

## Exercícios - Estatística e Delineamento - 2011/12

### 2 Regressão Linear Simples

1. Suponha que foi realizado um ensaio para avaliar o crescimento radicular de uma certa cultivar de uma espécie agrícola. Para o efeito, foi medido o comprimento (em *mm*) da raíz principal ( $y$ ), decorridos  $x$  dias. Obtiveram-se os seguintes resultados:

$x$	1	7	13	20	27	34	62
$y$	5	10	12	29	36	83	102

Utilize o programa R para responder às seguintes questões.

- Crie dois vectores com os valores das variáveis  $x$  e  $y$ .
  - Crie um objecto de tipo *data frame* com as duas variáveis.
  - Crie uma nuvem de pontos dos dados.
  - Ajuste uma recta de regressão de comprimento das raízes sobre número de dias. Discuta o significado biológico dos valores obtidos para o declive da recta e para a ordenada na origem. Comente.
  - Comente a qualidade da recta obtida, calculando o seu coeficiente de determinação, e interpretando o valor obtido.
  - Trace a recta de regressão obtida em cima da nuvem de pontos e comente.
  - Calcule a Soma de Quadrados Total (SQT), a partir do cálculo da variância amostral de  $y$ .
  - Calcule o valor da Soma de Quadrados da Regressão (SQR).
  - Calcule a Soma de Quadrados dos Resíduos (SQRE), directamente a partir dos resíduos, e verifique numericamente a relação fundamental da Regressão Linear:  $SQT=SQR+SQRE$ .
2. O ficheiro *Azeite.xls*, disponível na página *web* da disciplina (secção *material de Apoio*), é um ficheiro de tipo folha de cálculo, comum a aplicações de escritório como o LibreOffice, OpenOffice ou MicrosoftOffice. A folha de cálculo contém dados relativos à produção de azeite em Portugal no período 1995-2010, disponibilizados pelo Instituto Nacional de Estatística ([www.ine.pt](http://www.ine.pt)). As colunas “Azeitona” e “Azeite” correspondem à produção de azeitona oleificada (em t) e azeite (em hl), respectivamente.
- Abra o ficheiro *Azeite.xls* e guarde a folha de cálculo num ficheiro *Azeite.txt* (utilizando o *Save as* com a opção *Ficheiro de Texto*). Coloque esse ficheiro na pasta de trabalho do R.
  - Numa sessão do R, guarde os dados do ficheiro *Azeite.txt* (criado na alínea anterior) numa *data frame* de nome *azeite*, através do comando:
 

```
> azeite <- read.table("Azeite.txt", header=TRUE)
```
  - Crie a nuvem de pontos relacionando as produções de Azeite (eixo vertical, variável  $y$ ) e Azeitona (eixo horizontal, variável  $x$ ).
  - Com base na nuvem de pontos, sugira um valor para o coeficiente de correlação entre as duas variáveis. Avalie a sua sugestão calculando o valor de  $r_{xy}$ . Comente o valor obtido.
  - Calcule as estimativas de mínimos quadrados para os parâmetros da recta de regressão, e comente o seu significado.

- (f) Calcule a precisão da recta de regressão estimada de  $y$  sobre  $x$  e comente o valor obtido.
3. Demonstre as seguintes relações algébricas:
- (a)  $\sum_{i=1}^n (x_i - \bar{x}) = 0$ , para qualquer conjunto de  $n$  valores,  $\{x_i\}_{i=1}^n$ , de média  $\bar{x}$ .
- (b)  $(n-1)\text{cov}_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})y_i = \sum_{i=1}^n (y_i - \bar{y})x_i$ , para quaisquer conjuntos de  $n$  valores,  $\{x_i\}_{i=1}^n$ , e  $\{y_i\}_{i=1}^n$  de médias  $\bar{x}$  e  $\bar{y}$ , respectivamente.
4. Deduza as expressões para o declive e ordenada na origem da recta de regressão, resultantes de minimizar a soma dos quadrados dos resíduos:

$$\begin{aligned} b_1 &= \frac{\text{cov}_{xy}}{s_x^2} \\ b_0 &= \bar{y} - b_1\bar{x} \end{aligned}$$

