INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2017-18

29 de Janeiro de 2018 Segunda Chamada de EXAME Duração: 3h30

I [2,5 valores]

Um ensaio realizado em Elvas visou avaliar o nível de ataque da mosca da azeitona num olival, ao longo de várias fases do ciclo vegetativo das oliveiras. Em cada uma de quatro fases, foram dispostas 24 armadilhas igualmente espaçadas no interior do olival que, após um igual período de exposição foram retiradas, sendo contadas as moscas capturadas. As contagens foram agrupadas em classes de infestação `A` (0 a 2 moscas capturadas), `B` (3 ou 4 moscas), `C` (5, 6 ou 7 moscas) e `D` (8 ou mais moscas). Os resultados obtidos são indicados na tabela. Pretende-se saber se é admissível considerar que a distribuição das armadilhas pelas quatro classes de infestação é idêntica nas quatro fases do ciclo vegetativo.

	Classes de infestação			
	A	B	C	D
Fase 1	7	4	5	8
Fase 2	6	4	8	6
Fase 3	2	5	4	13
Fase 4	4	8	8	4

1. Diga, justificando, qual o teste de hipóteses adequado, explicitando as hipóteses em confronto, a estatística de teste e a sua distribuição assintótica, bem como a natureza da região crítica.
2. Considera que a amostra é suficientemente grande para se poder admitir a validade da distribuição assintótica da estatística do teste? Justifique.
3. Independentemente da sua resposta na alínea anterior, diga qual a conclusão do teste admitindo a validade da distribuição assintótica e sabendo que o valor da estatística calculada é 12.967. Comente.
4. Calcule a parcela que lhe parecer mais elevada na estatística do teste. Comente, tendo em conta o problema sob estudo.

II [8 valores]

Um estudo realizado no Alentejo visou modelar o índice de área foliar em oliveiras (variável `IAF`, em $m^2\, m^{-2}$). Dispõe-se de dados relativos a 30 árvores, nas quais, além do índice de área foliar, foram medidas cinco outras variáveis que podem ser usadas como preditores: o perímetro do tronco a $50\, cm$ de altura (variável `P`, em cm); a altura do tronco (variável `HS`, em cm); a altura da árvore (variável `HT`, em m); o diâmetro da copa (variável `D`, em m); e o volume da copa (variável `V`, em m^3). Eis alguns indicadores, bem como a matriz de correlações, relativas à totalidade das observações:

	IAF	P	HS	HT	D	V
Mínimo	2.310	24.00	55.00	2.850	2.400	8.47
Média	4.928	43.57	72.03	3.489	3.323	20.65
Máximo	9.760	78.00	97.00	4.650	5.400	52.56
Variância	3.8246	342.8057	171.2747	0.2617	1.0770	209.7101

```
       IAF       P      HS      HT       D       V
IAF  1.0000 -0.6798 -0.4328 -0.6088 -0.7746 -0.7247
P   -0.6798  1.0000  0.4328  0.8563  0.9608  0.9349
HS  -0.4328  0.4328  1.0000  0.4659  0.4112  0.3807
HT  -0.6088  0.8563  0.4659  1.0000  0.8721  0.9195
D   -0.7746  0.9608  0.4112  0.8721  1.0000  0.9831
V   -0.7247  0.9349  0.3807  0.9195  0.9831  1.0000
```

1. Com base na informação disponível até aqui, responda às seguintes questões:
   (a) Comente a afirmação "*a oliveiras de maiores dimensões correspondem, em geral, menores índices de área foliar*".
   (b) Foi ajustada uma regressão linear simples do índice de área foliar `IAF` sobre o diâmetro da copa `D`.
      i. Qual o valor do Coeficiente de Determinação? Interprete o valor obtido.
      ii. Qual a equação da recta ajustada?
      iii. Construa um intervalo a 95% de confiança para o valor esperado do índice de área foliar, associado a oliveiras cujo diâmetro da copa é 4 m.

