INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2018-19

10 de Janeiro de 2019 **Segundo TESTE** **Duração: 2h30**

I [9,5 valores]

Um estudo sobre Macieiras Bravo Esmolfe, realizado um ano após a transplantação das árvores para o campo, visa encontrar um modelo para prever a produção total (variável `producao`, em kg/árvore), a partir de cinco potenciais variáveis preditoras: o diâmetro das árvores 10cm acima do ponto de enxertia (variável `diametro`, em cm); a altura da árvore (variável `altura`, em m); e os números de frutos contados em três datas diferentes: em Junho (variável `nfrJun`); em Setembro (variável `nfrSet`); e à colheita (variável `nfrColh`). Cada uma das $n = 149$ observações individuais corresponde à média de 15 pseudo-repetições, dadas pelas árvores de um mesmo genótipo. Seguidamente são dados alguns indicadores relativos a cada variável, bem como a matriz de correlações entre cada par de variáveis.

Variável	diametro	altura	nfrJun	nfrSet	nfrColh	producao
Mínimo	14.13	1.499	1.333	1.067	0.433	0.0450
Máximo	23.47	2.117	16.600	14.000	9.900	1.2730
Média	18.82	1.842	7.364	6.156	4.025	0.5221
Desvio Padrão	1.8569648	0.1161166	3.7105911	3.0958839	2.3597379	0.2975462

```
> cor(macieira2)
          diametro    altura    nfrJun    nfrSet   nfrColh  producao
diametro 1.0000000 0.7207190 0.5783022 0.5950834 0.5834737 0.5936943
altura   0.7207190 1.0000000 0.4343267 0.4310161 0.3950239 0.3863919
nfrJun   0.5783022 0.4343267 1.0000000 0.9895452 0.9554149 0.9444425
nfrSet   0.5950834 0.4310161 0.9895452 1.0000000 0.9716697 0.9616189
nfrColh  0.5834737 0.3950239 0.9554149 0.9716697 1.0000000 0.9853879
producao 0.5936943 0.3863919 0.9444425 0.9616189 0.9853879 1.0000000
```

Decidiu-se ajustar um modelo de Regressão Linear Múltipla, que permitisse prever a produção antes da data da colheita. Eis os resultados de um modelo de regressão linear com os quatro preditores disponíveis no final de Setembro:

```
Call:
lm(formula = producao ~ diametro + altura + nfrJun + nfrSet , data = macieira2)
[...]
Coefficients:
             Estimate  Std.Error t value Pr(>|t|)
(Intercept)  0.086704   0.109663   0.791   0.4305
diametro     0.014061   0.005772   2.436   0.0161
altura      -0.210009   0.082022  -2.560   0.0115
nfrJun      -0.021418   0.012389  -1.729   0.0860
nfrSet       0.116200   0.015066   7.713      ???
---
Residual standard error: 0.07935 on ??? degrees of freedom
Multiple R-squared: 0.9308,  Adjusted R-squared: 0.9289
F-statistic: ??? on ?? and ??? DF,  p-value: < 2.2e-16
```

1. Discuta pormenorizadamente o ajustamento global deste modelo.

2. Interprete o valor estimado do coeficiente da variável `diametro` e construa um intervalo a 95% de confiança para o correspondente valor populacional.

3. Compare, através de um Teste de Hipóteses adequado, o modelo completo ajustado neste ponto e o submodelo com apenas os preditores `diametro` e `nfrJun`, ao qual corresponde um Coeficiente de Determinação de valor $R^2 = 0.8954$. Comente.

4. Partindo do modelo de quatro preditores, ajustado inicialmente, considere um algoritmo de exclusão sequencial baseado em testes t, ao nível de significância $\alpha = 0.05$.

   (a) Identifique o submodelo de três preditores resultante do primeiro passo do algoritmo. Comente, tendo em conta o coeficiente de correlação entre `producao` e a variável que excluiu.

   (b) Calcule o Coeficiente de Determinação do submodelo de três preditores que escolheu.

5. Considere a seguinte afirmação: "*O valor do Critério de Informação de Akaike (AIC) do modelo inicial de quatro preditores é maior do que o valor do AIC do modelo de regressão linear simples cujo único preditor é o número de frutos à data da colheita* (`nfrColh`)". Comente a afirmação sem calcular o valor dos AICs e diga qual o modelo preferido, com base nos AICs.

