6.39}{12.65289} = 0.505023$. A tabela-resumo vem assim:

	g.l.	SQs	QMs	F_{calc}
Factor adubação	2	12.78	6.39	0.505023
Residual	45	569.38	12.65	

O valor inferior a 1 da estatística do teste deixa antever a não rejeição da hipótese nula do teste, facto que pode ser confirmado confrontando $F_{calc} = 0.505023$ com a fronteira da região crítica, que é agora $f_{0.10(2,45)} \approx 2.44$. Assim, para o mesmo nível de significância $\alpha = 0.10$, a conclusão no teste à existência de efeitos de adubação é agora diferente da obtida no modelo a dois factores, com efeito de interacção. Trata-se de mais uma ilustração da ideia de que não contemplar na experiência e no modelo causas importantes de variabilidade (neste caso, a existência de dois diferentes tipos de grão, aos quais correspondem teores médios de zinco muito diferentes) pode mascarar a existência de reais efeitos (neste caso, efeitos de nível de adubação) que, com uma experiência e modelo mais apropriado, se evidenciariam.

IV

1. (a) O modelo de regressão linear múltipla relaciona uma variável resposta Y com p variáveis preditoras X_1, X_2, ..., X_p. Designando por $\vec{\mathbf{Y}}$ o vector das n observações da variável resposta Y, $\vec{\boldsymbol{\epsilon}}$ o vector dos n erros aleatórios correspondentes, $\vec{\boldsymbol{\beta}}$ o vector dos $p+1$ parâmetros do modelo, $\beta_0, \beta_1, ..., \beta_p$, e $\mathbf{X}$ a matriz $n \times (p+1)$, cuja primeira coluna é constituída por n uns e cada uma das restantes p colunas contém as n observações duma variável preditora,

tem-se:

$$\vec{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & x_{1_{(1)}} & x_{2_{(1)}} & \cdots & x_{p_{(1)}} \\ 1 & x_{1_{(2)}} & x_{2_{(2)}} & \cdots & x_{p_{(2)}} \\ 1 & x_{1_{(3)}} & x_{2_{(3)}} & \cdots & x_{p_{(3)}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1_{(n)}} & x_{2_{(n)}} & \cdots & x_{p_{(n)}} \end{bmatrix}, \ \vec{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \ \vec{\boldsymbol{\epsilon}} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

O modelo de regressão linear múltipla é então dado por:

i. $\vec{\mathbf{Y}} = \mathbf{X}\,\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}$

ii. $\vec{\boldsymbol{\epsilon}} \ \cap \ \mathcal{N}_n(\vec{\mathbf{0}},\ \sigma^2\mathbf{I}_n)$,

sendo $\vec{\mathbf{0}}$ o vector de n zeros e $\mathbf{I}_n$ a matriz identidade $n \times n$. Na segunda condição, indica-se que o vector dos erros aleatórios segue uma distribuição Multinormal, com vector médio dado pelo vector de zeros (ou seja, cada erro aleatório individual tem valor esperado zero) e matriz de variâncias-covariâncias diagonal, com os elementos diagonais todos iguais a σ^2. Uma vez que, numa matriz de (co-)variâncias os elementos diagonais representam as variâncias de cada componente do vector, esta condição indica que $V[\epsilon_i] = \sigma^2$, $\forall\, i$. O facto de os elementos não diagonais da matriz $\sigma^2\,\mathbf{I}_n$ serem todos nulos equivale a dizer que a covariância entre elementos diferentes do vector aleatório dos erros é sempre nula (ou seja, $Cov[\epsilon_i, \epsilon_j] = 0$, sempre que $i \neq j$) e, como sabemos, numa distribuição Multinormal tal facto implica a independência desses elementos.

(b) O vector $\vec{\hat{\boldsymbol{\beta}}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)^t$ dos estimadores dos $p+1$ parâmetros dum modelo linear é dado (ver formulário) por $\vec{\hat{\boldsymbol{\beta}}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}}$. Mas, pelo modelo, tem-se $\vec{\mathbf{Y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}$. Substituindo, tem-se:

$$\vec{\hat{\boldsymbol{\beta}}} \ = \ (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}) \ = \ \underbrace{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}}_{=\mathbf{I}}\vec{\boldsymbol{\beta}} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}} \ = \ \vec{\boldsymbol{\beta}} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}\,,$$

como se pedia para mostrar.