2. Foi decidido ajustar uma regressão linear simples do logaritmo natural do índice de área foliar sobre o logaritmo natural do diâmetro, tendo sido obtidos os seguintes resultados.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9442     0.1582  18.614  < 2e-16
log(D)       -1.2314     0.1326  -9.285 4.82e-10
---
Residual standard error: 0.205 on 28 degrees of freedom
Multiple R-squared:  0.7548,   Adjusted R-squared:  0.7461
F-statistic: 86.21 on 1 and 28 DF,  p-value: 4.822e-10
```

   (a) Qual a curva ajustada da relação de fundo ajustada entre o índice `IAF` e o diâmetro `D`?
   (b) Teste formalmente se é admissível afirmar que o índice de área foliar é inversamente proporcional ao diâmetro da copa.
   (c) Sabendo que a soma de quadrados das distâncias entre os `IAF` observados e os correspondentes valores de `IAF` ajustados pela curva da alínea II.2a é 42.0678, comente a seguinte afirmação: "*este modelo ajusta-se melhor do que o modelo de regressão linear entre `IAF` e `D` (sem transformação logarítmica)*".

3. Foi entretanto sugerido que o desempenho do submodelo ganharia com a introdução de novos preditores. Foi assim ajustado um modelo de regressão linear múltipla para modelar os logaritmos (naturais) do `IAF`, à custa dos logaritmos dos cinco preditores disponíveis. Eis os resultados obtidos:

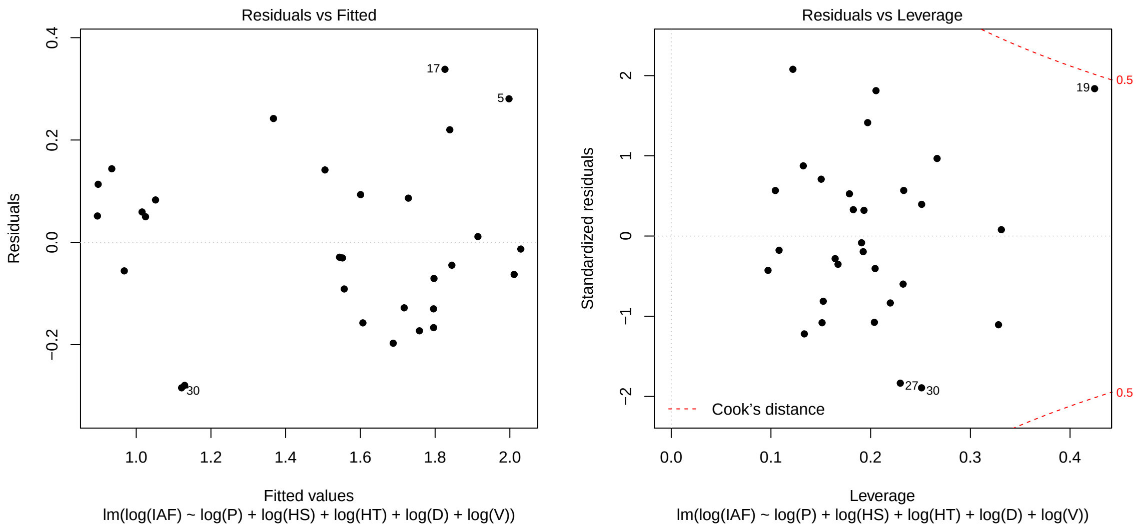
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3467     1.4810   0.234  0.81690
log(P)        0.7563     0.2619   2.888  0.00809
log(HS)       0.5898     0.5056   1.167  0.25487
log(HT)      -3.3226     1.8009  -1.845  0.07741
log(D)       -6.8870     2.2846  -3.014  0.00600
log(V)        2.8149     1.3347   2.109  0.04557
```

```
---
Residual standard error: 0.1736 on 24 degrees of freedom
Multiple R-squared:  0.8493,   Adjusted R-squared:  0.8179
F-statistic: 27.04 on 5 and 24 DF,  p-value: 3.987e-09
```

(a) Comente os valores dos R^2 usual e modificado.

