INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2021-22

19 Janeiro 2022 Primeira Chamada de EXAME Uma resolução possível

I

Trata-se duma regressão linear simples com logaritmização de ambas as variáveis.

1. O declive da recta é dado no enunciado: $b_1 = 0.219779$. Para calcular a ordenada na origem basta recorrer à fórmula, sem esquecer que as variáveis foram logaritmizadas: $b_0 = \overline{y^*} - b_1\,\overline{x^*} = 2.785572 - 0.219779 \times 2.771838 = 2.17638$. Logo, a equação da recta ajusta é $y^* = 2.17638 + 0.219779\,x^*$. Para calcular o respectivo coeficiente de determinação, começamos por lembrar que este é dado, numa regressão linear simples, pelo quadrado do coeficiente de correlação linear entre as variáveis. Por outro lado, o declive é dado por $b_1 = \frac{cov_{x^*y^*}}{s^2_{x^*}} = r_{x^*y^*} \cdot \frac{s_{y^*}}{s_{x^*}}$. Logo, $r_{x^*y^*} = b_1 \cdot \frac{s_{x^*}}{s_{y^*}} = 0.219779 \times \frac{\sqrt{13.2079}}{\sqrt{0.7682822}} = 0.9112608$. Assim, $R^2 = (r_{x^*y^*})^2 = 0.9112608^2 = 0.8303962$. A recta explica cerca de 83% da variabilidade observada dos log-comprimentos axiais dos olhos, tratando-se dum valor razoavelmente bom.

2. Vamos testar se o modelo difere significativamente do Modelo Nulo (sem preditores), sendo de prever que assim seja, dado o valor considerável de R^2. Tem-se:

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

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

   Nível de significância: $\alpha = 0.05$.

   Região Crítica: Unilateral direita. Rejeitar H_0 se $F_{calc} > f_{0.05[1,170]}$. Trata-se dum valor entre os valores tabelados mais próximos, $f_{0.05[1,120]} = 3.92$ e $f_{0.05[1,\infty]} = 3.84$.

   Conclusões: O valor calculado da estatística é dado por $F_{calc} = (n-2) \times \frac{R^2}{1-R^2} = 170 \times \frac{0.8303962}{1-0.8303962} = 832.336$. Assim, tem-se uma claríssima rejeição de H_0 (ao nível de significância usado mas, dado o valor enorme de F_{calc}, também para qualquer dos usuais níveis de significância), pelo que o modelo difere claramente do Modelo Nulo. Esta conclusão era expectável, dado o valor de R^2. O gráfico de log-comprimento axial *vs.* log-peso revela o bom ajustamento geral, embora com uma curvatura ligeira. Assim, é possível que uma transformação alternativa das variáveis pudesse linearizar melhor a relação.

3. Em qualquer modelo linear, a estimativa da variância σ^2 dos erros aleatórios é dada por $\hat{\sigma}^2 = QMRE$. Para calcular o Quadrado Médio Residual, registamos que por definição, $R^2 = \frac{SQR}{SQT} = 0.8303962$. Ora, $SQT = (n-1) \cdot s^2_{y^*} = 171 \times 0.7682822 = 131.3763$. Logo, $SQR = SQT \cdot R^2 = 131.3763 \times 0.8303962 = 109.0943$ e $SQRE = SQT - SQR = 131.3763 - 109.0943 = 22.282$. Finalmente, $\hat{\sigma}^2 = QMRE = \frac{SQRE}{n-2} = \frac{22.282}{170} = 0.1310706$.

4. A relação não linear pedida é uma relação potência. De facto,

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

   Tendo em conta os parâmetros b_0 e b_1 da recta ajustada, a relação potência para estes dados será $y = \mathrm{e}^{2.17638}\,x^{0.219779} = 8.814341\,x^{0.219779}$. Por outras palavras, o comprimento axial (y) é proporcional à potência 0.219779 do peso do corpo (x), ou seja, aproximadamente proporcional à raíz quinta de x.

