

INSTITUTO SUPERIOR DE AGRONOMIA
MATEMÁTICA E ESTATÍSTICA – 2009/10

13 de Novembro, 2009

PRIMEIRO TESTE

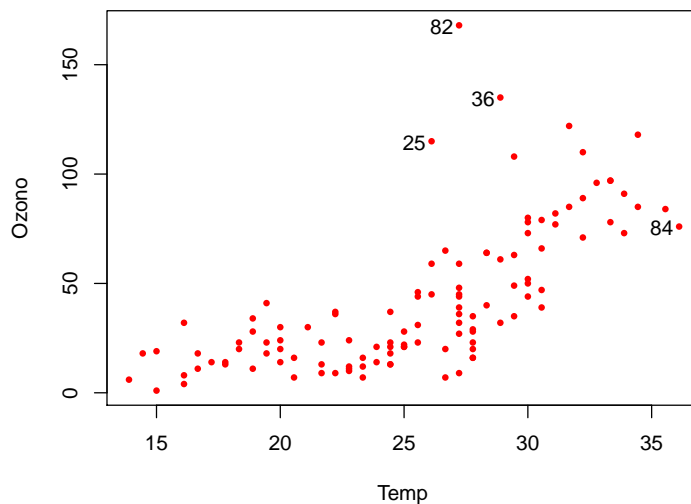
Duração: 2h30

I

Num estudo sobre poluição numa grande cidade, foram efectuadas medições, em 116 dias, da quantidade de ozono no ar (em partes por mil milhões) às 14h00 e da temperatura máxima (em °C) no respectivo dia. Alguns indicadores relativos às medições efectuadas são os seguintes:

	Temp	Ozono		Matriz de covariâncias
Min.	:13.89	Min.		Temp Ozono
1st Qu.	:21.67	1st Qu.		Temp 27.76989 121.4007
Median	:26.11	Median		Ozono 121.40067 1088.2005
Mean	:25.48	Mean		
3rd Qu.	:29.44	3rd Qu.		Matriz de correlações
Max.	:36.11	Max.		Temp Ozono
Var.	:27.77	Var.		Temp 1.0000000 0.6983603
				Ozono 0.6983603 1.0000000

O gráfico correspondente às observações efectuadas é o seguinte:



Tendo em conta a curvatura observada no gráfico, foi sugerido o ajustamento de um modelo exponencial, da forma $y = \alpha e^{\beta x}$.

1. Indique, justificando, como é possível linearizar este modelo.
2. Foi ajustado o *modelo linearizado*, tendo sido obtida a seguinte tabela de resultados:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.32212	0.26927	1.196	0.234
Temp	0.12150	0.01035	11.741	<2e-16

Residual standard error: 0.5848 on 114 degrees of freedom

- Indique, justificando, qual o teor médio de ozono (em partes por mil milhões) estimado pelo modelo ajustado, para um dia em que a temperatura máxima seja de 25°C.
- Calcule um intervalo de predição (95%) para o teor de ozono (em partes por mil milhões) num dia com a temperatura máxima de 25°C.

II

Num estudo sobre uma população experimental de clones da casta Aragonez, realizado em 2004 em Reguengos de Monsaraz, foram medidos os valores das seguintes variáveis em 255 genótipos (clones): grau brix (variável `grau.brix`); acidez total (em g de ácido tartárico por dm^3 , variável `acidez.total`); pH (variável `pH`); teor de antocianinas (em mg/dm^3 , variável `antocianinas`); e índice de polifenóis totais (variável `IPT`). Nesta casta, uma característica muito importante é a cor do bago, que está associada à pigmentação da película do bago, e portanto ao teor em antocianinas. Pretende-se modelar esta variável, à custa das restantes variáveis observadas. Eis algumas das observações e as médias e variâncias globais:

	<code>grau.brix</code>	<code>acidez.total</code>	<code>pH</code>	<code>antocianinas</code>	<code>IPT</code>
1	21.07	3.10	4.25	750.00	49.73
2	21.33	2.83	4.35	828.33	55.30
3	21.90	3.40	4.19	589.33	39.37
4	21.00	3.30	4.23	698.67	49.70
...
253	20.80	2.90	4.29	591.67	35.23
254	21.47	3.13	4.28	647.00	39.30
255	19.43	3.17	4.21	588.67	38.77

Média	20.72	3.14	4.25	593.55	39.40
Var.	1.4409	0.0336	0.0091	12960.3718	38.2689

A matriz de correlações entre as variáveis observadas é:

	<code>grau.brix</code>	<code>acidez.total</code>	<code>pH</code>	<code>antocianinas</code>	<code>IPT</code>
<code>grau.brix</code>	1.00000	-0.42275	0.47176	0.66949	0.44907
<code>acidez.total</code>	-0.42275	1.00000	-0.42345	-0.30279	-0.16163
<code>pH</code>	0.47176	-0.42345	1.00000	0.25292	0.16839
<code>antocianinas</code>	0.66949	-0.30279	0.25292	1.00000	0.78494
<code>IPT</code>	0.44907	-0.16163	0.16839	0.78494	1.00000