II [6 valores]

Pretende-se estudar os rendimentos dum dado genótipo da casta Tinta Francisca, em dois diferentes locais: Régua e Tabuaço. É sabido que os rendimentos podem variar muito em diferentes anos e decidiu-se levar esse aspecto em conta ao delinear a experiência. O ensaio foi realizado na Régua nos anos 1999, 2000, 2002, 2003 e 2004. No entanto, limitações de recursos impuseram que no Tabuaço a experiência só pudesse ser realizada em 1999 e 2003. Em cada situação experimental, foram medidos os rendimentos (variável `rend`, em kg/planta), em 8 parcelas. Registou-se, para a totalidade dos rendimentos medidos, um valor médio de 0.924375 $kg/planta$, e uma variância de 0.329922 $(kg/planta)^2$. Para cada situação experimental registaram-se os seguintes rendimentos médios:

Regua1999	Regua2000	Regua2002	Regua2003	Regua2004	Tabuaco1999	Tabuaco2003
0.291	0.917	1.327	1.682	0.743	0.725	0.786

1. Identifique o delineamento experimental usado no ensaio e descreva em pormenor o modelo ANOVA correspondente. Usando os parâmetros da equação do modelo, diga qual o rendimento esperado no Tabuaço em 1999, e no Tabuaço em 2003.

2. Sabendo que a Soma de Quadrados Residual é 8.311 e que a Soma de Quadrados associada à variabilidade entre o Tabuaço e a Régua é 0.6402, construa o quadro de síntese do modelo ANOVA, indicando como obtém cada um dos restantes valores.

3. Justifique, através dum teste adequado, se foi importante ter efectuado o ensaio em anos diferentes.

4. Justifique brevemente, com base num teste F, se é possível afirmar que os rendimentos médios populacionais na Régua e no Tabuaço são diferentes.

5. Utilize os testes de Tukey para dizer se é possível afirmar que o rendimento médio obtido na Régua, em 1999, é significativamente diferente dos rendimentos médios em todas as restantes situações experimentais. Comente o seu resultado.

6. Construa a tabela-resumo que resultaria de ajustar aos dados um modelo ANOVA que apenas prevesse a existência de diferentes locais, e tratasse as observações em anos diferentes como meras repetições. Quais seriam os resultados do(s) teste(s) F comparável(is)? Comente.

III [4,5 valores]

1. Seja dado um Modelo de Regressão Linear Múltipla de p variáveis preditoras, ajustado com base em n observações.

   (a) Descreva os pressupostos do Modelo, usando a notação vectorial/matricial e indicando a natureza de todas as quantidades que referir.

   (b) Mostre que, dado o Modelo, o vector de estimadores verifica $\vec{\hat{\boldsymbol{\beta}}} \cap \mathcal{N}_{p+1}\left(\vec{\boldsymbol{\beta}}\,,\, \sigma^2\,(\mathbf{X}^t\mathbf{X})^{-1}\right)$.

   (c) Considere o Coeficiente de Determinação modificado, R^2_{mod}.

      i. Mostre que $R^2_{mod} = 1-(1-R^2)\,\frac{n-1}{n-(p+1)}$, sendo R^2 o Coeficiente de Determinação usual.

      ii. Mostre que R^2_{mod} pode tomar valores entre $\frac{-p}{n-(p+1)}$ e 1.

      iii. Justifique a seguinte afirmação e comente as suas implicações: "*Se R^2_{mod} toma valores negativos, a vâriancia estimada das observações da variável resposta Y em torno da hipersuperfície de regressão é maior do que a variância amostral de Y ignorando a regressão sobre os preditores*".

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

   (a) Indique, justificando, qual o número total de parâmetros de cada modelo.

   (b) Justifique que só é possível ajustar o modelo com efeitos de interacção caso haja repetições nas células.

   (c) Justifique que o espaço das colunas da matriz do modelo M_{A+B} está contido no espaço das colunas da matriz do modelo M_{A*B}. Deduza daí que a Soma de Quadrados Residual do modelo M_{A+B} não pode ser menor que a Soma de Quadrados Residual do modelo M_{A*B}.