5. O intervalo a $(1-\alpha)\times 100\%$ de confiança para β_1 é da forma:

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

Sabemos que $b_1 = 0.219779$. Pelas tabelas da distribuição *t-Student* podemos considerar que $t_{0.025(170)} \approx 1.97$. Pelo formulário sabemos que $V[\hat{\beta}_1] = \frac{\sigma^2}{(n-1)\,s_x^2}$, pelo que (e tendo em conta que o nosso preditor foi logaritmizado) é estimada por $\hat{\sigma}^2_{\hat{\beta}_1} = \widehat{V[\hat{\beta}_1]} = \frac{QMRE}{(n-1)\,s^2_{x^*}}$. Logo, $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{0.1310706}{171\times 13.2079}} = 0.007617943$. Substituindo estes valores na fórmula do IC obtém-se o seguinte intervalo a 95% de confiança para β_1: $]\ 0.2047\ ,\ 0.2348\ [$. A afirmação do enunciado é que a quinta potência da variável y é proporcional a x ($y^5 \propto x$), ou seja, $y^5 = c\,x$, para alguma constante de proporcionalidade c. Esta afirmação é equivalente a dizer que $y = c^{\frac{1}{5}}\,x^{\frac{1}{5}} = c^*\,x^{0.2}$, ou seja que $y \propto x^{0.2}$. Assim, e na sequência da alínea anterior, o enunciado pede para verificar se o parâmetro populacional β_1 pode tomar o valor $\beta_1 = 0.2$. Este valor não está contido no intervalo a 95% de confiança para β_1, pelo que a afirmação do enunciado deve ser rejeitada.

II

Temos uma regressão linear múltipla de `brix` sobre $p=7$ preditores, ajustada com $n=219$ observações.

1. Pede-se um teste para comparar dois parâmetros populacionais, concretamente, pergunta-se se é admissível afirmar que $\beta_5 = \beta_6$, que equivale a $\beta_5 - \beta_6 = 0$. Vamos efectuar um teste de hipóteses para saber se é admissível que esta combinação linear de parâmetros seja nula.

   Hipóteses: $H_0 : \beta_5 - \beta_6 = 0$ vs. $H_1 : \beta_5 - \beta_6 \neq 0$.

   Estatística do Teste: $T = \frac{(\hat{\beta}_5-\hat{\beta}_6)-(\beta_5-\beta_6)_{|H_0}}{\hat{\sigma}_{\hat{\beta}_5-\hat{\beta}_6}} \frown t_{n-(p+1)}$, sob H_0.

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

   Região Crítica: (Bilateral) Rejeitar H_0 se $|T_{calc}| > t_{\frac{\alpha}{2}[n-(p+1)]} = t_{0.025(211)} \approx 1.97$, estando o valor exacto entre os valores tabelados $t_{0.025(120)} = 1.97993$ e $t_{0.025(\infty)} = 1.96234$.

   Conclusões: Para calcular o valor da estatística de teste, precisaremos do valor do erro padrão

$$\begin{aligned}
\hat{\sigma}_{\hat{\beta}_5-\hat{\beta}_6} &= \sqrt{\widehat{V[\hat{\beta}_5-\hat{\beta}_6]}} = \sqrt{\hat{V}[\hat{\beta}_5] + \hat{V}[\hat{\beta}_6] - 2\,\widehat{Cov}[\hat{\beta}_5,\hat{\beta}_6]} \\
&= \sqrt{1.729520\times 10^{-6} + 0.8459073\times 10^{-6} - 2*(-1.043154)\times 10^{-6}} \\
&= \sqrt{4.661735\times 10^{-6}} = 0.002159105.
\end{aligned}$$

   Logo, $T_{calc} = \frac{(0.0061633-0.0009412)-0}{0.002159105} = 2.418641$. Assim, rejeita-se H_0, concluindo-se que $\beta_5 \neq \beta_6$ (ao nível $\alpha = 0.05$).

