

INSTITUTO SUPERIOR DE AGRONOMIA  
ESTATÍSTICA E DELINEAMENTO – 2008/09  
**RESOLUÇÃO DE ALGUNS EXERCÍCIOS**

I

1. A equação de base do modelo pode ser re-escrita como:

$$\begin{aligned}y &= \frac{ax}{b+x} \Leftrightarrow \frac{1}{y} = \frac{b+x}{ax} \\ &\Leftrightarrow \frac{1}{y} = \frac{b}{a} \cdot \frac{1}{x} + \frac{1}{a} \\ &\Leftrightarrow y' = \alpha' + \beta'x'\end{aligned}$$

com  $y' = \frac{1}{y}$ ,  $x' = \frac{1}{x}$ ,  $\alpha' = \frac{1}{a}$  e  $\beta' = \frac{b}{a}$ .

2. Em relação à regressão linearizada:

(a) Valores omissos na tabela:

- Valor da estatística  $t$ : por definição,  $t_{calc} = \frac{b}{\hat{\sigma}_\beta}$ . Tendo em conta os valores da tabela, obtém-se  $t_{calc} = \frac{0.0002472}{0.0000321} \approx 7.70$ .
- Valor da estatística  $F$ : sabemos que numa Regressão Linear Simples, a estatística  $F$  é o quadrado da estatística  $t$  associada ao teste do declive da recta,  $\beta$ . Assim,  $F_{calc} \approx 7.70^2 \approx 59.3$ .
- Graus de liberdade da estatística  $F$ : sabemos que, nas Regressões Lineares em geral, são dados por  $p$  e  $n - (p + 1)$ , onde  $p$  indica o número de variáveis preditoras ( $p + 1$  é o número de parâmetros do modelo) e  $n$  o número de observações com base nas quais foi ajustado o modelo. Tem-se  $p = 1$  e  $n = 12$ . Assim, os graus de liberdade são, respectivamente, 1 e 10.

(b) O teste de ajustamento do Modelo de Regressão Linear consiste num teste em que se coloca como Hipótese Nula a equivalência do modelo com o modelo sem preditores,  $y = \alpha + \epsilon$ , e como Hipótese Alternativa a negação dessa hipótese. Assim, e no caso duma Regressão Linear Simples:

**Hipóteses:**  $H_0 : \beta = 0$ , vs.  $H_1 : \beta \neq 0$

**Estatística do Teste:**  $F = \frac{QMR}{QMRE} \cap F_{(p, n-(p+1))}$  sob  $H_0$ .

**Nível de Significância:**  $\gamma = 0.05$

**Região Crítica:** Unilateral direita, rejeitando-se  $H_0$  caso  $F_{calc} > f_{\gamma(1,10)}$ .

**Valor calculado da estatística:** Na alínea anterior verificou-se que este valor é 59.3, e que a sua significância empírica ( $p$ -value) é quase nula (0.0000164), pelo que se encontra claramente dentro de qualquer Região Crítica plausível (isto é, para qualquer nível de significância sensato).

Assim sendo, rejeita-se a Hipótese Nula, ou seja, o Modelo Linear ajustado é significativamente diferente do modelo sem preditores. Como é sabido, esta conclusão **não** equivale, por si só, a afirmar que o Modelo se ajusta bem aos dados, mesmo tendo em conta o baixíssimo *p-value* associado ao teste. Mas o valor relativamente elevado do Coeficiente de Determinação (0.8557) aponta para um ajustamento efectivamente bom.

