INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2018-19

28 de Janeiro de 2019 Segunda Chamada de EXAME Duração: 3h30

I [2,5 valores]

Pretende-se estudar a frequência do vírus do enrolamento foliar de tipo 3 (virus GLRaV3), um vírus com elevada frequência na videira. A fim de estudar se a presença do vírus, na casta Aragonez, difere consoante regiões vinícolas em Portugal e Espanha, foram aleatoriamente seleccionados um total de 664 genótipos. Cada genótipo foi classificado, quer de acordo com a região de proveniência, quer pela presença ou ausência do vírus. Os resultados obtidos estão na seguinte tabela.

	Alentejo	Dão	Douro	La Rioja	Valdepeñas	Total
Ausência	110	23	131	179	105	548
Presença	19	9	54	13	21	116
Total	129	32	185	192	126	664

1. Teste se é possível afirmar que a incidência do vírus depende das regiões. Explicite as hipóteses e estatística do teste, a natureza da região crítica, bem como o nível de significância usado. Sabendo que o valor calculado da estatística é 36.115 e admitindo válida a distribuição assintótica da estatística do teste, qual é a sua conclusão? Comente.
2. Justifique, com o mínimo de contas possível, se se verificam, ou não, as condições de Cochran.
3. Calcule a contribuição da presença do vírus no Douro para o valor da estatística do teste. Comente. Essa contribuição resulta duma associação positiva, ou negativa?

II [8,5 valores]

Um estudo sobre plantações de eucalipto visa modelar a altura média das árvores dominantes (variável `altura`, em m). Em 167 parcelas de eucalipto recolheram-se observações dessa variável resposta, bem como das seguintes variáveis: `idade` (em anos) das árvores da parcela; `diam`, o diâmetro médio (em cm) dos troncos; `vivas`, o número de árvores vivas por hectare, em cada parcela; e `areabasal` média, em m^2 por hectare. Eis alguns indicadores, e a matriz de correlações entre as variáveis observadas:

Indicador	idade	diam	altura	vivas	areabasal
mínimo	2.2	2.4	3.9	450.0	0.4
máximo	16.2	20.0	31.8	2811.0	38.6
média	8.340719	11.916766	17.529341	1271.005988	14.685629
desvio padrão	3.544709	3.353808	6.097297	499.721724	8.287654

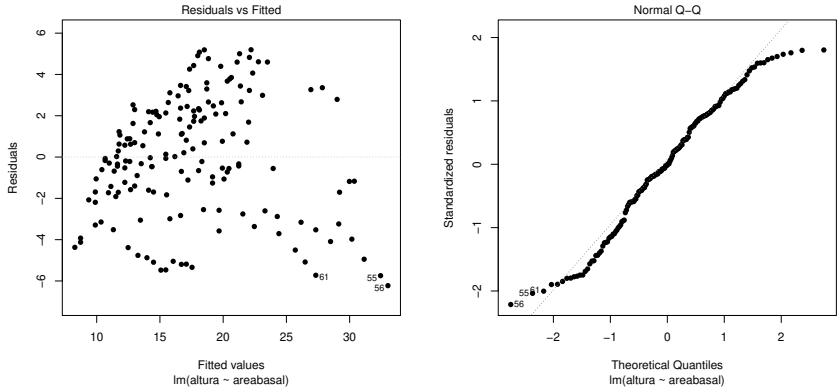
```
              idade       diam    altura       vivas areabasal
idade     1.00000000  0.7333571 0.7174714  0.06122611 0.7035585
diam      0.73335709  1.0000000 0.8744747 -0.16602431 0.7262391
altura    0.71747144  0.8744747 1.0000000  0.17699573 0.8808362
vivas     0.06122611 -0.1660243 0.1769957  1.00000000 0.4789466
areabasal 0.70355845  0.7262391 0.8808362  0.47894665 1.0000000
```

1. Inicialmente, consideraram-se apenas regressões lineares simples.

(a) Ajuste a recta de regressão linear que melhor explica a variabilidade observada na variável `altura`. Em particular, justifique a escolha de variável preditora e a forma como obteve o declive e ordenada na origem da recta. Qual a proporção de variabilidade observada nas alturas, que é explicada por essa recta?

