INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2017-18

11 de Janeiro de 2018 Primeira 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. Foram dispostas 24 armadilhas igualmente espaçadas no interior do olival. Numa data pré-fixada foram recolhidas as armadilhas e contadas as moscas capturadas, tendo sido obtidos os resultados indicados em baixo. Pretende-se saber se é admissível considerar que o número de moscas por armadilha segue uma distribuição de Poisson.

No. moscas	0	1	2	3	4	5	6	7
No. armadilhas	8	6	2	6	1	0	0	1

1. Calcule o valor do parâmetro que torna a distribuição de Poisson mais verosímil.
2. Verifique as condições de Cochran, e discuta o seu papel no teste que irá realizar.
   [**Nota:** se não resolveu a alínea anterior, use o valor $\hat{\lambda}=1.6$.]
3. Independentemente da sua resposta na alínea anterior, um investigador decidiu agrupar as cinco últimas classes da tabela, e realizar um teste apropriado (ao nível $\alpha=0.01$) para avaliar o ajustamento da distribuição de Poisson, tendo obtido um valor calculado da estatística de Pearson de $X^2_{calc}=6.8229$. Formule o teste em detalhe e discuta as suas conclusões, admitindo a validade do critério de Cochran.

II [8 valores]

1. Um estudo realizado por uma equipa do ISA visou caracterizar a relação existente entre um índice de vegetação, calculado com base em medições dum aparelho portátil, e a Produtividade Primária Bruta (PPB), medida em micromoles por metro quadrado, por segundo (μmole m^{-2} s^{-1}) em comunidades herbáceas mediterrânicas de Portugal. O índice de vegetação usado é o índice NDWI, um índice adimensional que toma valores entre -1 e 1 (e que é definido com base na reflectância nas bandas do verde e do infra-vermelho próximo). Recolheram-se 91 pares de observações, com os seguintes indicadores:

Variável	Mínimo	Máximo	Média	Variância
NDWI	-0.18446	0.19154	0.03286	0.007910756
PPB	7.173	33.966	19.715	54.4267635

   Após alguma análise, optou-se por ajustar uma regressão linear simples do logaritmo (natural) da Produtividade Primária Bruta sobre os valores do índice NDWI, com os seguintes resultados:

```
> summary(lm(log(ppb) ~ ndwi, data=gpp))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.77400    0.02872    96.6   <2e-16
ndwi         3.83488    0.30432    12.6   <2e-16
---
Residual standard error: 0.2568 on 89 degrees of freedom
Multiple R-squared: 0.6408,  Adjusted R-squared: 0.6368
F-statistic: 158.8 on 1 and 89 DF,  p-value: < 2.2e-16          AIC=-245.46
```

(a) Dado o modelo ajustado, será admissível considerar que, a um aumento de uma unidade no índice NDWI corresponde, em média, um aumento de 4 unidades na log-Produtividade Primária Bruta? Justifique através dum teste de hipóteses adequado.

(b) Calcule o intervalo a 95% de confiança para a ordenada na origem da recta populacional. Interprete o resultado em termos da Produtividade Primária Bruta (em μmole $m^{-2}\ s^{-1}$).

(c) Um intervalo de predição (95%) para uma observação individual de log-PPB, quando o índice NDWI toma o valor 0.1, é da forma $]2.64287, ???[$. Diga justificando,

i. qual o valor central desse intervalo de predição;

ii. qual o extremo direito do intervalo de predição;

iii. qual o erro padrão do estimador do valor esperado de log-PPB, quando o índice NDWI é 0.1, ou seja, o erro padrão de $\hat{\mu}_{Y|X=0.1}$.

(d) A que tipo de relação não linear entre a Produtividade Primária Bruta e o índice NDWI corresponde a regressão linear acima ajustada? Calcule a equação da curva ajustada, relacionando PPB e NDWI. Qual o valor estimado da taxa de variação relativa da Produtividade Primária Bruta, face aos valores de NDWI?

