INSTITUTO SUPERIOR DE AGRONOMIA

Modelos Matemáticos e Aplicações (2018-19)

2 de Maio, 2019 — **Teste – Modelo Linear** — Duração: 2h30

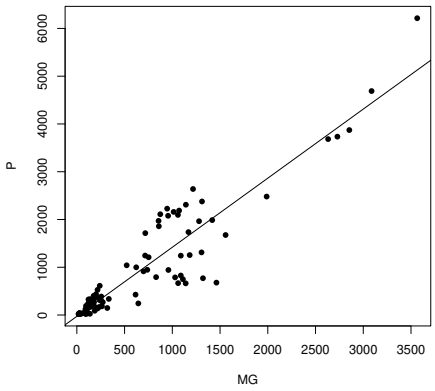
I [16 valores]

A Floresta Experimental H.J. Andrews, no Estado norte-americano do Oregon, disponibiliza numerosos conjuntos de dados florestais. Um subconjunto de dados tem medições de concentração de 12 nutrientes. As 92 medições foram efectuadas em três diferentes tipos de material lenhoso de árvores, que definem um factor (`TYPE`), com níveis identificados pelas siglas `TB` (casca), `TF` (folhagem) e `TW` (madeira). Os valores observados das concentrações de nutrientes são identificados pelos seus símbolos químicos. Estas concentrações são todas medidas em $mg\ kg^{-1}$, excepto o carbono (`C`), que é dado em percentagens. Em baixo indicam-se as médias, desvios padrão, valores mínimos e máximos de cada uma das variáveis numéricas.

	P	K	CA	MG	MN	CU	B	ZN	AL	FE	NA	C
$\overline{x}$	923.794	3633.663	7715.130	657.957	111.294	3.902	10.022	14.544	130.478	38.815	26.794	48.784
s	1149.549	3673.290	7881.323	736.304	151.833	1.562	4.622	19.417	188.951	36.020	17.379	2.405
min.	19.0	81.0	170.0	19.0	3.0	2.0	4.0	1.0	1.0	2.0	10.0	43.1
máx.	6214.0	16480.0	39600.0	3565.0	770.0	12.0	21.0	124.0	735.0	323.0	154.0	55.9

1. Pretende-se estudar a concentração de fósforo (`P`).

   Foi inicialmente ajustada uma Regressão Linear Simples sobre concentração de magnésio (`MG`), vendo-se na figura a nuvem de pontos e a recta de regressão respectivas. Sabe-se que o coeficiente de correlação entre estas duas variáveis é $r = 0.92633$.

   (a) Calcule a equação da recta de regressão, e discuta brevemente a qualidade desse ajustamento.

   (b) Calcule uma estimativa da variância σ^2 dos erros aleatórios do modelo.

   (c) A observação que surge no canto superior direito da nuvem de pontos é a observação 9. Calcule os valores do respectivo: (i) resíduo; (ii) efeito alavanca; (iii) resíduo (internamente) estandardizado; (iv) distância de Cook. Comente.

   (d) Indique uma característica expectável no gráfico de resíduos contra valores ajustados de y, para este modelo. Discuta as implicações dessa característica.

2. Foi efectuada uma ANOVA das concentrações de fósforo (variável `P`) sobre o tipo de material lenhoso (factor `TYPE`). A Soma de Quadrados associada aos efeitos do factor é $SQF = 72\,013\,571$.

   (a) Construa a tabela de síntese desta ANOVA.

   (b) Efectue o teste F correspondente e comente as suas conclusões.

