INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2021-22

2 Fevereiro 2022 Segunda Chamada de EXAME Uma resolução possível

I

1. Trata-se duma regressão linear múltipla de `antocianas` sobre $p=4$ preditores.

   (a) É pedido o valor (omisso no enunciado) de $R^2 = \frac{SQR}{SQT}$. É conhecido o valor de $F_{calc} = \frac{QMR}{QMRE} = 171.9$. Também disponível no enunciado está a estimativa do erro padrão dos erros aleatórios (sob a designação *Residual standard error*), ou seja, $\sqrt{QMRE} = 17.55$. Assim, $QMR = 171.9 \times 17.55^2 = 52945.63$, pelo que $SQR = QMR \cdot p = 52945.63 \times 4 = 211\,782.5$. Como $SQT = (n-1) \cdot s_y^2 = 122 \times 45.10046^2 = 248\,154.3$, tem-se $R^2 = \frac{211\,782.5}{248\,154.3} = 0.8534307$. Assim, o modelo explica um pouco mais de 85% da variabilidade observada nos teores de antocianas, o que é um bom valor.

   (b) A frase do enunciado chama a atenção que o coeficiente do preditor `fenois`, $b_4 = 0.25349$ é, em termos absolutos, o mais próximo de zero. Mas essa constatação não é relevante se não fôr acompanhada da medida do erro padrão associado, que neste caso é $\hat{\sigma}_{\hat{\beta}_4} = 0.02776$. Assim, num teste *t-Student* às hipóteses $H_0 : \beta_4 = 0$ versus $H_1 : \beta_4 \neq 0$, a estatística do teste terá valor calculado $T_{calc} = \frac{b_4 - \beta_{4|H_0}}{\hat{\sigma}_{\hat{\beta}_4}} = \frac{0.25349-0}{0.02776} = 9.132$ (disponível no enunciado). Este valor calculado é claramente muito superior aos valores tabelados para um teste de hipóteses aos níveis de significância usuais. Por exemplo, para $\alpha = 0.01$, tem-se $t_{\frac{\alpha}{2}(118)} \approx t_{0.005(120)} = 2.61742$ (e para valores de α maiores, este limiar da região crítica seria ainda menor). Assim, há uma rejeição clara da Hipótese Nula $H_0 : \beta_4 = 0$, pelo que $b_4 = 0.25349$ é significativamente diferente de zero, logo o preditor x_4 (`fenois`) *não* pode ser excluído do modelo sem afectar de forma significativa a qualidade do ajustamento. Curiosamente, trata-se do preditor para o qual o valor da estatística T_{calc} nos testes a $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ é maior. Desta forma, trata-se da estatística com o *menor p-value*, ou seja, trata-se do preditor cuja exclusão tem o *maior* efeito sobre a qualidade de ajustamento do modelo, precisamente o oposto daquilo que é afirmado na frase do enunciado.

   (c) i. O gráfico da esquerda é um *qq-plot*, ou seja, um gráfico de quantis empíricos dos resíduos estandardizados (cujos valores se podem ler no eixo vertical) versus quantis teóricos duma Normal reduzida. Sendo verdadeiro o pressuposto de Normalidade dos erros aleatórios do modelo, os pontos deste gráfico deverão ter uma forte colinearidade. Uma vez que esta disposição linear aproximada é observada de forma bastante clara no gráfico, considera-se válido o referido pressuposto de Normalidade dos ϵ_is. O gráfico da direita é um diagrama de barras indicando as distâncias de Cook (no eixo vertical) de cada uma das 123 observações. A distância de Cook D_i mede a influência duma dada observação, ou seja, a modificação nos valores de $\hat{y}_i$ produzida pela modificação na hipersuperfície ajustada que resultaria de excluir a i-ésima observação. A única observação com uma distância de Cook considerável é a observação 78, para a qual se tem $D_{78} \approx 0.35$. Embora seja um valor aquém do limiar de alerta 0.5, é um valor assinalável, que mereceria uma inspecção das possíveis causas para uma tão grande influência desta observação.