2. Estamos perante um gráfico de resíduos estandardizados (eixo vertical) contra valores do efeitos alavanca (eixo horizontal). As linhas a tracejado que surgem junto aos cantos direitos do gráfico correspondem a isolinhas de distâncias de Cook, de valor 0.5 e 1 (esta última apenas visível no canto superior). Ao abrigo do modelo, a quase totalidade das observações deveria ter resíduos estandardizados $R_i = \frac{E_i}{\sqrt{QMRE\cdot(1-h_{ii})}}$ contidos no intervalo $[-3,3]$. Como se pode constatar, existe uma única observação com valor de R_i fora deste intervalo, concretamente $R_{172} \approx 5$, correspondente à observação que surge no topo do gráfico. Trata-se dum valor muito elevado, indicando uma observação que se encontra anormalmente afastada da hipersuperfície ajustada.

Conviria analisar melhor esta observação a fim de identificar possíveis causas para um valor tão atípico. Por outro lado, os efeitos alavanca h_{ii} medem o grau de 'atracção' da observação i sobre a hipersuperfície ajustada, sendo tanto maior quanto mais próxima do valor máximo 1 estiver h_{ii}. Também neste caso, existe uma observação com um valor muito superior aos restantes, observação não legendada, à direita no gráfico, com $h_{ii} \approx 0.35$. Uma vez que o valor médio dos efeitos alavanca é $\overline{h} = \frac{p+1}{n} = \frac{8}{219} = 0.03652968$, trata-se dum valor cerca de 10 vezes maior que a média, o que é assinalável. Também neste caso, conviria conhecer melhor os valores desta observação. As restantes observações têm efeitos alavanca inferiores a metade deste valor máximo registado, sendo no entanto de referir que a observação atípica 172 tem um efeito alavanca cujo valor está entre os quatro maiores. Assim, não surpreende que a observação 172 tenha uma distância de Cook muito elevada, próxima do limiar de alerta: $D_{172} \approx 0.5$. De facto, sabemos que a distância de Cook, que mede a influência duma observação (ou seja, o impacto que a sua exclusão do conjunto de observações teria sobre a hipersuperficie ajustada), é dada por $D_i = R_i^2 \cdot \frac{h_{ii}}{1-h_{ii}} \cdot \frac{1}{p+1}$. Assim, D_i cresce com os valores dos resíduos estandardizados R_i e também dos efeitos alavanca h_{ii}. Como a observação 172 tem valores relativamente elevados em ambas estas quantidades, a sua distância de Cook resulta ser muito elevada. Já a observação mais à direita no gráfico, apesar de ter um efeito alavanca grande, tem um resíduo estandardizado próximo de zero, acabando por ter uma distância de Cook baixa (o ponto está afastado das isolinhas). Estas constatações reforçam a importância de se analisar mais de perto a observação 172 (algo que não podemos fazer por não se dispôr no enunciado dos seus valores observados).

3. A primeira parte da afirmação poderá, ou não, ser verdadeira, mas nada permite retirar essa conclusão a partir da informação disponível. É verdade que `pH` e `antocianas` são os dois preditores individualmente mais fortemente correlacionados com a variável resposta `brix`, mas esse facto por si só não implica que formem o melhor par de preditores. O que é seguramente possível afirmar é que a equação do plano de regressão que resultaria da regressão linear de `brix` sobre `pH` e `antocianas` *não* será a indicada no enunciado, uma vez que a equação mostrada resulta de reter, na equação da regressão linear múltipla com a totalidade dos $p=7$ preditores, as parcelas correspondentes a esses dois preditores, bem como a constante aditiva. Sabemos que não é assim que se obtêm as equações ajustadas por submodelos.

