INSTITUTO SUPERIOR DE AGRONOMIA
MATEMÁTICA E ESTATÍSTICA – 2009/10

1 de Fevereiro, 2010 **SEGUNDA CHAMADA DE EXAME** **Duração:** 1h45

(Parte de Estatística)

I

A relação entre quantidade de precipitação num dia (em *mm*) e quantidade de água que escorre pelo tronco duma árvore (escorrimento ao longo do tronco, em *ml*) é de interesse no estudo da intercepção da precipitação pelo coberto vegetal. No âmbito dum estudo destas relações, foi medida a precipitação em 28 dias de chuva, e os correspondentes escorrimentos para três árvores (numa mesma parcela experimental) da espécie *Liriodendron tulipifera*. Experiência prévia destes estudos sugere que as relações são melhor estudadas quando se procede à logaritmização das variáveis. Eis alguns dos dados logaritmizados obtidos em cada dia, bem como as respectivas médias e variâncias globais:

dia	ln(precipitação)	ln(escorrimento)		
		Árvore 1	Árvore 2	Árvore 3
1	2.4423	6.9613	6.4069	9.1407
2	2.7663	7.7446	6.0307	9.3384
3	2.0412	5.3375	5.1930	6.0426
⋮	⋮	⋮	⋮	⋮
27	1.1632	3.9512	3.4965	2.9444
28	1.9459	1.6094	4.9200	4.3820

Para as 84 observações de cada variável logaritmizada, tem-se:

	ln-precipit.	ln-escorrim.
Média	2.213738	6.454755
Variância	0.866739	7.3268

Foi ajustada uma regressão linear simples de logaritmos do escorrimento (variável `ln.esc`) sobre logaritmos da precipitação (`ln.prec`), utilizando as 84 observações, com os seguintes resultados:

```
>  summary(lm(ln.esc ~ ln.prec, data=dadosEx2))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9971     0.4084   2.442   0.0168
ln.prec       2.4654     0.1702  14.485   <2e-16
---
Residual standard error: 1.444 on 82 degrees of freedom
Multiple R-squared: 0.719,Adjusted R-squared: 0.7156
F-statistic: ??? on 1 and 82 DF,  p-value: < 2.2e-16
```

1. Qual o valor da estatística F no teste de ajustamento global do modelo?
2. Qual o coeficiente de correlação entre as variáveis (logaritmizadas) envolvidas nesta regressão? Justifique, em particular, o sinal dessa correlação.
3. Determine um intervalo a 95% de confiança para o valor esperado do log-escorrimento, associado aos dias em que se verifiquem 10 *mm* de precipitação.
4. Com base na referida regressão linear simples, diga qual a equação ajustada para a relação entre as variáveis precipitação e escorrimento (não logaritmizadas)?
5. Será admissível afirmar que o escorrimento (em *ml*) é proporcional ao quadrado da precipitação (em *mm*)? Responda utilizando um procedimento estatístico adequado.

II

Num estudo sobre uma população experimental de clones da casta Tinta Francisca, realizado em 2003 em Tabuaço, foram medidos os valores das seguintes variáveis em 24 videiras: teor de antocianas (em mg/dm^3, variável `antoci`); teor de açúcar (variável `acucar`, em g/l); pH (variável `pH`); acidez total (em g de ácido tartárico por dm^3, variável `actot`); e fenois totais (variável `fentot`). Uma característica muito importante de algumas castas é a côr do bago, que está associada à pigmentação da película do bago, e portanto ao teor em antocianas. Pretende-se modelar esta variável, à custa das restantes variáveis observadas. Foram obtidos os seguintes resultados.

```
Call: lm(formula = antoci ~ acucar + pH + actot + fentot, data = Ex2RM)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  908.6751   679.2284   1.338   0.1968
acucar         0.9885     0.7308   1.353   0.1920
pH          -234.0446   123.6247  -1.893   0.0737
actot        -17.3130    50.0359  -0.346   0.7331
fentot        20.8702     2.5662   8.133 1.31e-07
---
Residual standard error: 45.12 on 19 degrees of freedom
Multiple R-squared: 0.8225,Adjusted R-squared: 0.7851
F-statistic: 22.01 on 4 and 19 DF,  p-value: 6.501e-07
```

1. Comente a qualidade do modelo ajustado, efectuando um teste de hipóteses apropriado e discutindo o valor do coeficiente de determinação.