(c) A expressão da alínea anterior é a soma dum vector não aleatório, $\vec{\boldsymbol{\beta}}$, com um vector aleatório, $(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}$. Ora, para qualquer vector aleatório $\vec{\mathbf{W}}$ e vector não aleatório $\vec{\mathbf{a}}$ verifica-se $E[\vec{\mathbf{W}} + \vec{\mathbf{a}}] = E[\vec{\mathbf{W}}] + \vec{\mathbf{a}}$. Logo, no nosso caso, tem-se: $E[\vec{\hat{\boldsymbol{\beta}}}] = E[\vec{\boldsymbol{\beta}} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}] = \vec{\boldsymbol{\beta}} + E[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}]$. A segunda parcela é o vector esperado dum vector que resulta de multiplicar uma matriz não aleatória ($(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$) por um vector aleatório ($\vec{\boldsymbol{\epsilon}}$). Por outra propriedade operatória dos vectores esperados, tem-se $E[\mathbf{B}\vec{\mathbf{W}}] = \mathbf{B}\,E[\vec{\mathbf{W}}]$, onde $\mathbf{B}$ é uma matriz não aleatória. Assim, $E[\vec{\hat{\boldsymbol{\beta}}}] = \vec{\boldsymbol{\beta}} + E[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}] = \vec{\boldsymbol{\beta}} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\underbrace{E[\vec{\boldsymbol{\epsilon}}]}_{=\vec{\mathbf{0}}} = \vec{\boldsymbol{\beta}} + \vec{\mathbf{0}} = \vec{\boldsymbol{\beta}}$.

Por outro lado, tendo em conta a propriedade operatória geral de matrizes de (co-)variâncias, $V[\vec{\mathbf{W}} + \vec{\mathbf{a}}] = V[\vec{\mathbf{W}}]$, tem-se $V[\vec{\hat{\boldsymbol{\beta}}}] = V[\vec{\boldsymbol{\beta}} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}] = V[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}]$. Outra propriedade operatória de matrizes de (co-)variâncias diz-nos que $V[\mathbf{B}\vec{\mathbf{W}}] = \mathbf{B}V[\vec{\mathbf{W}}]\mathbf{B}^t$, para uma matriz não aleatória $\mathbf{B}$. Logo (e sendo no nosso caso $\mathbf{B} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$), tem-se:

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

Nota: Repare-se que, não tendo sido esta a forma como o vector esperado e a matriz de (co-)variâncias de $\vec{\hat{\boldsymbol{\beta}}}$ foram calculados nas aulas, as propriedades usadas para chegar ao resultado final são as mesmas.

2. (a) Usando como ponto de partida a expressão de R^2_{mod} contante do formulário, bem como as definições de $QMRE$ e QMT, tem-se:

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

já que o coeficiente de determinação usual verifica $R^2 = \frac{SQR}{SQT} = \frac{SQT-SQRE}{SQT} = 1 - \frac{SQRE}{SQT}$, pelo que $\frac{SQRE}{SQT} = 1 - R^2$.

(b) Tendo em conta a expressão da alínea anterior, tem-se:

$$\begin{aligned} R^2 - R^2_{mod} &= R^2 - \left[1 - \frac{n-1}{n-(p+1)}(1-R^2)\right] = (R^2-1) + (1-R^2)\frac{n-1}{n-(p+1)} \\ &= (1-R^2)\left[-1 + \frac{n-1}{n-(p+1)}\right] = (1-R^2)\left[\frac{-[n-(p+1)]+n-1}{n-(p+1)}\right] \\ &= (1-R^2)\left[\frac{\cancel{-n}+p+\cancel{1}+\cancel{n}-\cancel{1}}{n-(p+1)}\right] = (1-R^2)\frac{p}{n-(p+1)} \end{aligned}$$

Logo, tem-se o resultado pedido:

$$\frac{R^2}{R^2 - R^2_{mod}} = \frac{R^2}{(1-R^2)\frac{p}{n-(p+1)}} = \frac{n-(p+1)}{p}\frac{R^2}{1-R^2} = F\ .$$

(c) A partir da expressão para R^2_{mod} obtida na alínea 2a, tem-se:

$$\begin{aligned} R^2_{mod} < 0 &\Leftrightarrow 1 < (1-R^2)\frac{n-1}{[n-(p+1)]} \\ &\Leftrightarrow \frac{n-(p+1)}{n-1} < 1-R^2 \\ &\Leftrightarrow R^2 < 1 - \frac{n-(p+1)}{n-1} = \frac{p}{n-1} \end{aligned}$$

Quando o valor de R^2 é inferior a $\frac{p}{n-1}$, tem-se $R^2_{mod} < 0$ e, pela alínea anterior, o valor calculado da estatística do teste F de ajustamento global será inferior a 1. Uma rápida consulta às tabelas da distribuição F confirma que, nesse caso, nunca se rejeita a hipótese nula dum teste de ajustamento global, ou seja, o modelo de regressão linear múltipla será indistinguível do Modelo Nulo e, como tal, desinteressante.