   ii. No formulário dispõe-se da fórmula $D_i = R_i^2 \cdot \left(\frac{h_{ii}}{1-h_{ii}}\right) \cdot \frac{1}{p+1}$, sendo R_i o resíduo estandardizado da i-ésima observação e h_{ii} o respectivo efeito alavanca. É conhecido o número de variáveis preditoras, $p = 4$, e a partir dos gráficos podem obter-se valores

aproximados da distância de Cook e do resíduo estandardizado da observação 78, nomeadamente: $D_{78} \approx 0.35$ e $R_{78} \approx -3$. A partir da fórmula tem-se, aproximadamente, para a observação 78:

$$\begin{aligned} & \frac{h_{78,78}}{1-h_{78,78}} = \frac{(p+1)\cdot D_{78}}{R_{78}^2} \approx \frac{5\times 0.35}{9} = 0.1944444 \\ \Leftrightarrow \quad & \frac{1-h_{78,78}}{h_{78,78}} = \frac{1}{h_{78,78}} - 1 \approx \frac{1}{0.1944444} = 5.1425858 \\ \Leftrightarrow \quad & \frac{1}{h_{78,78}} \approx 6.142858 \quad \Leftrightarrow \quad h_{78,78} \approx \frac{1}{6.142858} = 0.1627907 \end{aligned}$$

Sabe-se que o valor médio dos efeitos alavanca numa regressão linear múltipla com p preditores e n observações é dado por $\overline{h} = \frac{p+1}{n}$, que no nosso caso dá $\overline{h} = \frac{5}{123} = 0.04065$. Logo, a observação 78 tem um efeito alavanca que é mais de quatro vezes superior ao efeito alavanca médio. Trata-se dum valor razoavelmente elevado, embora ainda distante do máximo valor possível (que é o valor 1).

(d) Agora consideram-se apenas regressões lineares simples com a variável resposta `antocianas`.

i. O melhor modelo será o que tem o preditor mais fortemente correlacionado com a variável resposta `antocianas`. Pelo enunciado constata-se ser o preditor `fenois`, com coeficiente de correlação linear $r = 0.8370163$. O coeficiente de determinação associado a esta regressão linear simples será $R^2 = 0.8370163^2 = 0.7005963$. Assim, essa regressão explica um pouco mais de 70% da variabilidade observada nos teores de antocianas.

ii. A equação da recta é $y = b_0 + b_1\, x$, sendo o declive $b_1 = \frac{cov_{xy}}{s_x^2} = r_{xy}\,\frac{s_y}{s_x} = 0.8370163 \times \frac{45.10046}{80.49320} = 0.4689815$ e a ordenada na origem $b_0 = \overline{y} - b_1\,\overline{x} = 179.09 - 0.4689815 \times 458.4 = -35.89112$. Assim, a equação da recta de regressão é $y = -35.89112 + 0.4689815\, x$.

iii. O maior valor do preditor `fenois` é, pelo enunciado, 755.1. Logo, o correspondente teor de antocianas previsto será $\hat{y} = -35.89112 + 0.4689815 \times 755.1 = 318.2368$. Sabendo que o correspondente valor observado do teor de antocianas é 230.49, o respectivo resíduo será $e = y - \hat{y} = 230.49 - 318.2368 = -87.7468$. O facto de se tratar dum resíduo negativo indica que o ponto correspondente a essa observação encontra-se abaixo da recta de regressão ajustada, na nuvem de n pontos nessas duas variáveis.

iv. Para calcular o valor de QMRE neste modelo, podemos partir do valor, calculado acima, do coeficiente de determinação: $R^2 = \frac{SQR}{SQT} = 0.7005963$. Como a variável resposta (`antocianas`) é igual à do modelo inicial, e como SQT não depende do modelo ajustado, mas apenas das observações da variável resposta (que são as mesmas), tem-se $SQT = 248\,154.3$, logo $SQR = SQT \times R^2 = 248\,154.3 \times 0.7005963 = 173856$. Assim, $SQRE = SQT - SQR = 248\,154.3 - 173856 = 74298.3$. Finalmente, $QMRE = \frac{SQRE}{n-2} = \frac{74298.3}{121} = 614.0355$.