4. É pedido um teste F parcial comparando o modelo completo com $p=7$ preditores e o subdmodelo com $k=5$ preditores, resultante da exclusão dos preditores `acidez` e `fenois`. Tem-se:

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

   Estatística do teste $F = \frac{n-(p+1)}{p-k} \cdot \frac{R_c^2 - R_s^2}{1-R_c^2} \frown F_{[p-k,n-(p+1)]}$, se H_0 verdade.

   Nível de significância: $\alpha = 0.05$

   Região Crítica: Unilateral direita. Rejeita-se H_0 se
   $F_{calc} > f_{\alpha[p-k,n-(p+1)]} = f_{0.05(2,211)} \approx 3.00$.

   Conclusões: O enunciado informa que $R_s^2 = 0.6952$. Logo, tem-se $F_{calc} = \frac{211}{2} \cdot \frac{0.7026-0.6952}{1-0.7026} = 2.625084$. Assim, não se rejeita H_0. Não se pode concluir que os dois modelos tenham qualidade de ajustamento significativamente diferente. Esta conclusão é expectável, dado o valor bastante próximo de ambos os coeficientes de determinação.

III

1. Trata-se duma ANOVA em que a variável resposta Y é o rendimento das parcelas, havendo um único factor para explicar diferenças nos rendimentos: o factor genótipos, com $k=7$ níveis.

Para cada um desses níveis (genótipos) existem $n_i = 5$ observações, pelo que se trata dum delineamento equilibrado, com $n_c = 5$ repetições em cada nível do factor. No total existem $n = k \times n_c = 7 \times 5 = 35$ observações. Eis o modelo ANOVA a um factor, com casualização total:

Equação do Modelo: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, onde $i = 1, ..., 7$ indica genótipo e $j = 1, ..., 5$ indica repetição (dentro de cada genótipo i); Y_{ij} é o rendimento da j-ésima repetição do genótipo i; ϵ_{ij} é o correspondente erro aleatório; e α_i indica o efeito associado ao genótipo i. Com a restrição $\alpha_1 = 0$, a constante aditiva comum a todas as observações (μ) será o rendimento populacional médio do primeiro genótipo (B020) (que podemos designar $\mu = \mu_1$).

Distribuição dos erros: $\epsilon_{ij} \frown \mathcal{N}(0, \sigma^2)$, para qualquer i, j.

Independência dos erros: $\{\epsilon_{ij}\}_{i,j}$ são variáveis aleatórias independentes.

2. A tabela apenas terá duas linhas, uma associada ao factor genótipo e outra residual. Sabemos que os graus de liberdade associados ao factor são $k - 1 = 6$ e os graus de liberdade residuais são $n - k = 35 - 7 = 28$. Pelo formulário sabemos que $SQRE = \sum_{i=1}^{k} (n_i - 1)\, s_i^2 = (n_c - 1) \sum_{i=1}^{k} s_i^2 = 4 \times (0.15680 + 0.24099 + 0.42720 + 0.24901 + 0.33339 + 1.05899 + 0.50835) = 11.89892$. O Quadrado Médio Residual será $QMRE = \frac{SQRE}{n-k} = \frac{11.89892}{28} = 0.4249614$. Como a Soma de Quadrados Total é $SQT = (n-1)\, s_y^2 = 34 \times 0.57780 = 19.6452$, tem-se $SQF = SQT - SQRE = 19.6452 - 11.89892 = 7.74628$. Logo, $QMF = \frac{SQF}{k-1} = \frac{7.74628}{6} = 1.291047$. Finalmente, a estatística do teste F associado ao único tipo de efeitos previstos no modelo toma valor $F_{calc} = \frac{QMF}{QMRE} = \frac{1.291047}{0.4249614} = 3.038033$. Assim, tem-se a seguinte tabela-resumo:

Fontes de Variação	gl	Somas de Quadrados	Quadrados Médios	F_{calc}
Genótipo (Factor)	k-1=6	SQF=7.74628	QMF=1.291047	F=3.038033
Residual	n-k=28	SQRE=11.89892	QMRE=0.4249614	—
Total	34	19.6452	–	–

3. É pedido o teste aos efeitos do factor, cuja hipótese nula equivale à igualdade de todos os rendimentos médios populacionais de genótipo ($\mu_1 = \mu_2 = ... = \mu_7$). Pode escrever-se:

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

Estatística do teste $F = \frac{QMF}{QMRE} \frown F_{[k-1,n-k]}$, se H_0 verdade.