- (c) O coeficiente da variável preditora indica a *variação esperada na variável resposta, dado um aumento unitário na variável preditora*. No nosso caso, e tendo em conta as transformações, tal significa que o valor esperado do recíproco da taxa de reacção, por cada aumento de uma unidade no recíproco da concentração, é de 0.0002472.
- (d) Pede-se para calcular um intervalo de confiança para  $\alpha$  (na relação entre as variáveis transformadas). Tal intervalo é da forma (onde, no extremo inferior se indicam as quantidades envolvidas, e no extremo superior os respectivos valores):

$$\left] a' - t_{0.025} \cdot \hat{\sigma}_{\hat{\alpha}'} \quad , \quad 0.0051072 + 2.228 \cdot 0.0007040 \left[ \right. \\ \left. \right] 0.00354 \quad , \quad 0.00668 \left[ \right. .$$

Tendo em conta que este é um intervalo a 95% de confiança para  $\alpha' = \frac{1}{a}$ , e que a assíntota horizontal à direita da curva  $y = \frac{ax}{b+x}$  é a recta  $y = a$ , temos que a hipótese do enunciado corresponde a admitir que  $a = 210$ , em cujo caso  $\frac{1}{a} = 0.004762$ . Este é um valor admissível à luz do intervalo de confiança obtido.

## II

- $b_3 = 0.61469$  corresponde à estimativa da variação esperada no peso da árvore (variável resposta) para um aumento de um *dm* no perímetro do tronco (variável preditora associada ao coeficiente  $\beta_3$ ).
- Pede-se para efectuar o seguinte teste:

**Hipóteses:**  $H_0 : \beta_1 = 0$  vs.  $H_1 : \beta_1 \neq 0$

**Estatística do Teste:**  $t_{calc} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \cap t_{n-(p+1)}$ , caso  $H_0$  seja verdadeira.

**Nível de Significância:**  $\gamma = 0.05$

**Região Crítica:** Bilateral, rejeitando-se  $H_0$  caso  $|t_{calc}| > t_{\frac{\alpha}{2}(n-(p+1))} = t_{0.025(44)} = 2.015368$ .

**Conclusões:** Pelo enunciado temos  $t_{calc} = -2.397$ , pelo que se rejeita  $H_0$  ao nível de significância  $\gamma = 0.05$ .

Assim, a variável  $Y1$  (perímetro do tronco) parece contribuir de forma não negligenciável (embora não particularmente enfática) para os valores da variável resposta (peso da árvore).

- O modelo ajustado tem três variáveis preditoras. Destas, apenas uma tem coeficiente associado que não difere significativamente de zero, como se depreende da leitura, no enunciado, da tabela que resume os resultados do ajustamento, em que apenas a variável  $Y2$  tem um *p-value* associado relativamente elevado: 0.12630. Assim, essa é a única variável que, individualmente, poderá ser excluída do modelo sem que tal decisão afecte de forma significativa o ajustamento. O melhor submodelo com duas variáveis preditoras será assim o submodelo com as variáveis  $Y1$  (perímetro do tronco) e  $Y3$  (perímetro do tronco aos 15 anos de vida) como preditores.

4. Nesta alínea pede-se para comparar dois Modelos, sendo um deles o Modelo Completo, com as três variáveis preditoras  $Y_1$ ,  $Y_2$  e  $Y_3$ , e o outro o submodelo que apenas tem a variável preditora  $Y_3$ . Utilizando o Teste aos modelos encaixados (teste  $F$  parcial), temos:

**Hipóteses:**  $H_0 : \beta_1 = \beta_2 = 0,$  vs.  $H_1 : \beta_1 \neq 0 \vee \beta_2 \neq 0$   
 $\Leftrightarrow$  ( Submodelo admissível) vs. ( Submodelo não admissível)

**Estatística do Teste:**  $F = \frac{R_C^2 - R_S^2}{1 - R_C^2} \cdot \frac{n - (p+1)}{p - k} \cap F_{(p-k, n-(p+1))}$ , caso  $H_0$  seja verdadeira, e onde  $R_C^2$  indica o coeficiente de Determinação do Modelo Completo,  $R_S^2$  indica o Coeficiente de Determinação do Submodelo,  $p$  e  $k$  indicam o número de preditores no modelo completo e no Submodelo, respectivamente, e  $n$  o número de observações.