(b) Teste formalmente se este modelo se ajusta melhor do que o modelo de regressão linear simples do ponto II.2.

(c) Descreva e discuta os seguintes gráficos relativos ao modelo agora ajustado. Em particular, diga se há observações influentes e/ou com um efeito alavanca importante.



III [5 valores]

Um ensaio com a casta Alfrocheiro visou estudar o comportamento dum genótipo, referenciado com o código `AF1810`, em três localidades: Mangualde, Nelas e Vidigueira. Pretendia-se medir o rendimento das videiras (em kg/planta). A fim de controlar a variabilidade, o ensaio foi repetido, de forma equilibrada, em diferentes anos, mas dificuldades logísticas e financeiras obrigaram a que a experiência fosse realizada em anos (quase) sempre diferentes nas várias localidades. Em cada localidade e ano foram observadas 8 repetições. Os valores médios obtidos (global, por localidade e por localidade e ano) são indicados na tabela seguinte. Sabe-se ainda que a variância da totalidade das 88 observações é $s_y^2 = 1.774264$.

```
  Grand mean                  local
    2.457648                  Mangualde  Nelas Vidigueira
                                  1.972  2.187      4.213

   local:ano
              ano
local      A1994 A1996 A1997 A1999 A2000 A2004 A2005 A2006 A2010 A2002
Mangualde  0.677 2.175 2.043 2.683 2.283
Nelas                                    1.557 1.245 2.402 3.542
Vidigueira                               3.645                   4.782
```

Foi ajustado uma ANOVA, a que corresponde a seguinte tabela-resumo:

```
            Df  Sum Sq  Mean Sq  F value     Pr(>F)
local       ??    ??      ??       ??      3.01e-15
local:ano   ??  49.11   6.139    10.71     5.44e-10
Residuals   ??  44.16   0.573
```

1. Diga, justificando, a que tipo de delineamento experimental corresponde a experiência realizada. Descreva pormenorizadamente o modelo ANOVA apropriado.

2. Calcule os seis valores omissos na tabela, indicando como os obtém.

3. Quais os tipos de efeitos que devem ser considerados significativos ao nível $\alpha = 0.05$? Responda, descrevendo em pormenor todos os testes que realizar.

4. Quais os rendimentos médios que diferem significativamente do rendimento obtido na Vidigueira em 2002? Justifique através dum teste adequado ($\alpha = 0.05$).

IV [4,5 valores]

Considere um modelo de regressão linear simples, ajustado com n pares de observações $\{(x_i, y_i)\}_{i=1}^n$.

1. Mostre que a recta de regressão, ajustada pelo método de mínimos quadrados, passa no centro de gravidade $(\overline{x}, \overline{y})$ da nuvem de n pontos em $\mathbb{R}^2$.

2. Considere uma *qualquer recta*, de equação $y = a + m\,x$. Mostre que exigir que seja nula a soma das n diferenças entre os valores observados y_i e os correspondentes valores de y ajustados pela recta $(a + m\,x_i)$, obriga a que a recta passe no centro de gravidade da nuvem de pontos. Discuta porque essa exigência não é suficiente para garantir uma boa aproximação à nuvem de pontos.

3. Qual o menor valor de R^2 para o qual rejeitaria a hipótese nula do teste de ajustamento global, ao nível de significância $\alpha = 0.05$, se $n = 10$? E se $n = 62$? E para amostras de muito grande dimensão? Comente.

4. Considere o vector dos estimadores dos parâmetros da recta, $\vec{\hat{\boldsymbol{\beta}}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$.

   (a) Diga, justificando, quais os elementos da matriz de (co-)variâncias de $\vec{\hat{\boldsymbol{\beta}}}$.

   (b) *Utilizando as ferramentas dadas no estudo das regressões lineares múltiplas* deduza, justificando, qual a distribuição de probabilidades de $\hat{\mu}_{Y|x} = \hat{\beta}_0 + \hat{\beta}_1\,x$, o estimador do valor esperado de Y dado $X = x$.