Nível de significância: $\alpha = 0.05$

Região Crítica: Unilateral direita. Rejeita-se H_0 se $F_{calc} > f_{\alpha[k-1,,n-k]} = f_{0.05(6,28)} \approx 2.42$.

Conclusões: Aquando da construção da tabela viu-se que $F_{calc} = 3.038033 > 2.42$. Logo, rejeita-se H_0. Assim, conclui-se que nem todos os genótipos têm igual rendimento médio. Existem efeitos de genótipo significativos.

4. No enunciado é indicado que o rendimento médio amostral do genótipo `B226` é 2.1526. Para comparar as médias de pares de genótipos usaremos os testes de Tukey. Concluimos que as médias populacionais de dois níveis são diferentes quando o módulo da diferença das respectivas médias amostrais exceder o termo de comparação, ou seja, conclui-se que $\mu_i \neq \mu_{i'}$ caso $|\overline{y}_{i.} - \overline{y}_{i'.}| > q_{\alpha(k,n-k)} \sqrt{\frac{QMRE}{n_c}}$. O único valor ainda não calculado é o quantil da distribuição de Tukey. Para o nível de significância $\alpha = 0.05$ tem-se $q_{0.05\,(7,28)} = 4.49$. Assim, o termo de comparação virá $q_{\alpha(k,n-k)} \sqrt{\frac{QMRE}{n_c}} = 4.49 \sqrt{\frac{0.4249614}{5}} = 1.308989$. Logo, os genótipos i com rendimento significativamente inferior ao rendimento do genótipo `B226` são aqueles para os quais $\overline{y}_{5.} - \overline{y}_{i.} > 1.308989 \Leftrightarrow \overline{y}_{i.} < \overline{y}_{5.} - 1.308989 = 2.1526 - 1.308989 = 0.843611$. Há apenas dois genótipos nestas condições: `B020` (para o qual $\overline{y}_{1.} = 0.5280$) e `B263` (com $\overline{y}_{7.} = 0.8172$). Assim, a afirmação do enunciado não é válida.

5. Nesta alínea não se alteram os dados, mas sim a descrição da forma como foram obtidos. Não havia cinco repetições para cada genótipo, mas sim cinco valores obtidos, para cada genótipo, em cada uma de cinco diferentes localidades, que constituem assim um novo factor.

   (a) Estamos perante uma ANOVA a dois factores: genótipos (Factor A com $a = 7$ níveis) e localidades (Factor B com $b = 5$ níveis). Como todos os genótipos foram observados em todas as localidades, o delineamento é factorial. Mas como apenas houve *uma* observação de cada genótipo em cada local, não há repetições, inviabilizando a utilização do modelo ANOVA com efeitos de interacção. A equação do modelo ANOVA a dois factores, sem efeitos de interacção, é: $Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$, onde i indica genótipo, j indica localidade e $k = 1$ indica repetição (como não há repetições, a utilização deste terceiro índice é dispensável). Nesta equação, as constantes β_j são os efeitos de localidade, convencionando-se que $\beta_1 = 0$.