v. 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 $b_1 = 0.4689815$. Pelas tabelas da distribuição *t-Student* podemos considerar que $t_{0.025(121)} \approx t_{0.025(120)} = 1.97993$. Pelo formulário sabemos que $V[\hat{\beta}_1] = \frac{\sigma^2}{(n-1)\,s_x^2}$, pelo que é estimada por $\hat{\sigma}^2_{\hat{\beta}_1} = \hat{V}[\hat{\beta}_1] = \frac{QMRE}{(n-1)\,s_x^2}$. Logo, $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{614.0355}{121\times 80.49320^2}} = 0.02798626$. Substituindo estes valores na fórmula do IC obtém-se o seguinte intervalo a 95% de confiança para β_1:] 0.4135707 , 0.5243923 [.

vi. É pedido um teste F parcial para comparar o modelo inicial, com $p = 4$ preditores, e este seu submodelo, com apenas $k = 1$ preditor. Tem-se:

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

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

Nível de significância: $\alpha = 0.05$

Região Crítica: Unilateral direita. Rejeita-se H_0 se
$F_{calc} > f_{\alpha[p-k,n-(p+1)]} = f_{0.05(3,118)} \approx f_{0.05(3,120)} = 2.68$.

Conclusões: Já se viu que $R_c^2 = 0.8534307$ e $R_s^2 = 0.7005963$. Logo, tem-se $F_{calc} = \frac{118}{3} \times \frac{0.8534307 - 0.7005963}{1-0.8534307} = 41.01464$. Rejeita-se claramente H_0, concluindo-se que os dois modelos têm qualidade de ajustamento significativamente diferente. A conclusão era expectável, dado o valor algo diferente dos dois coeficientes de determinação.

2. Tem-se agora uma regressão linear múltipla de `rend` sobre todas as restantes variáveis observadas.

   (a) O número p de variáveis preditoras é o primeiro dos dois parâmetros da distribuição F associada ao teste de ajustamento global do modelo. Como na última linha da listagem constante do enunciado são indicados os graus de liberdade dessa distribuição, conclui-se imediatamente que $p = 7$.

   (b) Pede-se um teste de ajustamento global do modelo. Tem-se:

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

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

   Nível de significância: $\alpha = 0.05$.

   Região Crítica: Unilateral direita. Rejeitar H_0 se $F_{calc} > f_{0.05[7,115]} \approx f_{0.05[7,120]} = 2.09$.

   Conclusões: O valor calculado da estatística é dado no enunciado: $F_{calc} = 0.9454$. Assim (e como acontece sempre que os valores de F_{calc} são inferiores a 1, seja qual fôr o nível de significância usual escolhido), não se rejeita H_0. Desta forma, o nosso modelo não difere significativamente do Modelo Nulo (sem preditores), não sendo por isso possível recomendar a sua utilização. Esta conclusão era expectável, dado o valor muito próximo de zero do coeficiente de determinação amostral, $R^2 = 0.05442$. Na realidade, o modelo ajustado apenas explica pouco mais de 5% da variabilidade nos rendimentos observados.

   (c) Pelo formulário, sabemos que $R_{mod}^2 = 1 - \frac{QMRE}{QMT}$. No enunciado está disponível (com a designação *Residual standard error*) o valor $\sqrt{QMRE} = 1.046$. Logo, tem-se $QMRE = 1.046^2 = 1.094116$. Por outro lado, $QMT = \frac{SQT}{n-1}$ não é mais que a variância dos valores observados da variável resposta, s_y^2. Mas este valor não está directamente disponível no enunciado (a variável resposta é `rend`). No entanto, é possível calculá-lo a partir do valor dado de $F_{calc} = \frac{QMR}{QMRE} = 0.9454$, donde $QMR = 0.9454 \times 1.094116 = 1.034377$. Logo, $SQR = QMR \times p = 1.034377 \times 7 = 7.240641$. Como $SQRE = QMRE \times [n-(p+1)] = 1.094116 \times 115 = 125.8233$. Assim, $SQT = SQR + SQRE = 7.240641 + 125.8233 = 133.064$ e $QMT = \frac{SQT}{n-1} = \frac{133.064}{122} = 1.090689$. Tem-se então $R_{mod}^2 = 1 - \frac{1.094116}{1.090689} = -0.003142051$. O facto de este valor ser negativo (o que é possível, como sabemos pelos Exercícios das aulas e a partir duma pergunta na Primeira Chamada de exame deste ano lectivo) reflecte o facto de $QMRE > QMT$. Ora $QMRE$ estima a variância dos erros aleatórios (σ^2), que é também a variância das observações de Y em torno do hiperplano em $\mathbb{R}^{p+1}$ associado ao modelo. Por outro lado, $QMT = s_y^2$ é a variância das observações de Y sem qualquer modelo explicativo. Um valor negativo de R_{mod}^2 indica assim que a variabilidade de Y sem preditores explicativos é menor do que a variabilidade (inexplicada) de Y num modelo em que se introduziram preditores na tentativa de explicar a variabilidade de Y. Por outras palavras, o modelo tem um desempenho pior do que a ausência dum modelo explicativo.

