INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2021-22

2 Fevereiro 2022 **Segunda Chamada de Exame** **Duração: 3h00**

I [10 v.]

Um estudo sobre uvas da casta Grenache gerou um conjunto de observações independentes, com várias variáveis observadas em 123 parcelas escolhidas ao acaso.

1. Foi inicialmente ajustado um modelo de regressão linear múltipla envolvendo apenas algumas das variáveis observadas, tendo-se o teor de antocianas como variável resposta (variável `antocianas`, em mg/l) e quatro variáveis preditoras: `brix` (em graus brix); `pH` (na sua escala); `acidez` (em g/l de ácido tartárico); e teor de fenóis total (variável `fenois`, em mg/l). Eis alguns indicadores, bem como a matriz de correlações das variáveis usadas neste modelo:

Variável	brix	pH	acidez	antocianas	fenois
mínimo	18.47	3.490	2.610	58.82	271.7
média	23.76	3.818	3.297	179.09	458.4
máximo	25.79	4.000	4.650	305.64	755.1
desvio padrão	1.56723	0.10890	0.34249	45.10046	80.49320

```
                brix         pH     acidez  antocianas    fenois
brix       1.0000000  0.7165871 -0.8079042  0.8189338  0.6563123
pH         0.7165871  1.0000000 -0.6108238  0.4305197  0.3219094
acidez    -0.8079042 -0.6108238  1.0000000 -0.7516607 -0.5998286
antocianas 0.8189338  0.4305197 -0.7516607  1.0000000  0.8370163
fenois     0.6563123  0.3219094 -0.5998286  0.8370163  1.0000000
```

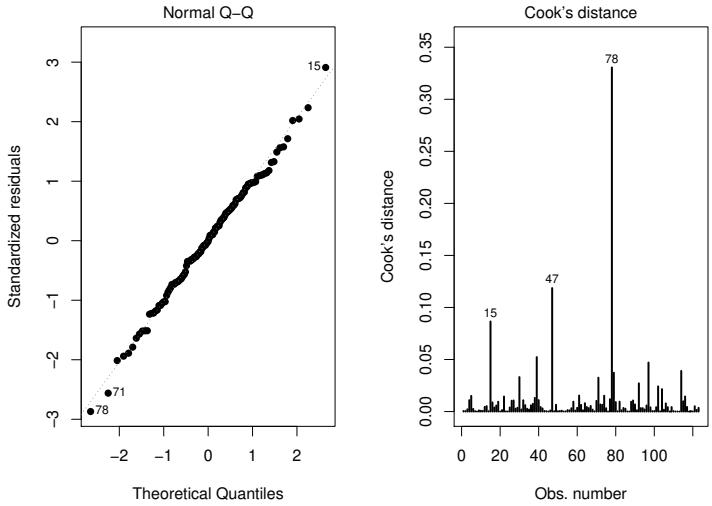
Eis os resultados obtidos:

```
  Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.32444   89.48851   1.155 0.250583
brix         14.71671    2.19840   6.694 7.72e-10
pH          -80.98132   21.98559  -3.683 0.000349
acidez      -24.56750    8.03729  -3.057 0.002768
fenois        0.25349    0.02776   9.132   ???
---
Residual standard error: 17.55 on 118 degrees of freedom
Multiple R-squared: ????,   Adjusted R-squared: 0.8486
F-statistic: 171.9 on 4 and 118 DF,  p-value: < 2.2e-16
```

   (a) Calcule a proporção da variância dos teores observados de antocianas que é explicada por esta regressão.

   (b) Comente, justificando adequadamente, a seguinte afirmação: *"como a estimativa do coeficiente do preditor* `fenois` *é, claramente, a mais próxima de zero, será a exclusão desse preditor que terá o menor efeito sobre a qualidade de ajustamento do modelo".*

   (c) Considere os seguintes gráficos:



i. Descreva e discuta os gráficos.

ii. Utilize estes gráficos para calcular um valor aproximado do efeito alavanca da observação 78. Trata-se dum valor acima ou abaixo do efeito alavanca médio?

