

INSTITUTO SUPERIOR DE AGRONOMIA
MATEMÁTICA E ESTATÍSTICA – 2009/10
PRIMEIRO TESTE

13 de Novembro, 2009

Uma resolução possível

I

1. $y = \alpha e^{\beta x} \Leftrightarrow \ln y = \ln \alpha + \beta x$, que é uma relação linear entre $y' = \ln y$ e x , de parâmetros $\alpha' = \ln \alpha$ e $\beta' = \beta$.

2. Com base no modelo linearizado e na informação do enunciado, tem-se:

(a) o valor estimado do logaritmo do teor de ozono é $\ln \hat{y} = 0.32212 + 0.12150(25) = 3.35962$. Logo, o teor médio de ozono estimado, nas unidades originais, é $\hat{y} = e^{3.35962} = 28.77825$ ppm.

(b) Vamos primeiro construir um intervalo de predição para o valor do logaritmo de teor de ozono, se $Temp = 25$, usando a expressão do formulário e tendo em conta a informação do enunciado:

$$\left[(a+bx) - t_{\frac{\gamma}{2}(n-2)} \sqrt{QMRE \left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2} \right]} \right. , \quad 3.35962 + t_{0.025(114)}(0.5848) \sqrt{1 + \frac{1}{116} + \frac{(25-25.48)^2}{115 \cdot 27.76989}} \left. \right] \\ \left[\quad 3.35962 - 1.98(0.5848)(1.004337) \quad , \quad 3.35962 + 1.162926 \quad \right] \\ \left[\quad 2.196694 \quad , \quad 4.522546 \quad \right]$$

Para ter um intervalo de predição para o teor de ozono nas unidades originais, tomamos os exponenciais dos extremos do intervalo, obtendo:

$$\left[8.995227 , 92.0697 \right] .$$

II

1. O melhor preditor linear será a variável mais correlacionada, em valor absoluto, com a variável resposta **antocianinas**. Tendo em conta a tabela dada, essa variável é IPT, para a qual $r = 0.78494$ e portanto o respectivo Coeficiente de Determinação será $R^2 = 0.6161308$. Neste modelo de Regressão Linear Simples, o teste de ajustamento global do modelo (cujas hipóteses se podem escrever como $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$) terá estatística do teste dada por $F = \frac{QMR}{QMRE} = (n-2) \frac{R^2}{1-R^2}$. A ser verdade H_0 , essa estatística tem distribuição $F_{(1,n-2)}$. A Região Crítica deste tipo de testes é unilateral direita, pelo que rejeitamos H_0 se $F_{calc} > f_{\gamma(1,n-2)} = f_{0.05(1,253)} \approx 3.90$ (os valores tabelados são $f_{0.05(1,120)} = 3.92$ e $f_{0.05(1,\infty)} = 3.84$). No nosso caso, $F_{calc} = 253 \times \frac{0.6161308}{1-0.6161308} = 406.0787$. De forma clara, rejeita-se H_0 e considera-se que o modelo não é inútil, apesar do valor pouco elevado de R^2 .

2. (a) É pedido um teste para comparar o

$$\text{Modelo Completo} \quad (C) \quad Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

com o

$$\text{Submodelo} \quad (S) \quad Y = \beta_0 \quad + \beta_2 x_2,$$

onde Y representa o teor de antocianinas, x_1 os graus brix, x_2 o IPT, x_3 o pH e x_4 a acidez total. Efectuamos um teste F parcial para comparar este modelo e submodelo.

Hipóteses: $H_0 : \beta_1 = \beta_3 = \beta_4 = 0$ vs. $H_1 : (\beta_1 \neq 0) \vee (\beta_3 \neq 0) \vee (\beta_4 \neq 0)$
i.e, [Modelo e Submodelo iguais] vs. [Modelo e Submodelo diferem]

Estatística: $F = \frac{(SQRE_S - SQRE_C)/(p-k)}{SQRE_C/(n-(p+1))} = \frac{n-(p+1)}{p-k} \cdot \frac{R_C^2 - R_S^2}{1 - R_C^2} \cap F_{(p-k, n-(p+1))}$, sob H_0 .

Nível de significância: $\gamma = P[\text{Erro do tipo I}] = P[\text{Rej. } H_0 | H_0 \text{ verdade}] = 0.05$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{\gamma(p-k, n-(p+1))} = f_{0.05(3, 250)} \approx 2.64$.