**Nível de Significância:**  $\gamma = 0.05$ .

**Região Crítica:** Unilateral. Rejeitar  $H_0$  se  $F_{calc} > f_{\gamma(2,44)} = 3.209278$ , para  $\gamma = 0.05$ .

**Conclusões:** No nosso caso, e tendo em atenção os valores dados no enunciado, tem-se  $R_C^2 = 0.908$  e  $\frac{n-(p+1)}{p-k} = 22$ . O valor que não está imediatamente disponível é o valor de  $R_S^2$ . Mas tendo em atenção que o Submodelo corresponde a uma Regressão Linear Simples, e que nesse tipo de regressão o coeficiente de determinação é o quadrado do coeficiente de correlação entre a (única) variável preditora e a variável resposta, temos (a partir da matriz de correlações dada no enunciado)  $R_S^2 = (0.9456171)^2 = 0.8941917$ . Assim, tem-se:  $F_{calc} = 3.301985$ , pelo que (por pouco) se rejeita  $H_0$  ao nível  $\gamma = 0.05$ , ou seja, pode considerar-se que os modelos diferem de forma significativa.

5. Apenas se altera a variável resposta, que passa a ser  $Y^* = Y_4 * 0.4536$ . Assim,
- $\mathbf{H}$  não se altera, uma vez que a matriz  $\mathbf{X}$  definida pelas variáveis preditoras (e o vector de  $n$  uns, associado à constante aditiva no modelo) é a mesma.
  - O novo vector ajustado corresponde a multiplicar o anterior vector ajustado pela constante 0.4536. De facto,  $\hat{\mathbf{Y}}^*$  é dado por  $\hat{\mathbf{Y}}^* = \mathbf{H}\mathbf{Y}^* = \mathbf{H}(\mathbf{Y}_4 \cdot 0.4536) = 0.4536 \hat{\mathbf{Y}}_4$ , onde  $\hat{\mathbf{Y}}_4 = \mathbf{H}\mathbf{Y}_4$  indica o vector ajustado expresso em libras.
  - O Coeficiente de Determinação  $R^2$  não sofre alterações. De facto, sabemos que  $R^2 = \frac{SQR}{SQT} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$ . Já se viu que a transformação dos valores observados  $y_i$  de libras em quilogramas implica que também os valores ajustados  $\hat{y}_i$  sofram uma transformação multiplicativa análoga. A mesma transformação afecta a média  $\bar{y}$  das observações. Assim, quer a Soma de Quadrados associada à Regressão ( $SQR$ ) quer a Soma de Quadrados Total  $SQT$  sofrem uma transformação multiplicativa com o mesmo factor  $0.4536^2$ , pelo que o quociente  $R^2 = \frac{SQR}{SQT}$  permanece inalterado.

### III

- O Modelo de Regressão Linear Simples admite que existem  $n$  pares de observações  $\{(x_i, y_i)\}_{i=1}^n$ , com as observações da variável  $X$  consideradas fixas e as observações da variável  $Y$  sendo realizações de v.a.s  $Y_i$  para as quais:
  - $Y_i = \alpha + \beta x_i + \epsilon_i, \forall i = 1, 2, \dots, n$ , ( $\alpha, \beta, x_i$  constantes;  $\{\epsilon_i\}_{i=1}^n$  variáveis aleatórias).

- Os *erros aleatórios* verificam  $\epsilon_i \cap \mathcal{N}(0, \sigma), \forall i = 1, 2, \dots, n$ .
- Os erros aleatórios  $\{\epsilon_i\}_{i=1}^n$  são variáveis aleatórias independentes.

Como se viu nas aulas, dado este modelo, temos:

- As variáveis  $Y_i$  são variáveis independentes, com distribuição  $Y_i \cap \mathcal{N}(\alpha + \beta x_i, \sigma)$ ;
- o estimador do parâmetro  $\alpha$  é dado por:  $\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{x} = \sum_{i=1}^n d_i Y_i$ , onde  $d_i = \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}$ , sendo  $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x})x_i$ .