5. Mostre que, numa Regressão Linear Simples, baseada em  $n$  pares de observações  $\{(x_i, y_i)\}_{i=1}^n$ , se verifica:
- (a) A igualdade da média dos valores observados e da média dos valores ajustados de  $y$ .
- (b) A média dos resíduos ( $e_i = y_i - \hat{y}_i$ ) é nula.
- (c)  $SQT = SQR + SQRE$ , sendo as três Somas de Quadrados definidas como:

$$\begin{aligned} SQT &= \sum_{i=1}^n (y_i - \bar{y})^2 = (n-1) \cdot s_y^2 \\ SQR &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = (n-1) \cdot s_{\hat{y}}^2 \\ SQRE &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 = (n-1) \cdot s_e^2, \end{aligned}$$

onde  $s_{\star}^2$  indica a variância amostral das quantidades representadas por  $\star$ .

- (d)  $SQR = b_1^2 \cdot (n-1) \cdot s_x^2$ , onde  $(n-1) \cdot s_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2$ .
6. Mostre que o declive da recta de regressão de  $y$  sobre  $x$  se pode escrever em termos do desvio padrão de cada variável e do coeficiente de correlação entre as duas variáveis, sendo dado por:

$$b_1 = r_{xy} \cdot \frac{s_y}{s_x}.$$

7. O programa R tem vários conjuntos de dados disponíveis. Um desses conjuntos de dados designa-se **anscombe** e pode ser visto apenas escrevendo o nome do objecto. Utilizando estes dados, determine, e comente os valores obtidos para:
- (a) As médias de cada variável  $x_i$  e  $y_i$  ( $i = 1 : 4$ ).
- (b) As variâncias de cada variável  $x_i$  e  $y_i$  ( $i = 1 : 4$ ).
- (c) O valor dos parâmetros  $b_0$  e  $b_1$  nas quatro rectas de regressão de  $y_i$  sobre  $x_i$  ( $i = 1, 2, 3, 4$ ).
- (d) Os Coeficientes de Determinação associados às quatro rectas indicadas na alínea anterior.

Após comentar os resultados obtidos, construa as quatro nuvens de pontos  $\{(x_i^{(j)}, y_i^{(j)})\}_{i=1}^{11}$ , para  $j = 1 : 4$ . Comente esses gráficos, à luz dos valores anteriormente obtidos.

8. Utilizando os dados das medições morfométricas sobre 150 lírios, contidos no objecto `iris` do R, responda às seguintes questões:
- Construa a nuvem de pontos de Comprimento das Pétalas (eixo horizontal, variável  $x$ ) e Largura das Pétalas (eixo vertical, variável  $y$ ).
  - Ajuste a recta de regressão de Largura ( $y$ ) sobre Comprimento ( $x$ ), e desenhe-a sobre a nuvem de pontos.
  - Ajuste a recta de regressão de Comprimento sobre Largura, mantendo os nomes de  $x$  (Comprimento) e  $y$  (Largura), ou seja, calcule a “recta de  $x$  sobre  $y$ ”.
  - Sobre a nuvem de pontos original, trace agora a recta de regressão de Comprimento sobre Largura - a “recta de  $x$  sobre  $y$ ”. (NOTA: Tenha em atenção que uma equação  $x = b_0 + b_1 y$  tem, na forma canónica, equação  $y = -\frac{b_0}{b_1} + \frac{1}{b_1} x$ ). Verifique que a recta de regressão de  $y$  sobre  $x$  é diferente da recta de regressão de  $x$  sobre  $y$ .
  - Explique o facto de as rectas obtidas nas alíneas anteriores serem diferentes.
9. O programa R tem um grande número de pacotes adicionais disponíveis. Um desses pacotes adicionais designa-se `MASS` e pode ser carregado mediante o comando `library(MASS)`.

Considere o conjunto de dados `Animals`, disponível no referido módulo `MASS`, onde se listam pesos médios dos cérebros (em  $g$ ) e dos corpos (em  $kg$ ) para 28 espécies animais. Pretende-se estudar uma eventual relação entre pesos do cérebro (variável resposta) e pesos do corpo (variável preditora).