Conclusões: Temos $n = 255$, $p = 4$, $k = 1$, $R_C^2 = 0.7467$ (dado no enunciado). Uma vez que o submodelo é uma regressão linear simples, o seu Coeficiente de Determinação é o quadrado do coeficiente de correlação entre o preditor e a variável resposta, isto é, entre IPT e antocianinas. Logo, $R_S^2 = (0.78494)^2 = 0.6161308$, e $F_{calc} = \frac{250}{3} \cdot \frac{0.7467 - 0.6161308}{1 - 0.7467} = 42.956$. Assim, rejeita-se H_0 ao nível $\gamma = 0.05$, ou seja, a regressão linear simples é significativamente pior que o modelo completo com 4 preditores.

- (b) O valor indicado é a estimativa de β_3 , o coeficiente que multiplica a variável pH no modelo de relação base $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$. Assim, estimamos que um aumento de 1 valor na escala pH esteja associado a uma diminuição esperada de 77.7083 g/dm^3 no teor de antocianinas.
- (c) É pedido um intervalo para a variação em $E[Y]$ associada a aumentar simultaneamente x_1 e x_4 em uma unidade. Ora,

$$\begin{aligned} E[Y|X_1=x_1+1, X_2=x_2, X_3=x_3, X_4=x_4+1] &= \beta_0 + \beta_1(x_1+1) + \beta_2 x_2 + \beta_3 x_3 + \beta_4(x_4+1) \\ - E[Y|X_1=x_1, X_2=x_2, X_3=x_3, X_4=x_4] &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \\ \hline & \beta_1 \qquad \qquad \qquad + \beta_4 \end{aligned}$$

Assim, o que se pede é um IC a 95% para $\beta_1 + \beta_4$. Trata-se dum caso particular duma combinação linear dos parâmetros β_i , $\mathbf{a}^t \boldsymbol{\beta}$, sendo neste caso $\mathbf{a}^t = (0, 1, 0, 0, 1)$. A partir do formulário o intervalo pedido tem a forma:

$$\left] \mathbf{a}^t \hat{\boldsymbol{\beta}} - t_{\gamma/2(n-(p+1))} \cdot \hat{\sigma}_{\mathbf{a}^t \hat{\boldsymbol{\beta}}} \quad , \quad (b_1 + b_4) + t_{0.025(250)} \cdot \hat{\sigma}_{\hat{\beta}_1 + \hat{\beta}_4} \left[.$$

Ora, pelo enunciado, $b_1 = 38.1179$, $b_4 = -38.8869$, $t_{0.025(250)} \approx 1.97$ e (tendo em conta a matriz de covariâncias associada à estimação dos parâmetros) $\hat{\sigma}_{\hat{\beta}_1 + \hat{\beta}_4} = \sqrt{\hat{V}[\hat{\beta}_1 + \hat{\beta}_4]} = \sqrt{\hat{V}[\hat{\beta}_1] + \hat{V}[\hat{\beta}_4] + 2 \hat{Cov}[\hat{\beta}_1, \hat{\beta}_4]} = \sqrt{15.368023 + 516.4850464 + 2(23.3648109)} = 24.05375$. Logo, o IC pedido é:

$$\left] -0.769 - 1.970(24.05375) \quad , \quad -0.769 + 47.38588 \left[\right. \\ \left. \right] -48.15488 \quad , \quad 46.61688 \left[\right.$$

O intervalo inclui o valor zero, pelo que é admissível (a 95% de confiança) a hipótese de que não há alteração no teor esperado de antocianinas.

- (d) Pelo formulário, $AIC = n \ln \left(\frac{SQRE_k}{n} \right) + 2(k+1)$. Neste caso, $k = 4$, $n = 255$ e $SQRE = (\sqrt{QMRE})^2 (255 - 5) = (57.75^2) 250 = 833765.6$. logo, $AIC = 2073.573$. Para ser escolhido, um submodelo com apenas 3 preditores tem de ter um AIC menor, ou seja, $AIC_3 < 2073.573$, indicando por AIC_3 o AIC dum submodelo com 3 preditores. Ora,

$AIC_3 = 255 \ln \left(\frac{SQRE_3}{n} \right) + 8$, onde $SQRE_3$ indica a Soma de Quadrados Residual do sub-modelo com 3 preditores. Assim,

$$\begin{aligned} AIC_3 < 2073.573 &\Leftrightarrow \ln \left(\frac{SQRE_3}{255} \right) < \frac{2073.573 - 8}{255} = 8.100287 \\ &\Leftrightarrow SQRE_3 < e^{8.100287} \cdot 255 = 840\,330.7 \end{aligned}$$

Logo, $SQRE_3 = 840\,330.7$ é o maior valor de $SQRE$ para o qual um submodelo com 3 preditores não difere significativamente do modelo de 4 preditores dado no enunciado, através do critério do AIC.