3. Foi seguidamente ajustado um modelo ANCOVA, para saber se seria preferível ajustar rectas diferentes de `P` sobre `MG`, para cada tipo de material. Eis alguns resultados parciais desse ajustamento:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  212.4407    98.6069   2.154   0.0340
MG             0.4492     0.1734   2.591   0.0112
TYPETF        25.5918   157.5196   0.162   0.8713
TYPETW      -178.4209   152.7362  -1.168   0.2460
MG:TYPETF      0.9285     0.1911   4.859 5.26e-06
MG:TYPETW      0.1997     0.9369   0.213   0.8317
```

```
---
Residual standard error: 353.2 on 86 degrees of freedom
Multiple R-squared:  0.9108,Adjusted R-squared:  0.9056
F-statistic: 175.6 on 5 and 86 DF,  p-value: < 2.2e-16
```

(a) Escreva a equação da recta de regressão entre concentrações de fósforo e magnésio, relativa ao material de folhagem.

(b) Calcule um intervalo de confiança (95%) adequado para dizer se é admissível considerar que as rectas populacionais relativas ao material de folhagem e ao material de casca são paralelas. Comente.

(c) Teste formalmente se este modelo tem um ajustamento significativamente melhor do que o modelo de uma única recta de regressão para todas as observações, que foi ajustado no ponto 1.

4. Foi finalmente ajustada uma regressão linear da concentração de fósforo sobre cinco preditores: além de `MG`, também as concentrações de alumínio (`AL`), boro (`B`), manganês (`MN`) e zinco (`ZN`).

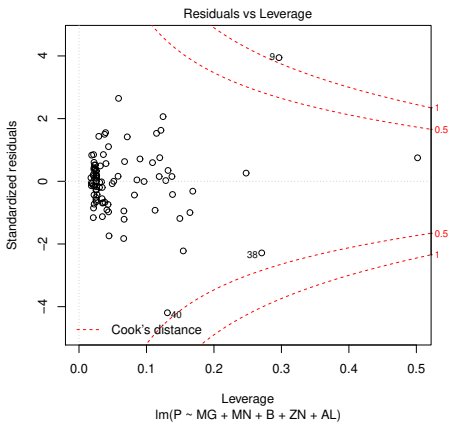
   (a) Comente a seguinte frase: "*o Quadrado Médio Residual deste modelo tem de ser mais pequeno que o Quadrado Médio Residual do modelo de Regressão Linear Simples ajustado no ponto 1*".

   (b) Na listagem produzida pelo `R`, resultante desse ajustamento, a linha acima do valor do coeficiente de determinação é a seguinte:

   ```
   Residual standard error: 249.7 on 86 degrees of freedom
   ```

   Calcule o valor do coeficiente de determinação deste modelo. Comente-o.

   (c) Descreva, e comente, o seguinte gráfico relativo ao modelo agora ajustado.



II [4 valores]

Considere uma regressão linear múltipla com p preditores e n observações, matriz do modelo $\mathbf{X}$, e equação $\vec{\mathbf{Y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}$. Considere o vector usual de estimadores, $\hat{\vec{\boldsymbol{\beta}}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}}$.

1. Mostre que o vector de valores ajustados, $\hat{\vec{\mathbf{Y}}} = \mathbf{X}\hat{\vec{\boldsymbol{\beta}}}$, é ortogonal ao vector dos resíduos, $\vec{\mathbf{E}}$.

2. Modifique o Modelo Linear, admitindo agora que a distribuição do vector dos erros aleatórios seja $\vec{\boldsymbol{\epsilon}} \frown \mathcal{N}_n(\vec{\mathbf{0}}, \Sigma)$, com a matriz Σ conhecida (não aleatória).

   (a) Diga, justificando, a que pressupostos corresponderia admitir que:

   i. A matriz Σ seja uma matriz *diagonal* genérica.

   ii. A matriz Σ seja uma matriz genérica, não diagonal.

(b) Sendo Σ uma matriz de (co-)variâncias genérica, determine as distribuições de probabilidades de:

   i. o vector $\vec{\mathbf{Y}}$ das observações;

   ii. o vector de estimadores $\vec{\hat{\boldsymbol{\beta}}}$.

(c) Considere um novo vector de estimadores dos parâmetros, dado por $\vec{\hat{\boldsymbol{\beta}}}_* = (\mathbf{X}^t \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^t \Sigma^{-1} \vec{\mathbf{Y}}$, com Σ genérica. Diga justificando, se se trata dum vector de estimadores centrado. Qual a respectiva matriz de (co-)variâncias?