(b) Construa um intervalo a 95% de confiança para a altura esperada, correspondente a observações com o valor médio amostral do preditor que escolheu. Comente.

(c) O modelo de regressão linear simples de `altura` sobre `areabasal` produziu os seguintes gráficos de resíduos. Descreva e comente-os, bem com as suas implicações no estudo do modelo.



(d) Foi seguidamente ajustado um modelo semelhante ao da alínea anterior, mas envolvendo a log-transformação das duas variáveis. Obtiveram-se os seguintes resultados.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.51442    0.04529   33.44   <2e-16
log(areabasal)  0.51361    0.01750   29.35   <2e-16
---
Residual standard error: 0.16 on 165 degrees of freedom
Multiple R-squared:  0.8393,   Adjusted R-squared:  0.8383
F-statistic: 861.6 on 1 and 165 DF,  p-value: < 2.2e-16
```

i. Deduza a equação da relação *não linear entre as variáveis originais* que corresponde ao modelo agora ajustado.

ii. É possível afirmar que este modelo explica uma maior proporção da variabilidade observada nas alturas das árvores do que o modelo da alínea anterior?

2. Foi decidido ajustar um modelo de regressão linear múltipla (sobre as variáveis não transformadas), utilizando todos os preditores disponíveis. Obtiveram-se os seguintes resultados:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0226249  1.5480836   0.015    0.988
idade       -0.0482364  0.0702374  -0.687    0.493
diam         1.0154064  0.1312800   7.735 1.04e-12
vivas        0.0005445  0.0006848   0.795    0.428
areabasal    0.3484105  0.0612311   5.690 5.80e-08
---
Residual standard error: 2.016 on 162 degrees of freedom
Multiple R-squared:  0.8933,   Adjusted R-squared:  0.8907
F-statistic: 339.2 on 4 and 162 DF,  p-value: < 2.2e-16
```

(a) Discuta pormenorizadamente a qualidade de ajustamento deste modelo.

(b) Execute o primeiro passo num algoritmo de exclusão sequencial baseado nos testes t e com nível de significância $\alpha = 0.05$. Comente.

(c) Teste formalmente se este modelo tem um ajustamento significativamente melhor do que o modelo de regressão linear simples que escolheu na alínea 1a). Comente.

(d) Calcule o valor do Critério de Informação de Akaike (AIC) deste modelo e do modelo que escolheu na alínea 1a). Comente.

3. Após o ajustamento dos modelos acima considerados, o investigador que recolheu os dados informou que as observações diziam respeito a apenas 19 parcelas diferentes. As 167 observações resultaram do facto de cada uma dessas parcelas ter sido observada por várias vezes, ao longo dos anos. Diga, justificando, se esta informação adicional afecta os modelos atrás ajustados.

III [5 valores]

Um estudo sobre Pera Rocha visava comparar eventuais efeitos de quatro diferentes sistemas de condução: Eixo, Palmeta, Solaxe e Tatura. Sabendo que os dois terrenos disponíveis para a produção estavam associados a diferentes condições edafo-climáticas, decidiu-se montar a experiência usando também um factor `bloco`, cujos níveis correspondem aos dois terrenos. Em cada bloco foram aleatoriamente associadas dez árvores a cada sistema de condução. Ao fim de quatro anos, procedeu-se à medição do comprimento médio dos ramos (variável `comp`, em cm) em cada árvore.

