

## Exercícios - Estatística e Delineamento - 2022-23

### 1 Regressão Linear Simples

1. Com base nos dados do Instituto Nacional de Estatística (INE), foi criado um ficheiro em formato CSV (*Comma separated values*) chamado `cereais.csv` e contendo a evolução da superfície agrícola utilizada anualmente na produção de cereais para grão (variável `area`, em  $\text{km}^2$ ) em Portugal, no período de 1986 a 2011 (variável `ano`). O ficheiro encontra-se disponível na página *web* da disciplina (secção *Materiais de apoio*, subsecção *Dados*). O ficheiro `cereais.csv` deve ser descarregado e guardado na directoria onde se localiza a sessão de trabalho do R, que convém ser a pasta (de preferência com o nome `AulasED`) onde tem estado a guardar o seu trabalho. Seguidamente o seu conteúdo deve ser lido para a sessão do R através do comando:

```
> Cereais <- read.csv("cereais.csv")
```

**Atenção:** Ao guardar o ficheiro na directoria da sua sessão de trabalho, não deve alterar o tipo (CSV) de ficheiro. A melhor forma de garantir isso será o de apenas *guardar* o ficheiro, sem o abrir. Se preferir abrir o ficheiro antes de o guardar, certifique-se de que está a utilizar um editor de texto que não altera o conteúdo do ficheiro. Os programas do tipo *Office* não oferecem essa garantia, pelo que se recomenda evitar a sua utilização.

- (a) Construa uma nuvem de pontos de superfície agrícola *vs.* ano e comente.
  - (b) A partir do gráfico obtido na alínea anterior, sugira um valor para o coeficiente de correlação entre superfície agrícola e ano. Depois, utilize os comandos do R para calcular esse mesmo coeficiente de correlação. Comente o seu significado.
  - (c) Ajuste uma recta de regressão de superfície agrícola utilizada sobre anos. Discuta o significado dos parâmetros da recta ajustada, no contexto do problema sob estudo.
  - (d) Comente a qualidade da recta obtida, calculando o respectivo coeficiente de determinação e interpretando o valor obtido.
  - (e) Trace a recta de regressão ajustada em cima da nuvem de pontos e comente.
  - (f) Calcule a Soma de Quadrados Total (SQT), a partir do cálculo da variância amostral de  $y$ .
  - (g) Calcule o valor da Soma de Quadrados da Regressão (SQR).
  - (h) Calcule a Soma de Quadrados dos Resíduos (SQRE), directamente a partir dos resíduos, e verifique numericamente a relação fundamental da Regressão Linear:  $\text{SQT} = \text{SQR} + \text{SQRE}$ .
  - (i) Altere as unidades de medida da variável `area`, de  $\text{km}^2$  para hectares (`area`  $\rightarrow$  `area` $\times$ 100). Ajuste novamente a regressão, após efectuar esta alteração. O que aconteceu aos parâmetros estimados e ao coeficiente de determinação  $R^2$ ? Comente.
  - (j) De novo a partir dos dados originais, transforme a variável `ano` num contador dos anos do estudo (`ano`  $\rightarrow$  `ano`-1985). Ajuste novamente a regressão, após efectuar esta alteração. O que aconteceu aos parâmetros estimados e ao coeficiente de determinação  $R^2$ ? Comente.
2. O ficheiro `azeite.xls`, disponível na página *web* da disciplina (secção *Materiais de apoio*, subsecção *Dados*), 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.

