INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2016-17

23 de Janeiro de 2017 Segunda Chamada de Exame Duração: 3h30

I [2,5 valores]

Num estudo sobre a tolerância à doença da murchidão do pinheiro, avaliou-se a sobrevivência de plantas de dois anos, provenientes de 8 famílias de meios-irmãos de pinheiro bravo, após inoculação do nemátodo *Bursaphelenchus xylophilus* em estufa. De cada família foram inoculadas 60 plantas e a sobrevivência foi avaliada 157 dias após a inoculação. Os resultados estão no quadro seguinte:

Família	A	B	C	D	E	F	G	H	Total
Sobreviventes	2	5	8	12	12	18	20	23	100
Não sobreviventes	58	55	52	48	48	42	40	37	380

1. Qual o teste adequado para determinar se a probabilidade de sobrevivência varia de forma significativa entre as famílias observadas? Explicite as hipóteses nula e alternativa, a estatística do teste e a sua distribuição assintótica, bem como a região crítica (para $\alpha = 0.05$).
2. Pode ser considerada válida a distribuição assintótica da estatística de Pearson? Justifique.
3. Sabendo que a estatística de Pearson tem valor $X^2_{calc} = 38.804$, qual a sua conclusão? Comente.
4. Calcule a contribuição da Família A para o valor da estatística do teste. Comente.

II [8,5 valores]

Um estudo vinícola com a casta Tinta Roriz, realizado em 2015 em Vila Real, avaliou, para 30 diferentes plantas, dados de rendimento (variável `rend`, em kg/planta), bem como várias características de qualidade do mosto: sólidos solúveis (variável `brix`, em graus brix); ácido tartárico (variável `acidez`, em g/l); pH (variável `ph`); volume do mosto (variável `volume`, em ml/60 bagos); teor de antocianas (variável `ant`, em g/l); taninos totais (variável `taninos`, em g/l); e polifenóis totais (variável `polifenois`, Índice de Folin). Pretende-se modelar o teor de taninos, usando regressões lineares. Eis alguns indicadores dos valores obtidos, bem como a matriz de correlações amostrais entre as variáveis:

	ant	polifenois	taninos	ph	acidez	brix	volume	rend
minimo	1.424	217.125	5.115	3.823	3.469	19.167	67.667	0.696
média	1.98210	251.29287	6.94150	4.07583	4.06800	21.55667	95.56663	2.76983
máximo	2.400	293.106	8.140	4.223	5.025	23.367	112.333	6.138
desvio padrão	0.24219	17.92042	0.78275	0.09688	0.38738	1.11332	11.70352	1.34786

	ant	polifenois	taninos	ph	acidez	brix	volume	rend
ant	1.000	0.690	-0.266	0.421	-0.508	0.709	-0.414	-0.460
polifenois	0.690	1.000	0.486	0.302	-0.363	0.537	-0.284	-0.230
taninos	-0.266	0.486	1.000	-0.080	0.162	-0.092	0.079	0.175
ph	0.421	0.302	-0.080	1.000	-0.655	0.631	-0.340	-0.601
acidez	-0.508	-0.363	0.162	-0.655	1.000	-0.781	0.225	0.237
brix	0.709	0.537	-0.092	0.631	-0.781	1.000	-0.329	-0.463
volume	-0.414	-0.284	0.079	-0.340	0.225	-0.329	1.000	0.435
rend	-0.460	-0.230	0.175	-0.601	0.237	-0.463	0.435	1.000

1. Inicialmente consideraram-se apenas regressões lineares simples.

   (a) Diga, justificando, qual seria a melhor variável preditora, e indique a proporção de variabilidade dos valores de taninos observados que se poderia explicar com esse modelo.

   (b) Efectue um teste de ajustamento global do modelo que escolheu. Comente os resultados, tendo em conta a sua resposta na alínea anterior.

   Independentemente da qualidade do modelo escolhido, responda às seguintes alíneas:

   (c) Calcule a equação da recta de regressão correspondente ao modelo que escolheu na alínea anterior. Interprete o valor do declive dessa recta, à luz do problema em estudo.

   (d) Calcule um intervalo a 95% de confiança para o teor esperado de taninos, correspondente ao valor médio da variável preditora que escolheu. Comente o resultado.

2. Decidiu-se seguidamente ajustar um modelo de regressão linear múltipla, usando todos os preditores disponíveis, tendo sido obtidos os seguintes resultados:

```
Call: lm(formula = taninos ~ ., data = rz2015)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.756285   3.875916  -1.227   0.2327
ant         -3.936354   0.288857 -13.627 3.33e-12
polifenois   0.055241   0.003125  17.677 1.73e-14
ph           0.236726   0.690835   0.343   0.7351
acidez       0.398472   0.194830   2.045   0.0530
brix         0.149043   0.073840   2.018   0.0559
volume      -0.001627   0.003892  -0.418   0.6799
rend        -0.008977   0.044484  -0.202   0.8419
---
Residual standard error: 0.212 on 22 degrees of freedom
Multiple R-squared:  0.9443, Adjusted R-squared:  0.9266
F-statistic: 53.31 on 7 and 22 DF,  p-value: 2.432e-12
```

   (a) Comente a seguinte afirmação: "*O coeficiente estimado do preditor* `polifenois` *é quase nulo, razão pela qual esse preditor pode ser excluído sem afectar de forma significativa a qualidade de ajustamento do modelo*".

   (b) Diga, justificando sinteticamente, qual o preditor cuja exclusão do modelo menos afectaria a qualidade de ajustamento. Calcule, justificando, o valor do coeficiente de determinação do submodelo resultante dessa exclusão.