- Construa uma nuvem de pontos de pesos do corpo (eixo horizontal) e pesos do cérebro (eixo vertical).
- Construa agora nuvens de pontos com as seguintes transformações de uma ou ambas as variáveis:
  - $\sqrt{y}$  vs.  $x$ ;
  - $\log(y)$  vs.  $x$ ;
  - $y$  vs.  $\log(x)$ ;
  - $\log(y)$  vs.  $\log(x)$ .
- Explicite a relação de base entre as variáveis originais (não logaritmizadas) associada a uma relação linear entre as variáveis logaritmizadas. Comente.

Nas alíneas seguintes considere sempre os *dados logaritmizados*.

- Considere a nuvem de pontos das variáveis logaritmizadas. Identifique os três pontos que se destacam na parte inferior direita da nuvem. (NOTA: explore o comando `identify` do R). Comente.
- Ajuste a recta de regressão de log-peso do cérebro sobre log-peso do corpo (utilizando a totalidade das observações). Trace essa recta sobre a nuvem de pontos e comente.
- Considere agora a estimativa para o declive da recta,  $b_1 = 0.49599$ . Qual o significado biológico deste valor, quer na relação entre variáveis logaritmizadas, quer na relação entre as variáveis originais (não logaritmizadas)?

Nas restantes alíneas, considere apenas os dados (logaritmizados) respeitantes a espécies que *não sejam de dinossáurios*.

- (g) Ajuste a recta de regressão de log-peso do cérebro sobre log-peso do corpo. Trace essa recta sobre a nuvem de pontos e comente. (NOTA: Aproveite a nuvem de pontos anterior, com a totalidade das espécies, para melhor compreender o efeito da exclusão das três espécies de dinossáurios sobre a recta ajustada).
- (h) Considere agora a estimativa para o declive da nova recta,  $b_1 = 0.75226$ . Qual o significado biológico deste valor, quer na relação entre variáveis logaritmizadas, quer na relação entre as variáveis originais (não logaritmizadas)?

Na resolução dos Exercícios seguintes, de natureza inferencial, considere o Modelo da Regressão Linear Simples.

10. Mostre que a média ( $\bar{Y}$ ) das observações de  $Y$ , é uma variável aleatória não correlacionada com o estimador do declive da recta,  $\hat{\beta}_1$ , ou seja, mostre que:

$$\text{cov}(\bar{Y}, \hat{\beta}_1) = 0.$$

11. Mostre que o estimador da ordenada na origem da recta tem a seguinte distribuição:

$$\hat{\beta}_0 \cap \mathcal{N}\left(\beta_0, \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{(n-1) \cdot s_x^2} \right]\right),$$

onde  $(n-1) \cdot s_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2$ .

12. Considere de novo os dados do Exercício 1, admitindo agora que se trata de uma amostra aleatória de raízes da cultivar em questão. Responda às seguintes alíneas.
- (a) Obtenha estimativas das variâncias e desvios padrões associados às estimativas dos parâmetros da recta,  $\beta_0$  e  $\beta_1$ .
- (b) Obtenha um intervalo a 95% de confiança para o declive  $\beta_1$  da correspondente recta populacional.
- (c) Obtenha um intervalo a 95% de confiança para a ordenada na origem  $\beta_0$  da recta populacional.
- (d) Utilize um teste de hipóteses para validar a seguinte afirmação: “por cada dia a mais, a raiz da cultivar cresce, em média, 2mm”.
- (e) Utilize um teste de hipóteses para validar a seguinte afirmação: “por cada dia a mais, a raiz da cultivar cresce, em média, menos do que 2mm”.
- (f) Utilize um teste de hipóteses sobre o declive da recta populacional  $\beta_1$  para validar a seguinte afirmação: “não existe uma relação linear significativa entre dias e comprimento da raiz, para a referida cultivar”.
- (g) Valide de novo a afirmação anterior, mas agora utilizando um teste de ajustamento global do Modelo (teste  $F$ ).
- (h) Para cada uma das seguintes transformações dos dados, verifique o que se altera e o que permanece igual nos resultados anteriores. Comente.
- a contagem do número de dias começa em zero, no dia da primeira observação ( $x \rightarrow x - 1$ ).
  - a contagem do tempo faz-se em número de horas ( $x \rightarrow 24 \cdot x$ ).

- iii. a medida do comprimento da raiz é feita em centímetros ( $y \rightarrow \frac{y}{10}$ ).
- iv. em simultâneo as transformações das alíneas (i) e (iii).