III

- Trata-se de um delineamento a um factor (camada de vegetação), sendo a variável resposta numérica (Y) a densidade das moscas. O delineamento é equilibrado, pois existe igual número ($n_c = 5$) de observações em cada um dos $k = 3$ níveis do factor. Para responder à questão, podemos admitir um modelo ANOVA a um factor, e testar a existência de efeitos de nível, i.e., testar se as médias populacionais são iguais nas três camadas de vegetação. Eis o modelo:

- (a) A observação j no i -ésimo nível do factor (camada vegetativa) é dada por:

$$Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij}, \quad \forall i = 1, 2, 3 \quad \text{e} \quad \forall j = 1, 2, 3, 4, 5,$$

tendo-se ainda $\alpha_1 = 0$. O parâmetro μ_1 representa a média populacional no primeiro nível do factor, enquanto que α_2 e α_3 representam os acréscimos a μ_1 , que conduzem às médias populacionais nos segundo e terceiro níveis do factor: $\mu_2 = \mu_1 + \alpha_2$ e $\mu_3 = \mu_1 + \alpha_3$. Finalmente, os ϵ_{ij} representam erros aleatórios aditivos.

- (b) $\epsilon_{ij} \cap \mathcal{N}(0, \sigma^2)$, $\forall i, j$.
(c) $\{\epsilon_{ij}\}_{i,j}$ são variáveis aleatórias independentes.

- Pelo formulário,

$$\begin{aligned} SQF &= \sum_{i=1}^k n_i (\bar{y}_i - \bar{y}_{..})^2 = 5 [(\bar{y}_1 - \bar{y}_{..})^2 + (\bar{y}_2 - \bar{y}_{..})^2 + (\bar{y}_3 - \bar{y}_{..})^2] \\ &= 5 [(10.82 - 7.7)^2 + (6.38 - 7.7)^2 + (5.90 - 7.7)^2] = 73.584. \end{aligned}$$

Tem-se ainda, $SQT = (n - 1) s_y^2 = 14(7.648571) = 107.080$ e $SQRE = SQT - SQF = 33.496$. Logo, a tabela-resumo da ANOVA é:

	SQs	gl	QMs	F
Factor	73.584	$k - 1 = 2$	$QMF = \frac{SQF}{k-1} = 36.792$	$F = \frac{QMF}{QMRE} = 13.181$
Residual	33.496	$n - k = 12$	$QMRE = \frac{SQRE}{n-k} = 2.791$	-
Total	107.080	$n - 1 = 14$	-	-

- O teste aos efeitos do factor:

Hipóteses: $H_0 : \alpha_2 = \alpha_3 = 0$ vs. $H_1 : (\alpha_2 \neq 0) \vee (\alpha_3 \neq 0)$
[Camada não afecta densidade] vs. [Camada afecta densidade]

Estatística do Teste: $F = \frac{QMF}{QMRE} \cap F_{(k-1, n-k)}$, sob H_0 .

Nível de significância: $\gamma = P[\text{Erro do tipo I}] = P[\text{Rej. } H_0 | H_0 \text{ verdade}] = 0.05$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{\gamma(k-1, n-k)} = f_{0.05(2, 12)} \approx 3.89$.

Conclusões: Como $F_{calc} = 13.181 > 3.89$, rejeita-se H_0 ao nível $\gamma = 0.05$, ou seja, concluímos que as moscas têm preferências diferentes por diferentes camadas.

4. Com base no teste de Tukey, duas médias populacionais de nível, μ_i e $\mu_{i'}$, devem ser consideradas diferentes se as respectivas médias amostrais excederem o termo de comparação, isto é, se

$$|\bar{y}_i - \bar{y}_{i'}| > q_{\gamma(k, n-k)} \sqrt{\frac{QMRE}{n_c}} = q_{0.05(3, 12)} \sqrt{\frac{2.7913}{5}} = 3.77(0.747168) = 2.82.$$

Ora,

$$\begin{aligned} |\bar{y}_1 - \bar{y}_2| &= |10.82 - 6.38| = 4.44 \\ |\bar{y}_1 - \bar{y}_3| &= |10.82 - 5.90| = 4.92 \\ |\bar{y}_1 - \bar{y}_2| &= |6.38 - 5.90| = 0.48 \end{aligned}$$

Logo, conclui-se que o primeiro nível (camada herbácea) tem média diferente das outras duas camadas, que por sua vez não diferem entre si, em média. Tendo em conta o valor das médias, podemos afirmar que as moscas revelam uma preferência pela camada herbácea.