Eis as médias geral, por sistema de condução, por terreno, e em cada situação experimental:

```
Grand mean  conducao                    bloco          conducao:bloco
95.74688     Eixo Palmeta Solaxe Tatura      1      2             bloco
            95.54   86.65 101.51  99.29  100.79  90.70  conducao     1      2
                                                        Eixo     98.63  92.45
                                                        Palmeta  99.83  73.48
                                                        Solaxe   99.37 103.65
                                                        Tatura  105.35  93.23
```

Foi ajustado um modelo ANOVA, com os resultados indicados no seguinte quadro de síntese.

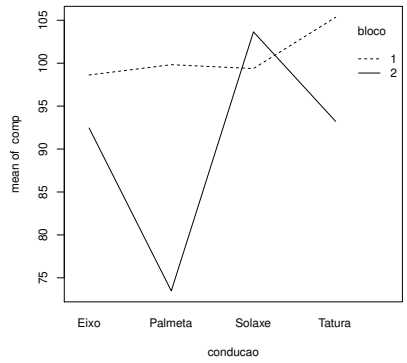
```
               Df Sum Sq Mean Sq F value  Pr(>F)
conducao       ??   2572    ???     ???    ???
bloco           1   2038  2037.7  11.570 0.00110
conducao:bloco  3   2451   817.0   4.639 0.00507
Residuals      72  12681   176.1
```

1. Identifique o delineamento experimental usado no ensaio e descreva em pormenor o modelo ANOVA ajustado. Qual o comprimento esperado dos ramos, previsto pelo modelo para o terceiro sistema de condução (Solaxe), no primeiro terreno? Qual o signficado do parâmetro α_3?

2. É possível afirmar que os diferentes sistemas de condução afectam os valores médios dos comprimentos dos ramos? Justifique, descrevendo em pormenor o teste de hipóteses que efectuar.

3. Diga, justificando de forma sintética, se foi importante prever a possibilidade de haver resultados médios diferentes nos dois terrenos.

4. Descreva o seguinte gráfico e interprete-o à luz da informação disponível. Sustente a sua discussão com os resultados de qualquer teste F que considere necessário.



5. Quais as situações experimentais em que o comprimento dos ramos é significativamente inferior ao obtido com o sistema de condução Solaxe, no segundo terreno? Justifique formalmente.

IV [4 valores]

1. Duas variáveis numéricas, X e Y, foram observadas em n entidades, produzindo os pares de observações $\{(x_i, y_i)\}_{i=1}^{n}$. Considere uma relação linear entre Y e X em que a recta é obrigada a passar na origem, ou seja, em que a tendência de fundo é dada por uma equação $y = bx$.

   (a) Defina o resíduo usual de cada observação, neste contexto.

   (b) Defina a Soma de Quadrados dos Resíduos, e deduza o valor de b correspondente a minimizar essa Soma de Quadrados.

   (c) Defina a matriz do modelo, $\mathbf{X}$, neste contexto. Utilize as fórmulas vectoriais/matricias dadas no estudo da Regressão Linear Múltipla para deduzir por essa via alternativa a expressão do estimativa de Mínimos Quadrados de b.

   (d) Mostre que, neste modelo forçado à origem, não é verdade que a média dos valores ajustados seja, em geral, igual à média dos valores observados da variável resposta.

2. Considere um delineamento factorial, a dois factores, com repetições nas células. Considere os modelos ANOVA correspondentes, com efeitos de interacção (M_{A*B}) e sem efeitos de interacção (M_{A+B}).

   (a) Mostre que a variância amostral das n observações da variável resposta é uma média ponderada dos vários Quadrados Médios de cada modelo, indicando quem são os pesos em cada parcela dessa média.

   (b) Mostre que, no modelo com efeitos de interacção, o Quadrado Médio Residual, $QMRE_{A*B}$, é maior que o Quadrado Médio associado aos efeitos de interacção se e só se fôr maior que o Quadrado Médio Residual no modelo sem efeitos de interacção, isto é, mostre que $QMRE_{A*B} > QMAB$ se e só se $QMRE_{A*B} > QMRE_{A+B}$. Interprete a condição, do ponto de vista da qualidade dos modelos com e sem efeitos de interacção.