13. Considere os estimadores  $\hat{\beta}_0$  e  $\hat{\beta}_1$  dos parâmetros duma recta de regressão.

- (a) Mostre que a covariância entre os dois estimadores dos parâmetros da recta é dada por:

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = - \frac{\bar{x}\sigma^2}{(n-1) \cdot s_x^2}.$$

- (b) Deduza da alínea anterior que *os estimadores de  $\beta_0$  e de  $\beta_1$  não são, em geral, independentes*. Indique uma condição necessária para que o sejam.

14. A estatística do teste de ajustamento global do modelo (teste  $F$ ) é dada por  $F = \frac{QMR}{QMRE}$ . O Coeficiente de Determinação define-se como  $R^2 = \frac{SQR}{SQT}$ . Com base nestas definições, e tendo em conta as propriedades das somas de quadrados,

- (a) Mostre que a estatística  $F$  se pode escrever também como:

$$F = (n-2) \cdot \frac{R^2}{1-R^2}$$

- (b) Verifique, a partir da expressão anterior, que a estatística  $F$  é (para  $n$  fixo) uma *função crescente do Coeficiente de Determinação*. Interprete esse facto, em termos do significado de  $R^2$  e a natureza do teste de ajustamento global.

15. Mostre que, numa Regressão Linear Simples, a estatística  $F = \frac{QMR}{QMRE}$  do teste de ajustamento global é o quadrado da estatística  $T = \frac{\hat{\beta}_1}{\sqrt{\frac{QMRE}{(n-1) \cdot s_x^2}}}$  do teste  $t$  para a hipótese  $H_0 : \beta_1 = 0$ . Tendo

em conta os resultados dados na disciplina de Estatística (dos 1<sup>os</sup> ciclos do ISA), relacionando as distribuições  $t$  e  $F$ , conclua que, numa Regressão Linear Simples, estes dois testes são equivalentes.

16. Considere os dados do Exercício 9. Trabalhe sempre com os *dados logaritmizados*, para a totalidade das espécies.

- (a) Considere a presença de erros aleatórios na relação linear entre as variáveis logaritmizadas:  $\log(Y) = \beta_0 + \beta_1 \log(x) + \epsilon$ . Qual a consequência para a relação entre as variáveis originais (não logaritmizadas) associada à presença dos erros aleatórios? E como se traduzem os restantes pressupostos do Modelo de Regressão Linear (normalidade, homogeneidade de variâncias, independência) em termos dessa relação entre as variáveis originais (não logaritmizadas)?

- (b) Analise os principais resultados associados à regressão ajustada na alínea 9e). Considere em particular os valores do Coeficiente de Determinação, e os resultados do teste  $F$  de ajustamento global. Como se explica:

- i. que o valor do Coeficiente de Determinação não seja particularmente bom, quando o teste  $F$  sugere que a rejeição da hipótese nula do teste de ajustamento é muito enfática?
- ii. que o Coeficiente de Determinação não seja particularmente elevado, sendo evidente a partir da nuvem de pontos que existe uma boa relação linear entre log-peso do corpo e log-peso do cérebro para a generalidade das espécies?

- (c) Construa um intervalo de confiança a 95% para o declive da recta que relaciona log-peso do corpo e log-peso do cérebro. É admissível falar-se numa relação isométrica entre peso do corpo e peso do cérebro?

- (d) Estude os gráficos dos resíduos para detectar a existência de eventuais problemas com os pressupostos do modelo. Em particular, veja como a presença das três espécies de natureza diferente das restantes está a afectar estes gráficos.

Nas restantes alíneas, considere apenas os dados (logaritmizados) respeitantes a espécies que *não sejam de dinossáurios*.