IV

1. Por definição, $\mathbf{E} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$. Logo, o vector \mathbf{E} dos resíduos é dado por transformações lineares dos valores do vector \mathbf{Y} . Ora, sabemos que a partir dos pressupostos do modelo ($\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \cap \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$) decorre que $\mathbf{Y} \cap \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Logo, $\mathbf{E} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$ também tem de ter distribuição Multinormal (Acetato 128). Os parâmetros dessa distribuição são (tendo em conta as propriedades dos vectores esperados e das matrizes de covariâncias de vectores aleatórios):

$$E[\mathbf{E}] = E[(\mathbf{I} - \mathbf{H})\mathbf{Y}] = (\mathbf{I} - \mathbf{H})E[\mathbf{Y}] = (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{H}\mathbf{X}\boldsymbol{\beta}.$$

Ora, \mathbf{H} é a matriz de projecção ortogonal sobre o subespaço $\mathcal{C}(\mathbf{X})$, e as colunas de \mathbf{X} ficam invariantes quando projectadas num subespaço ao qual já pertencem, ou seja, $\mathbf{H}\mathbf{X} = \mathbf{X}$. Logo, $E[\mathbf{E}] = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}$. Por outro lado, $V[\mathbf{E}] = V[(\mathbf{I} - \mathbf{H})\mathbf{Y}] = (\mathbf{I} - \mathbf{H})V[\mathbf{Y}](\mathbf{I} - \mathbf{H})^t = \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I}^t - \mathbf{H}^t)$. Tendo em conta que $\mathbf{I}^t = \mathbf{I}$, $\mathbf{I}^2 = \mathbf{I}$, $\mathbf{H}^t = \mathbf{H}$ e $\mathbf{H}^2 = \mathbf{H}$ (recorde-se que matrizes de projecção ortogonal são sempre simétricas e idempotentes), temos: $V[\mathbf{E}] = \sigma^2(\mathbf{I}^2 - \mathbf{H} - \mathbf{H} + \mathbf{H}^2) = \sigma^2(\mathbf{I} - \mathbf{H})$, como se queria mostrar.

2. (a) A estatística do teste F parcial é, como indicado no formulário,

$$\begin{aligned} F &= \frac{SQRE_S - SQRE_C}{SQRE_C} \cdot \frac{n - (p + 1)}{p - k} = \left(\frac{SQRE_S}{SQRE_C} - 1 \right) \frac{n - (p + 1)}{p - k} \\ \iff \frac{F \cdot (p - k)}{n - (p + 1)} + 1 &= \frac{SQRE_S}{SQRE_C} \\ \iff \frac{SQRE_C}{SQRE_S} &= \frac{1}{\frac{F \cdot (p - k)}{n - (p + 1)} + 1} \end{aligned}$$

- (b) Ora, por definição, $\frac{SQRE_C}{SQRE_S} > 0$. Por outro lado, o valor da estatística F não pode ser negativo (o que pode ser confirmado, quer pela expressão que define F , quer pelo facto de ter uma distribuição F), pelo que $1 + \frac{F(p-k)}{n-(p+1)} \geq 1 \Leftrightarrow \frac{SQRE_C}{SQRE_S} \leq 1$.
- (c) Quanto maior fôr o valor da estatística F , mais próximo de zero está o quociente $\frac{SQRE_C}{SQRE_S}$, isto é, tanto maior será $SQRE_S$, em relação a $SQRE_C$. Ora, se a Soma de Quadrados Residual ajustada do Submodelo fôr muito maior que a Soma de Quadrados Residual ajustada do modelo completo, é duvidoso que modelo e submodelo sejam equivalentes, como afirma a Hipótese Nula do teste F parcial.
- (d) Vimos nas aulas teóricas que $SQRE$ é o quadrado da distância entre o vector \mathbf{Y} das observações e o vector $\hat{\mathbf{Y}}$ que é a projecção ortogonal de \mathbf{Y} sobre o espaço coluna da matriz \mathbf{X} , $\mathcal{C}(\mathbf{X})$. Ora, o modelo completo tem associada uma matriz \mathbf{X}_C com uma coluna de uns e p colunas adicionais, cada uma das quais tem os valores observados duma das variáveis preditoras. A correspondente matriz do submodelo, \mathbf{X}_S , tem apenas algumas das colunas de \mathbf{X}_C . Logo, o subespaço coluna de \mathbf{X}_S está contido no subespaço coluna de \mathbf{X}_C : $\mathcal{C}(\mathbf{X}_S) \subset \mathcal{C}(\mathbf{X}_C)$. Como a projecção ortogonal de \mathbf{Y} em qualquer subespaço resulta no vector desse subespaço que está mais próximo de \mathbf{Y} (à menor distância de \mathbf{Y}), esta relação de inclusão de subespaços implica que \mathbf{Y} tem de estar tão ou mais próximo da sua projecção $\hat{\mathbf{Y}}_C$ sobre $\mathcal{C}(\mathbf{X}_C)$ do que da sua projecção $\hat{\mathbf{Y}}_S$ sobre $\mathcal{C}(\mathbf{X}_S)$. Ou seja, $SQRE_C \leq SQRE_S$.