   (c) Foi ajustado um novo modelo, com apenas dois preditores: `ant` e `polifenois`. Eis alguns resultados:

```
Call: lm(formula = taninos ~ ant + polifenois, data = rz2015)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.269530   0.581852   0.463    0.647
ant         -3.706825   0.233551 -15.872 3.25e-15
polifenois   0.055789   0.003156  17.675 2.28e-16
---
Residual standard error: 0.2206 on 27 degrees of freedom
Multiple R-squared:  0.9261, Adjusted R-squared:  0.9206
F-statistic: 169.1 on 2 and 27 DF,  p-value: 5.351e-16
```

i. Teste se este modelo com apenas dois preditores difere significativamente do modelo com sete preditores referido acima. Comente.

ii. Calcule os valores de R^2 nos dois modelos de regressão linear simples correspondentes a cada um dos preditores deste modelo. Compare-os com o coeficiente de determinação do modelo de dois preditores agora ajustado e comente.

iii. Interprete geometricamente os resultados da subalínea anterior, na nuvem de $n = 30$ pontos em $\mathbb{R}^3$.

III [4,5 valores]

Num estudo sobre características de crescimento de pinheiro manso, conduzido em Sines e em Tavira, avaliou-se a altura média de pinheiros de cinco diferentes proveniências (Marrocos, Grécia, Portugal e duas proveniências de Itália), dois anos após a plantação. Quer em Sines, quer em Tavira, foram plantados seis talhões com árvores de cada proveniência, gerando assim $n = 60$ valores de alturas (variável `alt2`, em cm), cuja variância amostral é $s^2 = 34.49584$. Eis algumas médias resultantes.

```
prov                                                    local
Grecia Italia-1 Italia-2 Marrocos Portugal              Sines Tavira
 28.81    32.75    30.23    35.13    31.90              28.14  35.38

 prov:local
            local                                       Grand mean
prov        Sines  Tavira                                 31.76298
  Grecia    22.52   35.10
  Italia-1  31.03   34.46
  Italia-2  26.91   33.56
  Marrocos  31.16   39.09
  Portugal  29.09   34.70
```

1. Identifique o delineamento experimental utilizado e o modelo ANOVA adequado. Descreva pormenorizadamente o modelo.

2. Sabendo que o Quadrado Médio Residual é 16.59 e que a Soma de Quadrados associada às cinco diferentes proveniências é 280.61, construa a tabela-resumo do modelo ANOVA adequado.

3. Use um teste F para avaliar a existência de efeitos de proveniência dos pinheiros. Comente as suas conclusões. Indique brevemente que outros tipos de efeitos devem ser considerados significativos. Considere $\alpha = 0.05$.

4. Na amostra, a maior altura média em Sines é inferior à menor altura média em Tavira. Independentemente das suas respostas nas alíneas anteriores, use o teste de Tukey para indicar se igual afirmação se pode estender à população. Comente.

IV [4,5 valores]

1. Considere uma regressão linear múltipla de equação $Y_i = \beta_0 + \beta_1\, x_{1(i)} + \beta_2\, x_{2(i)} + ... + \beta_p\, x_{p(i)} + \epsilon_i$, e ajustada com base em n observações das variáveis envolvidas.

   (a) Caracterize a matriz do modelo $\mathbf{X}$. Defina o conceito de espaço das colunas, $\mathcal{C}(\mathbf{X})$.

   (b) Considere o vector centrado das n observações de Y, $\vec{\mathbf{y}}^c = (y_1 - \overline{y}\,,\, y_2 - \overline{y}\,,\, ...\,,\, y_n - \overline{y})^t$.

      i. Mostre que se verifica $\vec{\mathbf{y}}^c = (\mathbf{I} - \mathbf{P})\vec{\mathbf{y}}$, onde $\vec{\mathbf{y}}$ é o vector das n observações da variável resposta e $\mathbf{P}$ é a matriz de projecção ortogonal sobre o subespaço gerado pelo vector dos n uns, $\vec{\mathbf{1}}_n$.

      ii. Caracterize o vector resultante da projecção ortogonal de $\vec{\mathbf{y}}^c$ sobre $\mathcal{C}(\mathbf{X})$, e diga qual o conceito estatístico correspondente ao quadrado da norma desse vector projectado.

      iii. *Com base em conceitos geométricos*, mostre que $SQR \leq SQT$, verificando-se a igualdade apenas quando todos os valores ajustados de Y coincidem com os respectivos valores observados (isto é, quando $y_i = \hat{y}_i$, $\forall\, i$).

2. Considere uma regressão linear múltipla com p variáveis preditoras, ajustada com base em n observações. Considere um algoritmo de exclusão sequencial baseado no Critério de Informação de Akaike (AIC). Convencione que, num dado passo do algoritmo, em que se considere um submodelo com k variáveis preditoras, a respectiva Soma de Quadrados Residual é representada por $SQRE_k$.

   (a) Mostre que num dado passo do algoritmo, haverá exclusão dum preditor caso haja um submodelo de $k-1$ preditores para o qual se verifique $\mathrm{e}^{2/n}\, SQRE_k > SQRE_{k-1}$. Comente as implicações deste resultado na aplicação do algoritmo com amostras de grande dimensão.

   (b) Considere também a variante do algoritmo baseada nos testes-t aos parâmetros β_j. Mostre, justificando adequadamente, que partindo dum mesmo submodelo, e caso haja exclusão dum preditor nas duas variantes do algoritmo, a variável a excluir tem de ser a mesma.