2. Comente a diferença entre os valores indicados para o R^2 e para o R^2 modificado (*adjusted*).

3. Construa um intervalo de confiança para a alteração esperada no teor de antocianas, associado a um aumento de $1\,g/l$ no teor de açúcar.

4. Foi efectuado um algoritmo de exclusão sequencial com o objectivo de estudar a possível simplificação do modelo. A listagem de resultados produzidos pelo programa `R` é a seguinte:

```
> step(lm(antoci ~ acucar + pH + actot + fentot, data=Ex2RM))
Start:  AIC=187.24
antoci ~ acucar + pH + actot + fentot
         Df Sum of Sq    RSS    AIC      NOTA: ``RSS'' são as
- actot   1       244  38920    185        iniciais em inglês
<none>                 38676    187        da Soma de Quadrados
- acucar  1      3724  42401    187        Residual.
- pH      1      7296  45972    189
- fentot  1    134631 173308    221

Step:  AIC=185.39
antoci ~ acucar + pH + fentot
         Df Sum of Sq    RSS    AIC
<none>                 38920    185
- acucar  1      4220  43140    186
- pH      1      9209  48129    188
- fentot  1    148138 187057    221
```

O submodelo seleccionado tem apenas três variáveis preditoras: `acucar`, `pH` e `fentot`.

(a) Com base na informação disponível, determine o valor do coeficiente de determinação associado ao submodelo seleccionado.

(b) A partir deste submodelo, seleccionado pelo programa `R`, qual a variável cuja exclusão menos afecta a qualidade do ajustamento? Justifique.

(c) Efectue um teste F parcial comparando o submodelo de dois preditores que escolheu na alínea anterior, com o submodelo de três preditores escolhido pelo programa `R`, e comente as conclusões deste teste, para um nível de significância 0.05.

III

Considere um modelo ANOVA, a um factor com $k = 3$ níveis. Admita que o modelo é equilibrado, e ajustado com n_c observações em cada nível.

1. Descreva a matriz $\mathbf{X}$ do delineamento.

2. Mostre que o subespaço das colunas desta matriz, $\mathcal{C}(\mathbf{X})$, é igual ao subespaço das colunas da matriz $\mathbf{Y}$, sendo esta última matriz constituída pelas três variáveis indicatrizes dos três níveis do factor, $\mathbf{I}_1$, $\mathbf{I}_2$ e $\mathbf{I}_3$.

3. Tendo em conta a alínea anterior, mostre que a matriz de projecção ortogonal sobre o subespaço das colunas de $\mathbf{X}$ é uma matriz da forma:

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{n_c} & \mathbf{0}_{n_c} & \mathbf{0}_{n_c} \\ \mathbf{0}_{n_c} & \mathbf{A}_{n_c} & \mathbf{0}_{n_c} \\ \mathbf{0}_{n_c} & \mathbf{0}_{n_c} & \mathbf{A}_{n_c} \end{bmatrix} ,$$

   sendo $\mathbf{A}_{n_c}$ uma matriz $n_c \times n_c$ cujos valores são todos iguais a $\frac{1}{n_c}$, e $\mathbf{0}_{n_c}$ uma matriz $n_c \times n_c$ cujos valores são todos iguais a zero.

4. A partir do resultado da alínea anterior, mostre que o valor ajustado de qualquer das observações do nível i é sempre igual à média amostral das observações nesse nível, ou seja, que $\hat{y}_{ij} = \overline{y}_{i.}, \forall\, i$.