II

1. Trata-se dum delineamento experimental factorial, com dois factores: variedade (com $a = 8$ níveis) e técnica de cultivo (com $b=4$ níveis). Em cada uma das $ab=32$ situações experimentais existem $n_{ij}=n_c=3$ observações, pelo que se trata dum delineamento equilibrado, com um total de $abn_c=96$ observações. Havendo repetições, é possível ajustar o modelo ANOVA com efeitos de interacção, a seguir indicado:

   Equação do Modelo: $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, onde $i=1,...,8$ indica variedade; $j = 1,2,3,4$ indica técnica de cultivo; $k = 1,2,3$ repetição (dentro de cada combinação variedade/técnica); Y_{ijk} indica o número de sementes germinadas da k-ésima repetição da variedade i com a técnica de cultivo j; ϵ_{ijk} é o correspondente erro aleatório. Com as restrições $\alpha_1=0$, $\beta_1=0$ e $(\alpha\beta)_{ij}=0$ se $i=1$ e/ou $j=1$, a constante aditiva comum a todas as observações, μ_{11}, representa o número esperado de sementes germinadas da primeira variedade com a primeira técnica de cultivo; α_i indica o acréscimo associado à variedade i; β_j indica o acréscimo associado à técnica de cultivo j e $(\alpha\beta)_{ij}$ indica o efeito de interacção entre variedade i e técnica de cultivo j.

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

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

   Assinale-se que, neste caso, a variável resposta Y_{ijk} é uma variável de contagem, logo apenas pode tomar como valores os números inteiros entre 0 e 100. Rigorosamente falando, trata-se duma variável aleatória discreta, logo seria impossível ter uma distribuição Normal. No entanto, e uma vez que o número de possíveis valores é bastante elevado, pode considerar-se que esse pressuposto do modelo é aproximadamente válido.

2. A tabela de síntese tem três linhas associadas aos tipos de efeitos no modelo (factor A, factor B, e interacção) e ainda a linha associada à variabilidade residual (sem contar com a linha correspondente à variabilidade total). A coluna dos graus de liberdade tem os valores $a-1=7$ (Factor A); $b-1=3$ (Factor B), $(a-1)(b-1)=21$ (Interacção); e $n-ab=96-32=64$ (Residual). São dados no enunciado três valores da coluna de Somas de Quadrados: $SQA=763.16$; $SQB=30774.28$; e $SQRE = 2578.67$. Facilmente se calcula a Soma de Quadrados Total: $SQT = s_y^2 \cdot (n-1) = 386.6973 \times 95 = 36736.24$. Daqui se conclui que a Soma de Quadrados associada à interacção é dada por: $SQAB = SQT - (SQA+SQB+SQRE) = 36736.24 - (763.16+30774.28+2578.67) = 2620.13$. Dividindo cada Soma de Quadrados pelos respectivos graus de liberdade obtêm-se os quatro Quadrados Médios: $QMA = \frac{SQA}{a-1} = 109.0229$; $QMB = \frac{SQB}{b-1} = 10258.09$; $QMAB = \frac{SQAB}{(a-1)(b-1)} = 124.7681$; e $QMRE = \frac{SQRE}{n-ab} = 40.29172$. Finalmente, dividindo os três Quadrados Médios associados a cada tipo de efeito pelos Quadrado Médio Residual, obtêm-se os valores calculados das estatísticas F associado a cada um dos três testes aos tipos de efeitos: $F_{A_{calc}} = \frac{QMA}{QMRE} = 2.705839$; $F_{B_{calc}} = \frac{QMB}{QMRE} = 254.5955$; e $F_{AB_{calc}} = \frac{QMAB}{QMRE} = 3.096619$. Eis a tabela-resumo obtida:

Fonte de variação	g.l.	Soma de Quadrados	Quadrado Médio	F_{calc}
Variedade (Factor A)	7	763.16	109.0229	2.705839
Técnica (Factor B)	3	30774.28	10258.09	254.5955
Interacção	21	2620.13	124.7681	3.096619
Residual	64	2578.67	40.29172	–

3. Há três tipos de efeitos previstos pelo modelo: os efeitos principais de variedade (α_i); os efeitos principais de técnicas de cultivo (β_j); e os efeitos de interacção ($(\alpha\beta)_{ij}$). A cada um corresponde um teste F, cuja Hipótese Nula é sempre a inexistência desse tipo de efeitos e a Hipótese

Alternativa é sempre que, em pelo menos um caso, esse tipo de efeito é não nulo. Por exemplo, no teste aos efeitos principais do Factor A, tem-se $H_0 : \alpha_i = 0$, $\forall\, i$ e $H_1 : \exists\, i$ tal que $\alpha_i \neq 0$. As estatísticas de teste são as razões entre o Quadrado Médio do respectivo tipo de efeito e o Quadrado Médio Residual, como indicado na alínea anterior, aquando da construção da tabela de síntese da ANOVA. As suas distribuições sob H_0 são F, com graus de liberdade associados ao numerador e denominador, respectivamente, da estatística. As regiões críticas são unilaterais direitas. Usando o nível de significância $\alpha = 0.05$, rejeita-se H_0 no teste aos α_i caso $F_{A_{calc}} > f_{0.05(7,64)}$, um valor compreendido entre os valores tabelados $f_{0.05(7,60)} = 2.17$ e $f_{0.05(7,120)} = 2.09$. Como se pode constatar a partir da tabela, o valor da estatística correspondente, $F_{A_{calc}} = 2.705839$ pertence à Região Crítica, pelo que se conclui pela existência de efeitos de variedade. No teste à existência de efeitos principais de técnica de cultivo, rejeita-se H_0 se $F_{B_{calc}} > f_{0.05(3,64)}$, um valor compreendido entre os valores tabelados $f_{0.05(3,60)} = 2.76$ e $f_{0.05(3,120)} = 2.68$. O enorme valor calculado da estatística ($F_{B_{calc}} = 254.5955$) implica uma claríssima rejeição de H_0, logo a existência de pelo menos uma técnica de cultivo com efeito principal não nulo. Finalmente, no teste aos efeitos de interacção, rejeita-se H_0 se $F_{AB_{calc}} > f_{0.05(21,64)}$, um valor compreendido entre os valores tabelados $f_{0.05(20,60)} = 1.75$ e $f_{0.05(25,120)} = 1.60$. Como $F_{AB_{calc}} = 3.096619$, também neste caso se rejeita H_0. Assim, os três tipos de efeitos previstos pelo modelo são significativos.

4. Pergunta-se se o menor número médio de sementes germinadas associado à técnica $T1$, que é com a variedade 7, $\overline{y}_{7,1} = 46.33$, se pode considerar significativamente diferente da maior germinação média registada com qualquer das outras técnicas de cultivo, que é $\overline{y}_{8,3} = 31.33$. Se assim fôr, será possível afirmar que, qualquer que seja a variedade, a técnica $T1$ produz sempre germinações médias superiores às restantes técnicas. Vamos comparar as médias de célula referidas usando o termo de comparação de Tukey, que é dado por $\tau = q_{\alpha(ab,n-ab)} \sqrt{\frac{QMRE}{n_c}}$. Ao nível de significância $\alpha = 0.05$ o quantil da distribuição de Tukey toma valor $q_{0.05(32,64)} \approx q_{0.05(30,60)} = 5.57$. Como $QMRE = 40.29172$ e $n_c = 3$, tem-se o termo de comparação $\tau = 5.57 \times \sqrt{\frac{40.29172}{3}} = 20.41279$. Ora a diferença entre as duas médias de célula referidas é $|\overline{y}_{7,1} - \overline{y}_{8,3}| = 15.00$, logo inferior ao termo de comparação. Assim, estas duas difereças não podem ser consideradas significativas (ao nível $\alpha = 0.05$). Analisando a tabela das médias de célula constante do enunciado, rapidamente se conclui que apenas três células, todas associadas à técnica de cultivo T3 (e, concretamente, às variedades V3, V6 e V8) têm germinações médias que não diferem de forma significativa da menor germinação média com a técnica T1. Assim, pode concluir-se que, para as variedades estudadas, o número médio de sementes germinadas obtido com a técnica T1 é sempre significativamente diferente dos obtidos com T2 e T4.