- Qual é a melhor variável preditora do teor de antocianinas, através duma regressão linear simples? Justifique e teste o ajustamento global do modelo resultante, ao nível de significância 0.05. Comente os seus resultados.
- Um modelo de regressão linear múltipla do teor de antocianinas sobre todas as restantes variáveis produziu os seguintes resultados:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-183.0714	217.7464	-0.841	0.4013
grau.brix	38.1179	3.9202	9.723	<2e-16
IPT	11.1392	0.6567	16.961	<2e-16
pH	-77.7083	44.9235	-1.730	0.0849
acidez.total	-38.8869	22.7263	-1.711	0.0883

Residual standard error: 57.75 on 250 degrees of freedom
 Multiple R-squared: 0.7467, Adjusted R-squared: 0.7426
 F-statistic: 184.2 on 4 and 250 DF, p-value: < 2.2e-16

Matriz de covariâncias estimadas entre os estimadores dos parâmetros

	(Intercept)	grau.brix	IPT	pH	acidez.total
(Intercept)	47413.494906	-92.392394	0.1908240	-8269.695174	-3302.1302779
grau.brix	-92.392394	15.368023	-1.0700515	-60.515625	23.3648109
IPT	0.190824	-1.070052	0.4313156	1.396217	-0.3018129
pH	-8269.695174	-60.515625	1.3962169	2018.124113	284.8631400
acidez.total	-3302.130278	23.364811	-0.3018129	284.863140	516.4850464

- Teste formalmente se este modelo é significativamente melhor que o modelo de regressão linear simples que escolheu no ponto 1.
- Como interpreta o valor -77.7083 que surge na primeira coluna da tabela?
- Com base neste modelo, construa um intervalo a 95% de confiança para a alteração no teor esperado de antocianinas resultante de um aumento *simultâneo* de um grau brix e de 1 g/dm^3 de acidez total (mantendo inalterados o IPT e o pH). Pode admitir-se que esses aumentos no brix e na acidez total não alteram o teor esperado de antocianinas?
- Qual o valor do Critério de Informação de Akaike (AIC) para este modelo? Diga, justificando, qual o maior valor da Soma de Quadrados Residual que um submodelo com apenas três preditores pode ter, para que seja considerado preferível a este modelo de quatro preditores, com base no AIC.

III

Um entomologista estudou a densidade de certa espécie de moscas em várias camadas de coberto vegetal duma floresta de folha caduca. Em particular, considerou três camadas de vegetação: herbácea, arbustiva e arbórea. Para cada camada, efectuou cinco medições de densidades das referidas moscas (em número de moscas por m^3). Obteve os seguintes resultados:

Camada	Densidades					Média	Variância
herbácea	14.0	12.1	9.6	8.2	10.2	10.82	5.122
arbustiva	8.4	5.1	5.5	6.6	6.3	6.38	1.637
arbórea	6.9	7.3	5.8	4.1	5.4	5.90	1.615

A média e variância da totalidade das observações são, respectivamente, 7.7 e 7.648571.

- Descreva o modelo estatístico adequado para determinar se as moscas têm preferência especial por alguma das camadas de vegetação.

2. Construa a tabela-resumo adequada ao estudo referido na alínea anterior, indicando como obtém cada um dos valores
3. Teste formalmente a hipótese de as moscas não terem preferência por qualquer das camadas de vegetação, indicando os vários passos do seu teste.
4. Independentemente dos resultados obtidos na alínea anterior, utilize o teste de Tukey para indicar quais os pares de camadas cuja densidade deve ser considerada significativamente diferente, ao nível de significância 0.05.

IV

Considere um Modelo de Regressão Linear Múltipla com p variáveis preditoras, e que é ajustado com base em n observações.

1. Seja \mathbf{E} o vector dos n resíduos associados ao modelo. Prove que a distribuição de \mathbf{E} é dada por:

$$\mathbf{E} \sim \mathcal{N}_n (\mathbf{0} , \sigma^2 (\mathbf{I} - \mathbf{H})) ,$$

onde \mathbf{I} indica a matriz identidade $n \times n$ e \mathbf{H} é a matriz de projecção ortogonal associada ao ajustamento do modelo.

2. Considere um submodelo com apenas k das p variáveis preditoras.
 - (a) Mostre que as Somas de Quadrados de Resíduos do modelo completo ($SQRE_C$) e do submodelo referido ($SQRE_S > 0$) verificam a seguinte igualdade:

$$\frac{SQRE_C}{SQRE_S} = \frac{1}{1 + \frac{F \cdot (p-k)}{n-(p+1)}} ,$$

onde F indica a estatística do teste F parcial para comparar o modelo e submodelo em questão.

- (b) Mostre que o quociente das Somas de Quadrados indicado na alínea anterior tem de pertencer ao intervalo $]0, 1]$.
- (c) Utilize o resultado da alínea 2a) para justificar a natureza unilateral direita da região crítica do referido teste F parcial.
- (d) Interprete geometricamente o facto de a Soma de Quadrados Residual de um modelo nunca poder ser superior à Soma de Quadrados Residual dum seu submodelo.