- (a) Abra o ficheiro `azeite.xls` com um programa do tipo *Office* e guarde a folha de cálculo num ficheiro de texto de nome `azeite.txt`, utilizando o *Save as* com a opção *Ficheiro de Texto*. Coloque esse ficheiro na pasta de trabalho do R.
- (b) 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)
```
- (c) Crie a nuvem de pontos relacionando as produções de Azeite (eixo vertical, variável  $y$ ) e Azeitona (eixo horizontal, variável  $x$ ).
- (d) 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.
- (e) 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  $n$  pares de valores,  $\{(x_i, y_i)\}_{i=1}^n$ , de médias  $\bar{x}$  e  $\bar{y}$ , respectivamente.
- (c)  $(n-1)s_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x})x_i$ , para qualquer conjunto de  $n$  valores,  $\{x_i\}_{i=1}^n$ , de média  $\bar{x}$ .

4. [Opcional] 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 verificam:

- (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) as três Somas de Quadrados da regressão são múltiplos de variâncias:

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

onde  $s_x^2$  indica a variância amostral das quantidades representadas por  $\star$ .

(d)  $SQR = b_1^2 \cdot (n - 1) \cdot s_x^2$ , onde  $(n - 1) \cdot s_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2$ .

(e)  $SQT = SQR + SQRE$ .

6. Considere uma regressão linear simples, ajustada com  $n$  pares de observações  $\{(x_i, y_i)\}_{i=1}^n$ . Mostre que:

(a) O declive da recta de regressão de  $y$  sobre  $x$  pode-se escrever em termos do desvio padrão de cada variável e do coeficiente de correlação entre as duas variáveis, sendo dado por:

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

(b) O coeficiente de determinação  $R^2$  é igual ao quadrado do coeficiente de correlação entre as observações da variável preditora  $x$  e da variável resposta  $y$ .

(c) O quadrado do coeficiente de correlação entre os  $n$  valores observados  $y_i$  e os  $n$  correspondentes valores ajustados,  $\hat{y}_i$ , é também igual ao coeficiente de determinação:  $(r_{y\hat{y}})^2 = R^2$ .

7. O programa R tem um grande número de pacotes adicionais disponíveis. Um desses pacotes adicionais designa-se MASS. 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 de animais terrestres. Pretende-se estudar uma relação entre pesos do cérebro (variável resposta,  $y$ ) e pesos do corpo (variável preditora,  $x$ ).

(a) Construa uma nuvem de pontos de pesos do corpo (eixo horizontal) e pesos do cérebro (eixo vertical). Calcule o coeficiente de correlação correspondente e comente.

(b) Construa nuvens de pontos com as seguintes transformações de uma ou ambas as variáveis:

i.  $\ln(y)$  vs.  $x$ ;

ii.  $y$  vs.  $\ln(x)$ ;

iii.  $\ln(y)$  vs.  $\ln(x)$ .

(c) Considere uma relação linear entre  $\ln(y)$  e  $\ln(x)$ . Explícite a relação de base correspondente entre as variáveis originais (não logaritmizadas). Comente.

Nas alíneas seguintes considere sempre os *dados logaritmizados*.

(d) Calcule os coeficientes de correlação e de determinação associados à relação entre  $\ln(x)$  e  $\ln(y)$ . Interprete os valores obtidos. Como se explica 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?

(e) 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.

(f) 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)?

(g) 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.

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

- (h) 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).
  - (i) 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)?
8. 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. Essas observações encontram-se num ficheiro em formato `csv` de nome `ozono.csv`, que se encontra disponível na página *web* da disciplina e pode ser descarregado para a directoria de trabalho da sessão do R, como indicado no Exercício 1. Seguidamente, o conteúdo desse ficheiro deve ser lido para dentro da sessão do R e armazenado num objecto de nome `ozono`, através do comando `read.csv`:

```
> ozono <- read.csv("ozono.csv")
```

- (a) Construa a nuvem de pontos de ozono (eixo vertical) *vs.* temperatura máxima (eixo horizontal).
- (b) Tendo em conta a curvatura observada no gráfico, foi sugerido o ajustamento dum modelo exponencial, da forma  $y = a e^{bx}$ .
  - i. Construa a nuvem de pontos com as transformações adequadas para verificar se o modelo exponencial é, efectivamente, uma boa opção.
  - ii. Ajuste o modelo *linearizado* recorrendo ao comando `lm` do R. Determine o respectivo coeficiente de determinação e comente.
  - iii. Interprete os parâmetros da recta que ajustou, directamente em termos do modelo exponencial.
  - iv. 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.
- (c) Considere novamente a nuvem de pontos original. Trace a curva exponencial correspondente ao ajustamento efectuado na alínea anterior.

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

9. Considere os estimadores  $\hat{\beta}_0$  e  $\hat{\beta}_1$  dos parâmetros duma recta de regressão.
- (a) 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:

$$\text{Cov}[\bar{Y}, \hat{\beta}_1] = 0.$$

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

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

- (c) 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 possam ser.

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

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

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

11. Considere os dados das medições morfométricas sobre 150 lírios, disponível em qualquer distribuição do  $\mathbb{R}$  e contidos no objecto `iris`. Admita que se trata duma amostra extraída aleatoriamente duma população mais vasta. Considere, em particular, a relação entre largura da pétala (`Petal.Width`, variável  $y$ ) e comprimento da pétala (`Petal.Length`, variável  $x$ ), ambas em *cm*. Responda às seguintes alíneas.

- 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 ( $x$ ) das pétalas, e desenhe-a sobre a nuvem de pontos.
- Obtenha estimativas das variâncias e dos desvios padrão dos estimadores dos parâmetros da recta populacional,  $\beta_0$  e  $\beta_1$ .
- Obtenha um intervalo a 95% de confiança para o declive  $\beta_1$  da correspondente recta populacional.
- Obtenha um intervalo a 95% de confiança para a ordenada na origem  $\beta_0$  da recta populacional.
- Utilize um teste de hipóteses para validar a seguinte afirmação: “por cada centímetro a mais no comprimento da pétala, a largura da pétala cresce, em média, 0.5cm”.
- Utilize um teste de hipóteses para validar a seguinte afirmação: “por cada centímetro a mais no comprimento da pétala, a largura da pétala cresce, em média, menos de 0.5cm”.
- 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 comprimentos e larguras das pétalas, nos lírios”.
- Comente de novo a afirmação anterior, mas agora utilizando um teste de ajustamento global do Modelo (teste  $F$ ).
- Estime o valor esperado da largura da pétala para lírios cuja pétala tenha comprimento 4.5cm. Construa um intervalo de confiança para esse valor esperado.
- Construa um intervalo de predição (95%) associado à largura duma pétala cujo comprimento seja 4.5cm. Compare com o intervalo de confiança obtido na alínea anterior e comente.
- Estude os gráficos dos resíduos para detectar a existência de eventuais problemas com os pressupostos do modelo. Comente as suas conclusões.
- Para cada uma das seguintes transformações dos dados, verifique os efeitos sobre os parâmetros ajustados e sobre o coeficiente de determinação. Comente.
  - os comprimentos das pétalas são dados em milímetros ( $x \rightarrow 10 \times x$ ), mantendo-se as larguras ( $y$ ) em centímetros.
  - as larguras das pétalas são dadas em milímetros ( $y \rightarrow 10 \times y$ ), mantendo-se os comprimentos ( $x$ ) em centímetros.

- iii. em simultâneo, larguras e comprimentos das pétalas são expressas em milímetros ( $x \rightarrow 10 \times x$  e  $y \rightarrow 10 \times y$ ).
12. 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.
13. A medição rigorosa de áreas foliares faz-se através de técnicas destrutivas. Deseja-se obter um modelo que permita estimar áreas foliares de castas de videiras, utilizando uma variável preditora que possa ser medidas sem arrancar as folhas da videira. Concretamente, deseja-se prever as áreas foliares a partir de medições do comprimento da nervura principal, e deseja-se obter uma relação que possa ser aplicada a diferentes castas.

Na Secção de Horticultura do ISA foram seleccionadas aleatoriamente 200 folhas de cada uma de três castas: Fernão Pires, Vital e Água Santa. Em cada folha mediu-se a área foliar (variável (**Área**), em  $cm^2$ ) e comprimento da nervura principal da folha (variável (**NP**), em  $cm$ ). Os dados estão no objecto **videiras**. Responda às seguintes alíneas para o conjunto dos 600 pares de observações, não distinguindo as castas.

- (a) Construa a nuvem de pontos das áreas foliares e comprimentos de nervura principal. Calcule o respectivo coeficiente de correlação linear amostral. Comente.
- (b) Considere as transformações logarítmicas de ambas as variáveis, e repita a alínea anterior usando os dados logaritmizados.
- (c) Ajuste a recta de regressão de log-áreas sobre log-comprimentos da nervura principal e trace-a sobre a nuvem de pontos obtida na alínea anterior. Comente.
- (d) Qual a relação não linear entre **Area** e **NP** que corresponde à recta ajustada na alínea anterior? Usando o comando `curve`, trace-a sobre a nuvem de pontos construída na alínea 13a. Comente.
- (e) Construa um intervalo a 95% de confiança para o declive da recta de regressão populacional relacionando as variáveis log-transformadas. Interprete o significado biológico do resultado obtido.
- (f) Um investigador comenta que, sendo a área foliar uma medida bidimensional e o comprimento da nervura unidimensional, seria plausível que a área seja proporcional ao quadrado do comprimento. Efectue um teste de hipóteses para avaliar a validade desta tese.
- (g) Valide os pressupostos do modelo de regressão linear ajustado, analisando o comportamento dos resíduos e restantes ferramentas de diagnóstico. Comente. Considere agora a regressão linear entre as variáveis originais (não logaritmizadas) e repita o estudo dos resíduos. Comente.
14. 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.

15. Um estudo realizado por uma equipa do ISA visou caracterizar a relação existente entre um índice de vegetação, calculado com base em medições dum aparelho portátil, e a Produtividade Primária Bruta (PPB), medida em micromoles por metro quadrado, por segundo ( $\mu \text{ mole } m^{-2} s^{-1}$ ) em comunidades herbáceas mediterrânicas de Portugal. O índice de vegetação usado é o índice NDWI, um índice adimensional que toma valores entre  $-1$  e  $1$  (e que é definido com base na reflectância nas bandas do verde e do infra-vermelho próximo). Recolheram-se 91 pares de observações, com os seguintes indicadores:

| Variável | Mínimo   | Máximo  | Média   | Variância   |
|----------|----------|---------|---------|-------------|
| NDWI     | -0.18446 | 0.19154 | 0.03286 | 0.007910756 |
| PPB      | 7.173    | 33.966  | 19.715  | 54.4267635  |

Após alguma análise, optou-se por ajustar uma regressão linear simples do logaritmo (natural) da Produtividade Primária Bruta sobre os valores do índice NDWI, com os seguintes resultados:

```
> summary(lm(log(ppb) ~ ndwi, data=gpp))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.77400    0.02872   96.6  <2e-16
ndwi         3.83488    0.30432   12.6  <2e-16
---
Residual standard error: 0.2568 on 89 degrees of freedom
Multiple R-squared:  0.6408,  Adjusted R-squared:  0.6368
F-statistic: 158.8 on 1 and 89 DF,  p-value: < 2.2e-16
```

- Dado o modelo ajustado, será admissível considerar que, a um aumento de uma unidade no índice NDWI corresponde, em média, um aumento de 4 unidades na log-Produtividade Primária Bruta? Justifique através dum teste de hipóteses adequado.
  - Calcule o intervalo a 95% de confiança para a ordenada na origem da recta populacional. Comente.
  - Um intervalo de predição (95%) para uma observação individual de log-PPB, quando o índice NDWI toma o valor 0.1, é da forma  $[2.64287, ???]$ . Diga justificando,
    - qual o valor central desse intervalo de predição;
    - qual o extremo direito do intervalo de predição;
    - qual a estimativa  $\hat{\sigma}_{indiv}$  da variabilidade associada a uma observação individual de  $Y$  quando  $x=0.1$ ? E qual o erro padrão do estimador do valor esperado de log-PPB, quando o índice NDWI é 0.1, ou seja, o erro padrão de  $\hat{\mu}_{Y|X=0.1}$ ?
    - indique um intervalo de predição para a Produtividade Primária Bruta (PPB, não logaritmizado) quando o índice NDWI toma o valor 0.1.
  - A que tipo de relação não linear entre a Produtividade Primária Bruta e o índice NDWI corresponde a regressão linear acima ajustada? Calcule a equação da curva ajustada, relacionando PPB e NDWI. Qual o valor estimado da taxa de variação relativa da Produtividade Primária Bruta, face aos valores de NDWI?
  - É possível ajustar um modelo potência para relacionar PPB e NDWI, com base numa regressão linear simples? Justifique a sua resposta.
16. Considere os dados do Exercício 7 (Animals). 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 em termos dessa relação entre as variáveis originais (não logaritmizadas)?
- (b) Efectue um teste de ajustamento global da regressão de log-pesos do cérebro sobre log-pesos do corpo. Como se explica que o teste  $F$  rejeite enfaticamente a hipótese nula do teste, quando o valor do coeficiente de determinação não é particularmente elevado?
- (c) 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*.

- (d) 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?
  - (e) 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.
  - (f) Construa um intervalo de predição associado ao log-peso do cérebro duma espécie cujo peso do corpo seja 250kg. Como obter um intervalo de predição associado *ao peso* do cérebro?
  - (g) 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 7.
- (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 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 9.
19. No contexto do Modelo de Regressão Linear Simples,
- Mostre que a covariância entre o  $i$ -ésimo valor observado e o correspondente valor ajustado da variável resposta é dada por  $Cov[Y_i, \hat{Y}_i] = \sigma^2 h_{ii}$ , onde  $h_{ii} = \left[ \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1) \cdot s_x^2} \right]$  (que é também o efeito alavanca da  $i$ -ésima observação). SUGESTÃO: Recorde que  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ .
  - Calcule a covariância entre cada observação de  $Y$  e o respectivo resíduo, ou seja,  $Cov[Y_i, E_i]$ .
  - 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  $Y$ ).
  - Com o auxílio dos resultados anteriores, mostre que os resíduos têm a distribuição indicada nas aulas teóricas, ou seja, mostre que  $E_i \sim \mathcal{N}(0, \sigma^2(1 - h_{ii}))$ .
20. No contexto da Regressão Linear Simples,
- Determine o valor esperado da soma de quadrados associada à Regressão ( $SQR$ ).  
**SUGESTÃO:** Utilize a fórmula para  $SQR$  obtida na alínea d) do Exercício 5.
  - 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  $F = \frac{QMR}{QMRE}$ .