5. É pedido para calcular a estimativa do parâmetro β_2. A partir do formulário verifica-se: $b_2 = \overline{y}_{12\cdot} - \overline{y}_{11\cdot} = 11.67 - 66.33 = -54.66$. Este valor indica que, na passagem da técnica de cultivo T1 para a técnica de cultivo T2, usando a variedade de referência V1, regista-se uma quebra média de 54.66 sementes germinadas. Este valor é coerente com a discussão da alínea anterior, onde se verificou que as germinações médias associadas a T1 eram sempre significativamente diferentes das verificadas com T2. Ajuda igualmente a compreender o resultado do teste F à existência de efeitos principais significativos do factor B, sendo natural que se tenha $\beta_2 \neq 0$. Repare-se que considerações análogas poderiam ser feitas para os efeitos β_3 e β_4, já que as estimativas b_3 e b_4 são semelhantes em valor a b_2.

6. Estamos perante um gráfico de interacção, cujo eixo horizontal está associado aos oito diferentes níveis do factor A (variedades) e o eixo vertical corresponde às germinações médias observadas em cada célula. Por cima de cada marcador do eixo horizontal (ou seja, de cada variedade) há quatro pontos, cujas alturas correspondem às germinações médias obtidas, para essa variedade,

com cada uma das técnicas de cultivo. Os pontos correspondentes a uma mesma técnica de cultivo são unidos por segmentos de rectas, com estilos indicados na legenda que é visível no canto superior direito do gráfico. Talvez a característica mais evidente deste gráfico seja o posicionamento das médias associadas à técnica T1 na parte superior do gráfico, reflectindo as germinações maiores obtidas com a técnica T1 que, como vimos em alíneas anteriores são quase sempre significativamente maiores das obtidas com outras técnicas de cultivo. No entanto, o próprio facto de haver três células onde essas diferenças não são significativas é indiciador de outra conclusão da nossa ANOVA: a existência de efeitos significativos de interacção. Estes efeitos (cuja significância não foi, nos testes F, tão clara como a dos efeitos principais do factor B) pode ser vista na diferença relativamente menor entre as germinações médias das técnicas T1 e T3 na parte direita do gráfico (em particular, com as variedades V6, V7 e V8), quando comparada com as diferenças análogas na parte esquerda do gráfico. Os efeitos principais do factor variedade, que eram significativos, mas por pouco, no respectivo teste F, são os mais difíceis de visualizar no gráfico.