Logo,

- $\hat{\alpha}$  é uma combinação linear de Normais independentes, e portanto Normal;
- Pelas propriedades do valor esperado (e como  $E[\hat{\beta}] = \beta$  e  $E[Y_i] = \alpha + \beta x_i$ ), tem-se:  
 $E[\hat{\alpha}] = E[\bar{Y}] - \bar{x} E[\hat{\beta}] = E\left[\frac{1}{n} \sum_{i=1}^n Y_i\right] - \beta \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n (\alpha + \beta x_i) - \beta \bar{x} = \alpha + \beta \bar{x} - \beta \bar{x} = \alpha$ .
- Pelas propriedades da variância (e tendo em conta que os  $Y_i$  são v.a.s independentes e as expressões disponíveis no Formulário:  $V[\hat{\beta}] = \sigma^2/S_{xx}$  e  $Cov[\hat{\alpha}, \hat{\beta}] = -\sigma^2 \bar{x}/S_{xx}$ ), tem-se:

$$\begin{aligned} V[\bar{Y}] &= V[\hat{\alpha} + \hat{\beta} \bar{x}] \\ \Leftrightarrow V\left[\frac{1}{n} \sum_{i=1}^n Y_i\right] &= V[\hat{\alpha}] + V[\hat{\beta} \bar{x}] + 2 \cdot Cov[\hat{\alpha}, \bar{x} \hat{\beta}] \\ \Leftrightarrow \frac{1}{n^2} \cdot \sum_{i=1}^n V[Y_i] &= V[\hat{\alpha}] + \bar{x}^2 \cdot V[\hat{\beta}] + 2 \bar{x} \cdot Cov[\hat{\alpha}, \hat{\beta}] \\ \Leftrightarrow \frac{1}{n^2} \cdot n \sigma^2 &= V[\hat{\alpha}] + \bar{x}^2 \cdot \frac{\sigma^2}{S_{xx}} - 2 \cdot \frac{\bar{x}^2 \sigma^2}{S_{xx}} \\ \Leftrightarrow V[\hat{\alpha}] &= \frac{\sigma^2}{n} - \bar{x}^2 \cdot \frac{\sigma^2}{S_{xx}} + 2 \cdot \frac{\bar{x}^2 \sigma^2}{S_{xx}} \\ \Leftrightarrow V[\hat{\alpha}] &= \sigma^2 \cdot \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]. \end{aligned}$$

2. Sabemos que, em  $\mathbb{R}^n$ , o Coeficiente de Determinação  $R^2$  é o **quadrado do** cosseno do ângulo  $\theta$  entre o vector centrado das observações de  $Y$ ,  $\mathbf{y}^c$ , e a sua projecção ortogonal sobre o subespaço gerado pelas colunas da matriz  $\mathbf{X}$  (definida à custa das variáveis predictoras),  $\mathbf{Hy}^c$ . Nesse caso, a razão  $\frac{R^2}{1-R^2}$  é a cotangente ao quadrado do ângulo  $\theta$ . Então,

$$\begin{aligned} F_{calc} &= \frac{n - (p + 1)}{p} \cdot ctg^2 \theta > f_\gamma \\ \Leftrightarrow ctg \theta &> \sqrt{\frac{p}{n - (p + 1)}} \cdot f_\gamma \\ \Leftrightarrow tg \theta &< \sqrt{\frac{n - (p + 1)}{p \cdot f_\gamma}} \\ \Leftrightarrow \theta &< arctg \sqrt{\frac{n - (p + 1)}{p \cdot f_\gamma}} \end{aligned}$$

Assim, a condição para a rejeição da Hipótese Nula (ou seja, para a rejeição da inutilidade do modelo de regressão linear) é que o ângulo entre  $\mathbf{y}^c$  e a sua projecção ortogonal sobre  $\mathcal{C}(\mathbf{X})$  não seja demasiado grande.