   (b) A nova tabela resumo terá uma linha adicional, correspondente à variabilidade imputável ao novo factor (localidades). Nesta nova tabela, os graus de liberdade, Soma de Quadrados e Quadrado Médio correspondente ao factor genótipo (Factor A) são iguais, tal como igual permanece SQT (que não depende do modelo ajustado). No enunciado é dado o valor $SQB = 3.104$. Como os graus de liberdade associados ao novo factor B são $b-1 = 4$, tem-se $QMB = \frac{SQB}{b-1} = \frac{3.104}{4} = 0.776$. A Soma de Quadrados que é agora imputável ao factor B (localidade) tem de ser retirada à anterior Soma de Quadrados Residual (uma vez que a soma das 3 novas SQs continua a ter de ser SQT). Logo, $SQRE = SQT - (SQA + SQB) = 19.6452 - (7.74628 + 3.104) = 8.79492$. Os graus de liberdade residuais são (como em qualquer modelo linear) dados pela diferença entre o número de observações (n) e o número de parâmetros do modelo (que é agora $a + b - 1$), logo são $n - (a + b - 1) = 35 - 11 = 24$. O Quadrado Médio Residual é agora $QMRE = \frac{SQRE}{a+b-1} = \frac{8.79492}{24} = 0.366455$. As estatísticas F dos testes são agora $F_A = \frac{QMA}{QMRE} = 3.523071$ e $F_B = \frac{QMB}{QMRE} = 2.117586$

Fontes de Variação	gl	Somas de Quadrados	Quadrados Médios	F_{calc}
Genótipo (Factor A)	6	SQA=7.74628	QMA=1.291047	$F_A = 3.523071$
Localidade (Factor B)	4	SQB=3.104	QMB= 0.776	$F_B = 2.117586$
Residual	24	SQRE=8.79492	QMRE=0.366455	—
Total	34	19.6452	–	–

   (c) No novo modelo, o teste aos efeitos do factor genótipo (que continua a ter Hipótese Nula $\alpha_i = 0$ para todo o i), vai agora ter estatística com valor calculado $F_{A_{calc}} = 3.523071$. O limiar da região crítica muda (dada a mudança dos segundos graus de liberdade da distribuição F), sendo agora $f_{\alpha\,[a-1, n-(a+b-1)]} = f_{0.05\,(6,24)} \approx 2.49$. Mas a conclusão mantém-se: rejeita-se H_0, concluindo-se que existem efeitos significativos de genótipo.

IV

1. (a) Pelo formulário,

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

(b) A relação entre R^2_{mod} e R^2 obtida na alínea anterior é uma relação linear crescente (trata-se da equação duma recta relacionando $y = R^2_{mod}$ e $x = R^2$, com declive positivo: $\frac{n-1}{n-(p+1)} > 0$). Logo, o menor valor possível de R^2_{mod} corresponde ao menor valor possível de R^2, tendo-se $R^2 = 0 \Rightarrow R^2_{mod} = \frac{-p}{n-(p+1)}$. Analogamente, o maior valor possível de R^2_{mod} corresponde ao maior valor possível de R^2, tendo-se $R^2 = 1 \Rightarrow R^2_{mod} = \frac{-p}{n-(p+1)} + \frac{n-1}{n-(p+1)} = \frac{n-p-1}{n-(p+1)} = 1$. Assim, os valores possíveis de R^2_{mod} são os do intervalo $\left[\frac{-p}{n-(p+1)}, 1\right]$.

2. (a) O modelo RLM em notação vectorial é constituido pela equação do modelo, e pela indicação dos pressupostos exigidos ao vector dos erros aleatórios. Mais concretamente,

   i. $\vec{\mathbf{Y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}$ (equação do modelo).

   ii. $\vec{\boldsymbol{\epsilon}} \frown \mathcal{N}_n(\vec{\mathbf{0}}, \sigma^2 \mathbf{I}_n)$ (pressupostos sobre os erros aleatórios).

   onde:

   - $\vec{\mathbf{Y}} = (Y_1, ..., Y_n)^t$ é o vector aleatório das n observações da variável resposta;
   - $\mathbf{X}$ é a matriz do modelo (não aleatória) de dimensões $n \times (p+1)$, tendo-se uma primeira coluna de uns, associada a constante aditiva do modelo (β_0) e p colunas adicionais, cada uma das quais contém as n observações de cada variável preditora;
   - $\vec{\boldsymbol{\beta}} = (\beta_0, \beta_1, ..., \beta_p)^t$ é o vector (não aleatório) dos $p+1$ parâmetros do modelo:;
   - $\vec{\boldsymbol{\epsilon}} = (\epsilon_1, ..., \epsilon_n)^t$ é o vector aleatório dos n erros aleatórios;
   - $\mathbf{I}_n$ é a matriz identidade de dimensão $n \times n$;
   - σ^2 é uma constante, que corresponde à variância comum de todos os erros aleatórios.

(b) Sabemos que o vector dos valores ajustados, $\hat{\vec{\mathbf{Y}}}$, é dado por $\hat{\vec{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$. Substituindo a equação do modelo, tem-se:

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

tendo em conta a expressão da matriz de projecção ortogonal e a definição de matriz inversa.

(c) i. Tendo em conta as propriedades operatórias das matrizes de (co-)variâncias, e o facto de o vector $\mathbf{X}\vec{\boldsymbol{\beta}}$ e a matriz $\mathbf{H}$ serem não aleatórias, tem-se:

$$V[\hat{\vec{\mathbf{Y}}}] = V[\mathbf{X}\vec{\boldsymbol{\beta}} + \mathbf{H}\vec{\boldsymbol{\epsilon}}] = V[\mathbf{H}\vec{\boldsymbol{\epsilon}}] = \mathbf{H} \cdot V[\vec{\boldsymbol{\epsilon}}] \cdot \mathbf{H}^t = \mathbf{H} \cdot \sigma^2\mathbf{I}_n \cdot \mathbf{H}^t = \sigma^2\,\mathbf{H}\mathbf{H}^t\ .$$

Ora, sabemos das aulas que a matriz de projecção ortogonal é simétrica ($\mathbf{H} = \mathbf{H}^t$) e idempotente ($\mathbf{H}\mathbf{H} = \mathbf{H}$). Logo, a expressão final é equivalente a ter-se $V[\hat{\vec{\mathbf{Y}}}] = \sigma^2\,\mathbf{H}$, como se pedia para mostrar.

ii. O elemento genérico do vector $\hat{\vec{\mathbf{Y}}}$ é $\hat{Y}_i$, e a correspondente variância é dada pelo i-ésimo elemento diagonal da matriz obtida na alínea anterior, ou seja, $V[\hat{Y}_i] = \sigma^2\,h_{ii}$. Mas h_{ii} é o efeito alavanca da i-ésima observação. Assim, $V[\hat{Y}_i]$ é o produto da variância comum dos erros aleatórios do modelo (σ^2) vezes o efeito alavanca da observação correspondente ao valor ajustado $\hat{Y}_i$. Como σ^2 é comum a todas as observações, pode dizer-se que as variâncias dos valores ajustados $\hat{Y}_i$ são proporcionais aos correspondentes efeitos alavanca.

iii. Sabemos que os efeitos alavanca satisfazem as desigualdades $\frac{1}{n} \leq h_{ii} \leq 1$. Logo, multiplicando por $\sigma^2 > 0$, tem-se: $\frac{\sigma^2}{n} \leq \sigma^2\,h_{ii} = V[\hat{Y}_i] \leq \sigma^2$. Por outro lado, ao abrigo do modelo linear, σ^2 é não apenas a variância dos erros aleatórios ϵ_i, mas dos

correspondentes valores observados $Y_i = \beta_0 + \beta_1 x_1 + ...\beta_p x_p + \epsilon_i$, já que as constantes aditivas $\beta_0 + \beta_1 x_1 + ...\beta_p x_p$ não alteram as variâncias. Assim, $V[\hat{Y}_i] \leq \sigma^2 = V[\epsilon_i] = V[Y_i]$, como se pedia para mostrar.