(e) É possível ajustar um modelo potência para relacionar PPB e NDWI, com base numa regressão linear simples? Justifique a sua resposta.

2. O aparelho portátil que mediu os $n = 91$ valores acima considerados mediu simultaneamente a reflectância para diferentes regiões do espectro electromagnético. Dispõe-se assim de valores da reflectância em 10 bandas, correspondentes às usadas no sensor MSI do satélite europeu Sentinel-2: as bandas 2, 3, 4 e 8 na resolução espacial $10m$ e as bandas 5, 6, 7, 8, 11 e 12 na resolução espacial de $20m$. Eis o resultado duma regressão linear múltipla de `log(PPB)` sobre estas 10 variáveis:

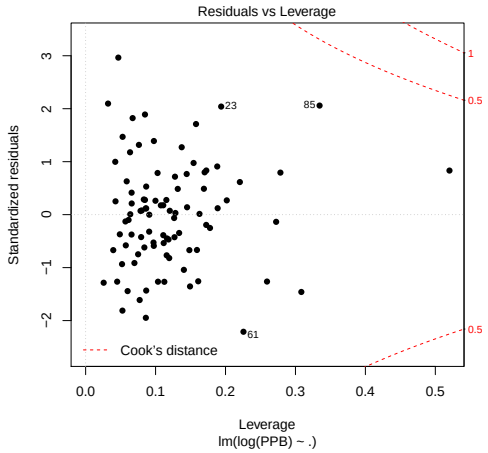
```
Residual standard error: 0.2367 on 80 degrees of freedom
Multiple R-squared: 0.7257,  Adjusted R-squared: 0.6914
F-statistic: 21.16 on 10 and 80 DF,  p-value: < 2.2e-16         AIC= ???
```

(a) Discuta a qualidade de ajustamento do modelo.

(b) Calcule o valor do Critério de Informação de Akaike (AIC) e use-o para comparar o modelo agora ajustado com o modelo de regressão linear simples ajustado no ponto anterior.

(c) Discuta, justificando, a afirmação: "*Um teste F parcial permite testar a hipótese de o ajustamento desta regressão linear múltipla ser significativamente melhor que o da regressão linear simples do ponto anterior*".

(d) Comente o seguinte gráfico, indicando a sua natureza e principais conclusões. Em particular, calcule um valor aproximado da distância de Cook da observação mais à direita.



(e) Foi utilizado um algoritmo de selecção que produziu o seguinte submodelo, em que os nomes dos preditores começam pela letra `B`, seguida do número de banda, da letra `s` e finalmente da resolução espacial (`10` ou `20` metros). Teste ($\alpha = 0.05$) se este submodelo difere significativamente do modelo completo com 10 preditores. Comente.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0212     0.2569  11.762  < 2e-16 ***
B8s10        23.9896    11.9434   2.009 0.047791 *
B5s20       -13.9554     4.3528  -3.206 0.001904 **
B6s20        27.2611    10.6493   2.560 0.012260 *
B7s20       -41.2228    18.1158  -2.276 0.025421 *
B11s20      -18.7217     5.4057  -3.463 0.000842 ***
B12s20       24.9354     9.3553   2.665 0.009220 **
---
Residual standard error: 0.2413 on 84 degrees of freedom
Multiple R-squared: 0.7007,  Adjusted R-squared: 0.6793
F-statistic: 32.77 on 6 and 84 DF,  p-value: < 2.2e-16
```

III [5 valores]

No âmbito dum estudo sobre o teor de amido em abóboras, a Secção de Horticultura do ISA realizou um ensaio em Casével, no distrito de Santarém. Nesse ensaio cruzaram-se quatro diferentes tratamentos de produção (variável `tratamento`) com quatro datas de colheita (variável `data`). Os tratamentos dizem respeito à utilização dum maior nível de cálcio (situação referenciada pela letra `A`); dum menor teor de cálcio (`B`); de condições de *stress* hídrico (`C`); ou condições de rega normal (`D`). As quatro datas de colheita ensaiadas foram 13 de Setembro (referenciada por `Set1`); 29 de Setembro (`Set2`); 11 de Outubro (`Out`); e 23 de Novembro (`Nov`). Foi medido o teor de amido na matéria fresca (variável `amido`, em g por $100g$ de abóbora). Para cada tratamento e data observaram-se três parcelas, tendo sido obtidos os seguintes valores médios global, por data, por tratamento, e por cruzamento de cada tratamento e data. A variância amostral da totalidade das observações é 0.8238201.

