Modelos Lineares

Teste - 26/01/2017

Duração: 2h30m

Mestrado em Estatística

Departamento de Matemática e Aplicações

Nome: _____ Número: _____

Grupo I

1. Considere os dados sobre a fertilidade na Suiça:

$$> data(swiss)$$

Nota: Em todas as questões que utilizar o programa R apresentar os comandos usados.

(a) Ajuste um modelo de regressão com fertility como variável resposta e as restantes como variáveis explicativas aplicando o método backward com o critério AIC. Teste o ajustamento global do modelo.

(b) Interprete o valor da estimativa do coeficiente de uma variável explicativa e determine um intervalo de confiança.

(c) Teste se o modelo com as cinco variáveis explicativas e o modelo só com as variáveis Agriculture e Education diferem ou não significativamente quanto ao ajustamento aos dados. Explicite as hipóteses nula e alternativa´.

(d) Determine um intervalo a 99% de confiança para o valor esperado da fertility para a primeira observação da base de dados (corresponde provincia Courtelary).

2. Considere os dados do ficheiro *hospital.txt* de 113 hospitais contendo as seguintes variáveis:
id- identificação do hospital;
length- duração média de internamento no hospital (em dias);
age- idade média dos pacientes (em anos);
inf- risco médio de infecção (em percentagem);
cult- número de exames por doente x 100;
xray- número de raio-X por doente x 100;
beds- número de camas ;
school hospital universitário- 1=yes 2=no;
region- região geográfica 1=NE 2=N 3=S 4=W;
pat- número médio de pacientes por dia;
nurs- número de enfermeiros;
serv- percentagem de serviços disponiveis;

Usando as variáveis age, inf, region, beds, pat, nurs como variáveis explicativas e length como variável resposta, responda às seguintes questões:

(a) Calcula a matriz de correlação entre estas variáveis. Comente os resultados.

(b) Ajuste um modelo de regressão aplicando o método backward com teste F.

Nas alineas seguintes, considere o modelo ajustado na alinea (b). (Caso não tenha resolvido considere o modelo com todas as variáveis explicativas)

(c) Estime a duração média de internamento de um hospital com age=50, inf=3, region=W, beds=200, pat=200, nurs=200.

(d) Um investigador afirma: "a variação na *length* é idêntica na região N e S."Concorda com esta afirmação. Justifique aplicando um teste de hipóteses.

Grupo II

1. Pretende-se modelar a variável Y em função de 8 variáveis explicativas. Tem-se que a matriz de correlação é:

	X_1	X_2	X_3	X_4	Y	X_5	X_6	X_7	X_8
X_1	1.000	-0.125	0.463	0.340	0.060	0.099	-0.294	0.445	0.027
X_2	-0.125	1.000	0.318	-0.201	-0.165	0.050	0.227	-0.326	-0.111
X_3	0.463	0.318	1.000	0.293	0.011	0.058	-0.144	0.023	-0.018
X_4	0.340	-0.201	0.293	1.000	0.112	0.029	-0.117	0.380	-0.170
Y	0.060	-0.165	0.011	0.112	1.000	0.692	0.017	0.047	0.079
X_5	0.099	0.050	0.058	0.029	0.692	1.000	0.075	-0.028	0.068
X_6	-0.294	0.227	-0.144	-0.117	0.017	0.075	1.000	-0.293	-0.048
X_7	0.445	-0.326	0.023	0.380	0.047	-0.028	-0.293	1.000	-0.092
X_8	0.027	-0.111	-0.018	-0.170	0.079	0.068	-0.048	-0.092	1.000

Foi ajustada uma regressão linear múltipla para a totalidade das variáveis preditoras acima referidas. Foram obtidos os seguintes resultados gerais.

$> summary(lm(Y\ ., data = dados))$

Coefficients:

	Estimate	Std. Error	t value
(Intercept)	-0.0330805	0.3544070	-0.093
X_1	-0.0436050	0.0787920	-0.553
X_2	-0.6497245	0.4631393	-1.403
X_3	0.0784129	0.2000816	0.392
X_4	0.0517223	0.1307570	0.396
X_5	0.7943703	0.1435978	5.532
X_6	0.0004273	0.0072603	0.059
X_7	0.0038168	0.0327531	0.117
X_8	0.0060019	0.0343731	0.175

$Residual\ standard\ error:\ 0.06056\ on\ 30\ degrees\ of\ freedom$
$Multiple\ R-squared:\ 0.5281,\ Adjusted\ R-squared:???$
$F-statistic:\ ???\ on\ ???\ and\ ???\ DF$

(a) Complete os resultados obtidos, indicando os valores em falta (???). Justifique as suas respostas.

(b) Qual o número total de observações? Qual a variância amostral de Y?

(c) Um algoritmo de exclusão sequencial baseado no Critério de Informação de Akaike (AIC) optou por um submodelo com apenas dois preditores: X_2 e X_5. Com base na informação disponível até aqui, justifique que o coeficiente de determinação deste submodelo de dois preditores está entre 0.4788 e 0.5281.

(d) Considere que o R^2 do submodelo é 0.52. Calcule o valor do R^2 ajustado deste submodelo. Compare e comente os valores dos R^2 e R^2 ajustado, quer deste submodelo, quer do modelo original.

2. Pretende-se estudar a relação entre HC e ADF e Dieta, sendo Dieta uma variável categórica (3 diferentes dietas). Obteve-se:

```
Call: lm(formula = HC ~ ADF * Dieta, data = leitoes.ex109

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.53119    0.02197  24.182  < 2e-16
ADF          0.43018    0.05718   7.523 2.19e-08
Dieta2      -0.06750    0.05545  -1.217  0.23301
Dieta3      -0.29361    0.03596  -8.166 4.08e-09
ADF:Dieta2   0.07278    0.12795   0.569  0.57372
ADF:Dieta3   0.26871    0.08462   3.175  0.00345
---
Residual standard error: 0.02103 on 30 degrees of freedom
Multiple R-squared:  0.9483,Adjusted R-squared: 0.9396
F-statistic:   110 on 5 and 30 DF,  p-value: < 2.2e-16
```

(a) Indique o modelo de regressão ajustado para cada dieta.

(b) É admissível considerar que as retas populacionais das dietas 1 e 2 são paralelas? Justifique.

(c) Discuta a seguinte afirmação:"tendo em conta os valores obtidos, é possivel admitir que as retas de regressão populacionais para as dietas 1 e 2 são iguais".

Grupo III

1 Mostre que a estatística do teste F de ajustamento global do modelo se pode escrever apenas à custa de R^2 e R^2_{ajus}

2 Mostre que o coeficiente de determinação ajustado é negativo quando $R^2 < \frac{p}{n-1}$. Comente as implicações desta condição para a estatística do teste F de ajustamento global.