(d) Considere agora o modelo de regressão linear simples com o maior valor de R^2 que se poderia construir a partir destes dados para modelar a variável `antocianas`.

i. Diga justificando, qual o preditor nesse modelo e qual o valor do correspondente coeficiente de determinação.

ii. Deduza a equação da recta de regressão.

iii. Calcule o valor ajustado pelo modelo para a observação com o maior valor do preditor. Sabendo que o verdadeiro teor de antocianas, para essa observação, é 230.49, calcule o respectivo resíduo (usual).

iv. Calcule o valor do Quadrado Médio Residual.

v. Construa um intervalo a 95% de confiança para o declive da recta populacional. **Nota:** Caso não tenha resolvido a alínea anterior, pode usar o valor QMRE=600.

vi. Teste se a qualidade de ajustamento deste modelo difere significativamente da do modelo inicial.

2. Seguidamente, ajustou-se uma regressão linear múltipla da variável rendimento (`rend`, em kg/planta) sobre todas as restantes variáveis observadas no estudo (e não apenas as usadas no modelo inicial). Obtiveram-se os seguintes resultados:

```
Residual standard error: 1.046 on 115 degrees of freedom
Multiple R-squared: 0.05442,  Adjusted R-squared: ????
F-statistic: 0.9454 on 7 and 115 DF,  p-value: ????
```

Responda às seguintes perguntas, justificando as suas respostas:

(a) Quantas variáveis preditoras existiam no modelo?

(b) Teste a qualidade de ajustamento do modelo. Comente.

(c) Calcule o valor do R^2 modificado. Interprete o valor obtido, tendo em conta a definição de R^2_{mod}.

II [6 valores]

Uma forma habitual de medir a viabilidade da germinação consiste em contar o número de sementes germinadas ao fim dum dado período de tempo, em condições controladas. Pretendendo-se comparar o desempenho de oito diferentes variedades duma espécie e quatro diferentes técnicas de cultivo, repetiu-se por três vezes o procedimento acima indicado, com 100 sementes em todas as possíveis combinações de técnica e variedade. Eis os números de sementes germinadas:

Variedade	Técnica de cultivo				Média	Variância
	T1	T2	T3	T4		
V1	66, 63, 70	12, 10, 13	13, 13, 11	6, 12, 7	24.67	638.6061
V2	77, 47, 66	26, 11, 18	27, 5, 11	15, 4, 15	26.833	574.1515
V3	51, 81, 63	8, 16, 14	20, 30, 29	10, 14, 10	28.833	564.333
V4	52, 40, 59	4, 15, 11	19, 16, 7	13, 9, 7	21.00	347.2727
V5	45, 51, 52	20, 13, 16	9, 10, 12	12, 12, 11	21.917	284.0833
V6	59, 66, 49	8, 8, 8	28, 32, 29	14, 21, 16	28.167	404.6970
V7	56, 38, 45	12, 16, 16	26, 16, 24	15, 8, 7	23.250	238.2045
V8	49, 41, 54	14, 20, 25	30, 28, 36	9, 15, 12	27.750	218.9318
Média	55.833	13.9167	20.042	11.417	Média global: 25.30208	
Variância	129.9710	28.6015	85.3460	15.2971	Variância global: 386.6973	

As médias das três repetições da experiência em cada combinação de técnica de cultivo e variedade (resultantes da tabela acima indicada) são:

	T1	T2	T3	T4
V1	66.33	11.67	12.33	8.33
V2	63.33	18.33	14.33	11.33
V3	65.00	12.67	26.33	11.33
V4	50.33	10.00	14.00	9.67
V5	49.33	16.33	10.33	11.67
V6	58.00	8.00	29.67	17.00
V7	46.33	14.67	22.00	10.00
V8	48.00	19.67	31.33	12.00

1. Descreva o delineamento experimental utilizado. Identifique e descreva o melhor modelo de Análise de Variância para estudar este problema.