III

1. Como indicado no enunciado, o vector $\vec{\mathbf{y}}^c = \vec{\mathbf{y}} - \overline{y} \cdot \vec{\mathbf{1}}_n$ tem elemento genérico $y_i - \overline{y}$. Ora, a projecção ortogonal deste vector sobre $\mathcal{C}(\mathbf{X})$ é dada por $\mathbf{H}\vec{\mathbf{y}}^c = \mathbf{H}(\vec{\mathbf{y}} - \overline{y} \cdot \vec{\mathbf{1}}_n) = \mathbf{H}\vec{\mathbf{y}} - \overline{y} \cdot \mathbf{H}\vec{\mathbf{1}}_n$. Sabemos que $\mathbf{H}\vec{\mathbf{y}} = \vec{\hat{\mathbf{y}}}$: recorde-se que a designação em inglês da matriz de projecção ortogonal $\mathbf{H}$, *hat matrix*, resulta de o produto $\mathbf{H}\vec{\mathbf{y}}$ 'colocar o chapéu' em $\vec{\mathbf{y}}$, ou seja, converter o vector dos valores observados de y ($\vec{\mathbf{y}}$) no vector dos correspondentes valores ajustados ($\vec{\hat{\mathbf{y}}}$). Por outro lado, sabemos igualmente que qualquer vector que pertença ao espaço das colunas de $\mathbf{X}$ fica invariante quando projectado nesse mesmo espaço. Ora, $\vec{\mathbf{1}}_n \in \mathcal{C}(\mathbf{X})$, logo $\mathbf{H}\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n$. Assim, $\mathbf{H}\vec{\mathbf{y}}^c = \vec{\hat{\mathbf{y}}} - \overline{y} \cdot \vec{\mathbf{1}}_n$, ou seja, o vector $\mathbf{H}\vec{\mathbf{y}}^c$ tem como elemento genérico $\hat{y}_i - \overline{y}$. A sua norma ao quadrado é dada pela soma de quadrados dos seus elementos, ou seja, $\|\mathbf{H}\vec{\mathbf{y}}^c\|^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$. Mas este somatório é, por definição, a Soma de Quadrados associada à Regressão, ou seja, $SQR = {}_{(n-1)} \cdot s^2_{\hat{y}} = \|\mathbf{H}\vec{\mathbf{y}}^c\|^2$.

2. Por definição, a norma ao quadrado de qualquer vector é o produto interno do vector com ele próprio, ou seja, $\|\vec{\mathbf{x}}\|^2 = \vec{\mathbf{x}}^t\vec{\mathbf{x}}$. Usando a propriedade de que a transposta dum produto de matrizes é o produto das transpostas pela ordem inversa ($(\mathbf{AB})^t = \mathbf{B}^t\mathbf{A}^t$) e ainda o facto de a matriz de projecção ortogonal $\mathbf{H}$ ser simétrica ($\mathbf{H}^t = \mathbf{H}$) e idempotente ($\mathbf{HH} = \mathbf{H}$), tem-se:

$$\|\mathbf{H}\vec{\mathbf{y}}^c\|^2 = (\mathbf{H}\vec{\mathbf{y}}^c)^t\,\mathbf{H}\vec{\mathbf{y}}^c = (\vec{\mathbf{y}}^c)^t\mathbf{H}^t\mathbf{H}\vec{\mathbf{y}}^c = (\vec{\mathbf{y}}^c)^t\mathbf{H}\mathbf{H}\vec{\mathbf{y}}^c = (\vec{\mathbf{y}}^c)^t\mathbf{H}\vec{\mathbf{y}}^c\ .$$

   A expressão final é o produto interno do vector $\vec{\mathbf{y}}^c$ com o vector $\mathbf{H}\vec{\mathbf{y}}^c$. Ora, para qualquer par de vectores $\vec{\mathbf{x}}$ e $\vec{\mathbf{z}}$, o seu produto interno define-se como $\vec{\mathbf{x}}^t\vec{\mathbf{z}} = \sum_{i=1}^{n} x_i z_i$. Logo, o produto interno de $\vec{\mathbf{y}}^c$ (de elemento genérico $y_i - \overline{y}$) com $\mathbf{H}\vec{\mathbf{y}}^c$ (de elemento genérico $\hat{y}_i - \overline{y}$) pode escrever-se também como a soma dos produtos dos elementos correspondentes de cada vector: $(\vec{\mathbf{y}}^c)^t\mathbf{H}\vec{\mathbf{y}}^c = \sum_{i=1}^{n} (y_i - \overline{y})(\hat{y}_i - \overline{y})$. Comparando este somatório com a definição da covariância dada no enunciado, verifica-se que se trata de $(n-1) \cdot cov_{y,\hat{y}}$, como se pedia para mostrar.

