INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2017-18

11 de Janeiro de 2018 **Segundo TESTE** **Duração: 2h30**

I [8,5 valores]

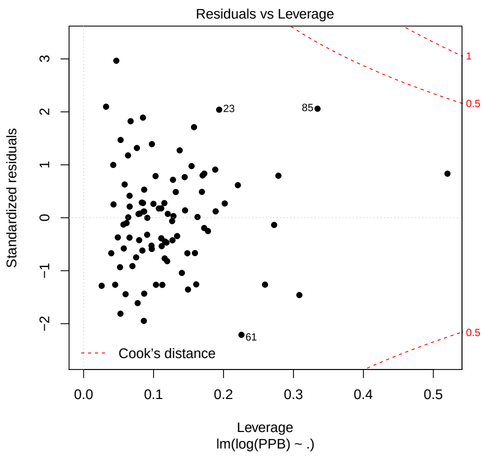
Um estudo realizado por uma equipa do ISA visou modelar o índice de Produtividade Primária Bruta (PPB), medida em micromoles por metro quadrado, por segundo (μ mole m^{-2} s^{-1}) em comunidades herbáceas mediterrânicas de Portugal. Foram medidos $n = 91$ valores da variável `PPB` e, simultaneamente, com um aparelho portátil, 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 do logaritmo (natural) de PPB sobre estas 10 variáveis:

```
Residual standard error: 0.2367 on 80 degrees of freedom
Multiple R-squared: 0.7257,  Adjusted R-squared: ???
F-statistic: 21.16 on 10 and 80 DF,  p-value: < 2.2e-16        AIC=-251.98
```

1. Discuta a qualidade de ajustamento do modelo.

2. Calcule o valor do R^2 modificado. Comente o valor obtido.

3. 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.



4. Foi utilizado um algoritmo de selecção que produziu o seguinte submodelo. Nos resultados, 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). É também dada a matriz de (co-)variâncias estimadas dos estimadores dos sete parâmetros do submodelo.

```
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       AIC= ???

> vcov(gppsoB.lm3)
            (Intercept)     B8s10    B5s20     B6s20     B7s20   B11s20    B12s20
(Intercept)      0.0660   -1.6206   0.0521   -0.3801    1.8874  -0.1210    0.0465
B8s10           -1.6206  142.6454 -34.8016   79.0325 -203.0525 -22.7008   47.4785
B5s20            0.0521  -34.8016  18.9467  -38.9915   63.9155  14.5439  -30.9054
B6s20           -0.3801   79.0325 -38.9915  113.4074 -164.5279 -43.9221   80.1549
B7s20            1.8874 -203.0525  63.9155 -164.5279  328.1829  54.6572 -105.7394
B11s20          -0.1210  -22.7008  14.5439  -43.9221   54.6572  29.2211  -48.8892
B12s20           0.0465   47.4785 -30.9054   80.1549 -105.7394 -48.8892   87.5207
```

(a) Teste ($\alpha = 0.05$) se este submodelo difere significativamente do modelo completo inicial. Comente.

(b) Calcule o valor do Critério de Informação de Akaike (AIC). Compare com o obtido no modelo completo e comente.

(c) Pode afirmar-se que os parâmetros populacionais correspondentes às bandas 6 e 12 (na resolução de $20m$) são iguais? Justifique, usando um intervalo de confiança adequado.

(d) Um utilizador repara que, neste submodelo, apenas sobrou uma banda correspondente à resolução espacial de $10m$ e sugere que se considere a exclusão desse preditor, o que simplificaria a obtenção e processamento dos dados necessários à aplicação do modelo. Calcule o Coeficiente de Determinação do submodelo resultante dessa exclusão. Justifique e comente.

II [6,5 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.

```
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. Estime o parâmetro populacional β_2. Interprete o significado dessa estimativa no contexto do problema. O que pode concluir sobre o parâmetro populacional?
6. O teor de amido da terceira observação efectuada em Novembro, com o tratamento `D` foi 0.11194. Qual é o respectivo resíduo?

III [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.