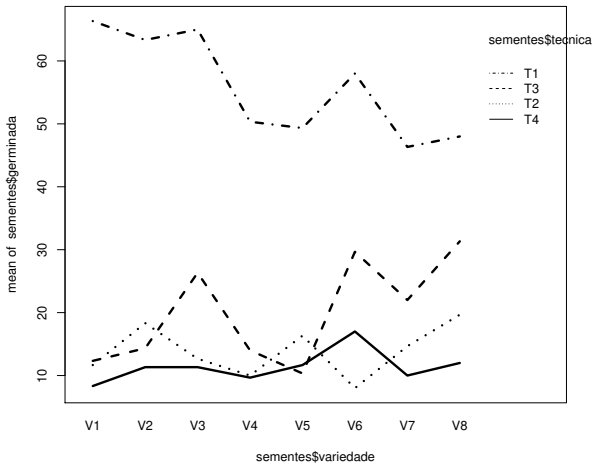
2. Construa a Tabela de Análise de Variância associada ao modelo indicado na alínea anterior, sabendo que a Soma de Quadrados associada à variação entre técnicas de cultivo é de 30774.28; que a Soma de Quadrados associada à variação entre variedades é 763.16; e que a Soma de Quadrados Residual é 2578.67.

3. Diga, justificando, quais os tipos de efeitos previstos no modelo que devem ser considerados significativos (ao nível $\alpha = 0.05$).

4. Foi sugerido que a menor das germinações associadas à técnica T1, em qualquer variedade, difere significativamente da maior das germinações com qualquer das restantes técnicas de cultivo e variedades. Teste se esta hipótese é admissível. Comente.

5. Calcule a estimativa do efeito principal associado ao tratamento T2. Interprete o valor obtido.

6. Discuta o seguinte gráfico, tendo também em conta toda a informação disponível.



III [4 valores]

Considere uma regressão linear múltipla com p preditores e ajustada com n observações, sendo $\mathbf{X}$ a respectiva matriz do modelo e $\mathbf{H}$ a matriz de projecção ortogonal sobre o respectivo espaço das colunas, $\mathcal{C}(\mathbf{X})$. Seja $\vec{\mathbf{y}}$ o vector das n observações da variável resposta e $\vec{\mathbf{y}}^c = \vec{\mathbf{y}} - \overline{y} \cdot \vec{\mathbf{1}}_n$ o correspondente vector centrado, de elemento genérico $y_i - \overline{y}$, onde $\vec{\mathbf{1}}_n$ é o vector de n uns e $\overline{y}$ a média das n observações de y.

1. Mostre que a Soma de Quadrados associada à Regressão, $SQR = (n\text{–}1)\, s^2_{\hat{y}}$, é dada pelo quadrado da norma da projecção do vector $\vec{\mathbf{y}}^c$ sobre $\mathcal{C}(\mathbf{X})$, ou seja, $SQR = \|\mathbf{H}\vec{\mathbf{y}}^c\|^2$.

2. Mostre que a covariância entre os valores observados, y_i, e os valores ajustados pelo modelo, $\hat{y}_i$, ou seja, $cov_{y\hat{y}} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})(\hat{y}_i - \overline{y})$, verifica a relação $(n\text{–}1)\, cov_{y\hat{y}} = \|\mathbf{H}\vec{\mathbf{y}}^c\|^2$.

3. Tendo em conta as alíneas anteriores, mostre que uma regressão linear simples dos valores observados y_i sobre os valores ajustados pela regressão linear múltipla, $\hat{y}_i$:

   (a) produz como recta de regressão a bissectriz $y = \hat{y}$;

   (b) tem resíduos iguais aos da regressão linear múltipla original;

   (c) tem um Coeficiente de Determinação igual ao R^2 da regressão linear múltipla original.