```
Grand mean          data                            tratamento
1.390549         Nov    Out   Set1   Set2             A      B      C      D
              0.3205 1.1439 2.1549 1.9429       1.5939 1.2162 1.2567 1.4954
```

```
      tratamento
data    A      B      C      D
Nov  0.2751 0.3839 0.2813 0.3419
Out  1.9299 1.1968 0.7277 0.7213
Set1 1.9696 1.8131 2.2853 2.5514
Set2 2.2008 1.4711 1.7325 2.3670
```

Foi ajustado um modelo ANOVA, com os seguintes resultados:

```
                 Df Sum Sq Mean Sq F value   Pr(>F)
data             ??    ??     ??    33.015 6.48e-10
tratamento       ??  1.208    ??     1.586   0.2120
data:tratamento  ??  4.250   0.472     ??    0.0951
Residuals        ??  8.122   0.254
```

1. Diga, justificando, a que tipo de delineamento experimental corresponde a experiência realizada. Descreva pormenorizadamente o modelo ANOVA apropriado.
2. Calcule os oito valores omissos na tabela, indicando como os obtém.
3. Um utilizador afirma que bastava considerar as datas como único factor a afectar o teor de amido na matéria fresca. Comente esta afirmação, utilizando testes de hipóteses adequados. Descreva pormenorizadamente pelo menos um dos testes que tiver de efectuar.
4. Independentemente da sua resposta na alínea anterior, é possível afirmar que, ao nível $\alpha = 0.05$, o teor de amido é significativamente maior na primeira data de Setembro, com o tratamento `D`, do que em qualquer outra situação experimental? Justifique com base num teste apropriado.
5. Qual o valor estimado para o parâmetro β_2? Interprete o significado dessa estimativa no contexto do problema.

IV [4,5 valores]

1. Considere uma regressão linear múltipla, onde se relaciona a variável resposta Y com p preditores, e que é ajustada com base em n observações das variáveis envolvidas.
   (a) Descreva a matriz do modelo, $\mathbf{X}$, e defina o conceito de subespaço das colunas de $\mathbf{X}$, $\mathcal{C}(\mathbf{X})$.
   (b) Mostre que a matriz $\mathbf{H}$ de projecção ortogonal sobre o subespaço das colunas de $\mathbf{X}$ é simétrica e idempotente.
   (c) Mostre que a Soma de Quadrados Residual se pode escrever como $SQRE = \vec{\mathbf{Y}}^t(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}$, onde $\vec{\mathbf{Y}}$ é o vector das n observações de Y e $\mathbf{I}_n$ é a matriz identidade $n \times n$.
2. Considere um delineamento experimental hierarquizado com dois factores: um factor dominante A com a níveis, e um factor subordinado B, com b_i níveis para cada nível $i = 1, 2, ..., a$ do factor dominante. Designe o Quadrado Médio Residual deste modelo por $QMRE_{A/B}$ e a estatística no teste à existência de efeitos do factor subordinado B por $F_{B(A)}$.
   (a) Admita que às mesmas n observações com que ajustou o modelo anterior foi agora ajustado um modelo apenas com o Factor A, de a níveis. Designe o respectivo Quadrado Médio Residual por $QMRE_A$. Mostre que $QMRE_A < QMRE_{A/B}$ se e só se $F_{B(A)} < 1$.
   (b) Se $F_{B(A)} < 1$, o que pode afirmar sobre os valores das estatísticas dos testes aos efeitos do factor A nos dois modelos acima considerados?
   (c) Comente as implicações dos resultados das alíneas anteriores.