- (e) Analise os principais resultados associados à regressão ajustada na alínea 9g). Compare com os resultados obtidos na alínea 9e) e comente. Em particular, como se explica a elevação considerável no valor do Coeficiente de Determinação?
- (f) Construa um intervalo de confiança a 95% para o declive da recta que relaciona log-peso do corpo e log-peso do cérebro. Perante o novo valor de  $b_1$ , será agora admissível falar-se numa relação isométrica entre peso do corpo e peso do cérebro?
- (g) Preveja o valor esperado do log-peso do cérebro para espécies com peso de corpo igual a 250kg. Construa um intervalo de confiança para esse valor esperado.
- (h) Construa um intervalo de predição associado ao peso do cérebro duma espécie cujo peso do corpo seja 250kg. Compare com o intervalo de confiança obtido na alínea anterior e comente.
- (i) Estude os gráficos dos resíduos para detectar a existência de eventuais problemas com os pressupostos do modelo. Comente as suas conclusões, tendo presente os gráficos análogos obtidos com a presença das 3 espécies de dinossáurios.
17. Considere agora o conjunto de dados relativos a 62 espécies de mamíferos, que é dado no objecto *mammals* do pacote *MASS*, e cuja natureza é semelhante aos dados do Exercício 9.
- (a) Construa uma nuvem de pontos de pesos do corpo (eixo horizontal) e pesos do cérebro (eixo vertical).
- (b) Tendo em vista uma relação alométrica entre as duas variáveis, construa agora uma segunda nuvem de pontos, desta vez entre os *logaritmos* de cada variável. Comente os dois gráficos.
- (c) Explícite a relação de base entre as variáveis originais (não logaritmizadas) associada a uma relação linear entre as variáveis logaritmizadas. Comente.

Nas alíneas seguintes considerar os *dados logaritmizados*, para a totalidade das espécies.

- (d) Ajuste a recta de regressão de log-peso do cérebro sobre log-peso do corpo. Trace essa recta sobre a nuvem de pontos e comente.
- (e) Analise os principais resultados associados à regressão ajustada. Considere em particular os valores do Coeficiente de Determinação, e os resultados do teste  $F$  de ajustamento global.
- (f) Qual o significado biológico da estimativa do declive da recta, quer na relação entre variáveis logaritmizadas, quer na relação entre as variáveis originais (não logaritmizadas)?
- (g) Construa um intervalo de confiança a 95% para o declive da recta que relaciona log-peso do corpo e log-peso do cérebro. Será agora admissível falar-se numa relação isométrica entre peso do corpo e peso do cérebro?
- (h) Estude os gráficos dos resíduos para detectar a existência de eventuais problemas com os pressupostos do modelo.
18. Dado o Modelo de Regressão Linear Simples, considere o estimador do valor esperado de  $Y$ , associado a  $X = x$ , ou seja, o estimador  $\hat{\mu}_{Y|x} = \hat{\beta}_0 + \hat{\beta}_1 x$ . Mostre que a sua variância é  $V[\hat{\mu}_{Y|x}] = \sigma^2 \left[ \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1) \cdot s_x^2} \right]$ . NOTA: Tenha em atenção o Exercício 13.

19. No contexto do Modelo de Regressão Linear Simples,

- (a) Mostre que a covariância entre valores observados e ajustados da variável resposta é dada por  $cov(Y_i, \hat{Y}_i) = \sigma^2 \left[ \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1) \cdot s_x^2} \right]$ . SUGESTÃO: Recorde que  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ .
- (b) Calcule a covariância entre cada observação de  $Y$  e o respectivo resíduo, ou seja,  $cov(Y_i, E_i)$ .
- (c) Mostre que a covariância entre cada valor ajustado de  $Y$  e o respectivo resíduo é nula, ou seja, mostre que  $cov(\hat{Y}_i, E_i) = 0, \forall i = 1, \dots, n$ . Com base neste resultado, justifique a utilização do gráfico de resíduos vs. valores ajustados de  $Y$  para estudar o comportamento dos resíduos (em vez de, por exemplo, o gráfico de resíduos vs. valores observados de  $Y$ ).
- (d) Com o auxílio dos resultados anteriores, mostre que os resíduos têm a distribuição indicada nas aulas teóricas, ou seja, mostre que  $E_i \sim \mathcal{N}(0, \sigma^2 \cdot (1 - h_{ii}))$ , onde  $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1) \cdot s_x^2}$ .

20. No contexto da Regressão Linear Simples,

- (a) Determine o valor esperado da soma de quadrados associada à Regressão ( $SQR$ ).  
**SUGESTÃO:** Utilize a fórmula para  $SQR$  obtida na última alínea do Exercício 5.
- (b) Compare os valores esperados dos quadrados médios  $QMR$  e  $QMRE$ . Com base nessa comparação, justifique a natureza unilateral direita da região de rejeição associada ao teste de ajustamento global, cuja estatística de teste é  $F = \frac{QMR}{QMRE}$ .