Modelos Lineares e Aplicações

Teste - 09/06/2014

Duração: 2h

Departamento de Matemática e Aplicações

Nome: _____ Número: _____

1. Considere a base de dados "iris", automaticamente disponível no R (`> names(iris)`). Tratam-se de dados sobre 150 flores, onde a variável "species" identifica a espécie da flor (setosa, versicolor ou virginica) e a variável "Petal.Length" identifica o comprimento da pétala (em centímetro).

   (a) Execute os seguintes comandos em ambiente R:

   ```
   > attach(iris)
   > boxplot(Petal.Length ~ Species)
   ```

   Copie o gráfico para a Folha de Teste e interprete-o.

   (b) Analise se a v.a. "species" pode ser considerada uma variável explicativa da variável dependente "Petal.Length". Em caso afirmativo, escreva o modelo encontrado e interprete as estimativas dos repetivos coeficientes de regressão.

   (c) Escreva agora um novo modelo que considere simultaneamente a v.a. "species" e a v.a. "Petal.Width" (largura da pétala) como possíveis variáveis explicativas da v.a. "Petal.Length". Interprete a estimativa do coeficiente de regressão associado à v.a. "Petal.Width".

   (d) Qual o modelo preferível o da alínea b) ou c) ? Justifique.

2. (a) Explique qual a diferença entre os três seguintes tipos de observações discordantes: *outliers*, *leverages* e pontos *influentes*.

   (b) Considere, agora, que a procura de um determinado produto no mercado português tem evoluido da seguinte forma nos últimos 7 anos.

Ano (x)	2007	2008	2009	2010	2011	2012	2013
Aumento (y)	1.62	1.63	1.9	2.64	2.05	2.13	1.94

   i. Apresente um diagrama de dispersão, e estime a respetiva reta de regressão.

   ii. Prossiga com uma análise habitual de diagnóstico para os dados acima apresentados, identificando eventuais *outliers*, *leverages* e pontos *influentes*. Justifique os vários resultados apresentados.

   NOTA: Alguns comandos úteis em ambiente R ?rstudent, ?influence e ?dffits

3. Considere a base de dados "swiss", automaticamente disponível no R, com informação sobre indicadores de fertilidade (variável resposta) e indicadores sócio-económicos (variáveis explicativas) recolhidos em 47 províncias da Suiça em 1888.

   Suponha que pretende ajustar aos referidos dados um modelo

   $$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

   onde $E[\epsilon_i \epsilon_j] = 0$ para $i \neq j$ e $E[\epsilon_i^2] = \sigma^2$, e para cada provincia:

   - Y identifica um indicador de fertilidade ("Fertility");
   - X_1 identifica a % de recrutas com classificação máxima no exame do exército ("Examination");
   - X_2 identifica a % de recrutas com formação superior à escola primária ("Education");
   - X_3 identifica uma proporção de nados-vivos que faleceram com menos de 1 ano ("Infant.Mortality").

   (a) Obtenha as estimativas de mínimos quadrados de $\vec{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^t$ e interprete os valores obtidos para $\hat{\beta}_2$ e $\hat{\beta}_3$.

   (b) Considere a matriz de covariâncias e de variâncias dos estimadores dos parâmetros β (**sem a calcular**). Quais os valores encontrados na diagonal desta matriz ?

   (c) Apresente uma estimativa para σ^2.

   (d) Avalie a qualidade de ajustamento global, aplicando o teste de Fisher adequado. Especificando as hipóteses nula (H_0) e alternativa (H_1), diga qual o p-valor obtido e o que conclui ?

   (e) Aplique um teste de Fisher Parcial para avaliar se β_1 é nulo. Especificando as hipóteses nula (H_0) e alternativa (H_1), diga qual o p-valor obtido e o que conclui ?

   (f) Recorrendo ao cálculo de coeficientes de determinação, diga qual dos dois modelos é preferível: o modelo completo (X_1, X_2, X_3) ou modelo aninhado (X_2, X_3) ?

4. Considere a estatística de teste F, calculada na avaliação da qualidade de ajustamento global de um dado modelo (resultante de um Teste de Fisher). Prove que $F = \dfrac{n-p-1}{p} \dfrac{R^2}{1-R^2}$.