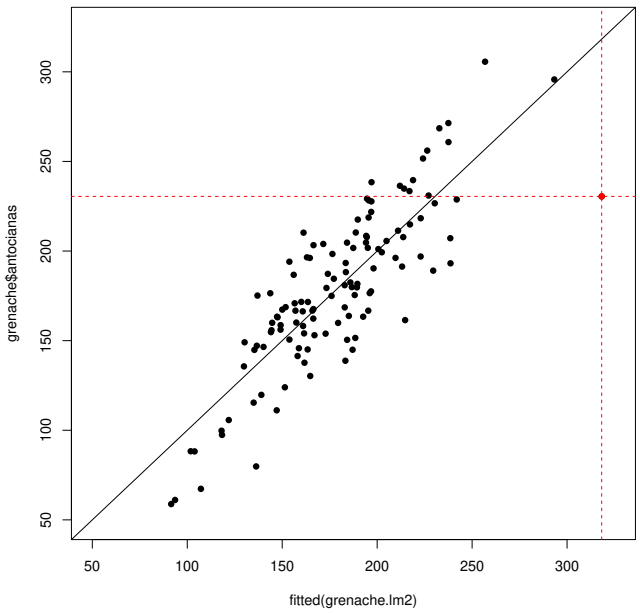
3. Pede-se para ajustar a recta de regressão para os pares de valores $\{(\hat{y}_i, y_i)\}_{i=1}^n$ (ou seja, com os valores ajustados pelo modelo de regressão linear múltipla, $\hat{y}_i$ a desempenhar o papel dos valores dum preditor x).

(a) Adaptando a notação ($x \to \hat{y}$), e tendo em conta as duas alíneas anteriores, a recta de regressão terá equação $y = b_0 + b_1\,\hat{y}$ com declive $b_1 = \frac{cov_{y,\hat{y}}}{s^2_{\hat{y}}} = \frac{\frac{\|\mathbf{H}\tilde{\mathbf{y}}^c\|^2}{n-1}}{\frac{\|\mathbf{H}\tilde{\mathbf{y}}^c\|^2}{n-1}} = 1$ e ordenada na origem $b_0 = \overline{y} - b_1\,\overline{\hat{y}}$. Mas numa regressão linear (simples ou múltipla) a média dos valores observados de y ($\overline{y}$) e dos correspondentes valores ajustados ($\overline{\hat{y}}$) é igual. Logo, e como $b_1 = 1$, tem-se $b_0 = 0$. A recta de regressão será assim a bissectriz $y = \hat{y}$.

(b) Por definição, os resíduos são a diferença entre os valores observados da variável resposta e os correspondentes valores ajustados pela recta de regressão. Logo, no nosso caso, os resíduos da recta de regressão são dados por $e_i = y_i - (b_0 + b_1\,\hat{y}_i) = y_i - \hat{y}_i$. Mas estes são, por definição, também os resíduos da regressão linear múltipla que produziu os valores $\hat{y}_i$.

(c) Se os resíduos nas duas regressões (a múltipla e a simples usando os $\hat{y}_i$ como preditor) são iguais, as respectivas Somas de Quadrados Residuais têm de também ser iguais. Uma vez que a Soma de Quadrados Total é o numerador da variância amostral dos valores observados de y, que são iguais nos dois modelos, também os SQT tomam o mesmo valor. Logo, $R^2 = \frac{SQR}{SQT} = 1 - \frac{SQRE}{SQT}$ terá de ser igual nas duas regressões.

A implicação destes resultados é que uma parte das características de uma regressão linear múltipla (cuja nuvem de pontos não é visualizável se $p > 2$) pode ser visualizada numa nuvem de pontos dos valores y_i sobre os valores de $\hat{y}_i$. Para ilustrar esta ideia, construa-se a nuvem de pontos referida, respeitante aos dados do modelo inicialmente ajustado no grupo I (dados Grenache, com 4 preditores para a variável resposta `antocianas`):



O valor de $R^2 = 0.7005963$ e de cada um dos resíduos são iguais neste gráfico e na nuvem de pontos em $\mathbb{R}^5$ associada à regressão linear múltipla de `antocianas` sobre os 4 preditores do modelo inicial. À direita no gráfico, a vermelho e debaixo da recta, encontra-se o ponto correspondente ao resíduo -87.7468 calculado na subalínea 1 d) iii) do grupo I, no cruzamento da recta horizontal $y = 230.49$ e da recta vertical $\hat{y} = 318.2368$. Embora não fosse possível sabê-lo com a informação disponível no enunciado, torna-se agora natural que se trata do ponto correspondente à observação 78, à qual está associada a maior distância de Cook.