
INSTITUTO SUPERIOR DE AGRONOMIA
ESTATÍSTICA E DELINEAMENTO – 2011/12
Resoluções de exercícios de Regressão Linear Simples

1. Os comandos do R pedidos nas alíneas iniciais são:

- (a) `> x <- c(1, 7, 13, 20, 27, 34, 62)`
`> y <- c(5, 10, 12, 29, 36, 83, 102)`
- (b) `> raiz <- data.frame(x,y)`

Para ver o conteúdo do objecto acabado de criar, escrevemos o seu nome:

```
> raiz
  x  y
1  1  5
2  7 10
3 13 12
4 20 29
5 27 36
6 34 83
7 62 102
```

NOTA: As variáveis individuais da *data frame* podem ser acedidas através duma indexação análoga à utilizada para objectos de tipo matriz:

```
> raiz[,1]
[1]  1  7 13 20 27 34 62
> raiz[,2]
[1]  5 10 12 29 36 83 102
```

Alternativamente, as variáveis que compõem uma *data frame* podem ser acedidas através do nome da *data frame*, seguido dum cifrão e do nome da variável:

```
> raiz$x
[1]  1  7 13 20 27 34 62
> raiz$y
[1]  5 10 12 29 36 83 102
```

- (c) `> plot(raiz)`
Repare-se que este comando funciona porque: (i) a *data frame* `raiz` apenas tem duas variáveis; e (ii) a ordem dessas variáveis coincide com a ordem desejada no gráfico: a primeira variável no eixo horizontal e a segunda no eixo vertical.
- (d) Os parâmetros da recta podem ser calculados, quer a partir da sua definição, quer utilizando o comando do R que efectua o ajustamento duma regressão linear: o comando `lm` (as iniciais, pela ordem em inglês, de *modelo linear*). Sabemos que:

$$b_1 = \frac{cov_{xy}}{s_x^2} \quad \text{e} \quad b_0 = \bar{y} - b_1 \bar{x} .$$

Utilizando o R, é possível calcular cada um dos indicadores estatísticos presentes nas definições:

```

> cov(raiz$x, raiz$y)
[1] 739.881
> var(raiz$x)
[1] 417.6190
> 739.881/417.6190
[1] 1.771665
> mean(y)
[1] 39.57143
> mean(x)
[1] 23.42857
> 39.57143 - 1.771665*23.42857
[1] -1.936147

```

Alternativamente, o comando `lm` devolve directamente os parâmetros da recta de regressão:

```

> lm(y ~ x, data=raiz)
Call: lm(formula = y ~ x, data = raiz)
Coefficients:
(Intercept)          x
      -1.936         1.772

```

NOTA: A fórmula $y \sim x$ indica que a variável do lado esquerdo do til é a variável resposta, e a do lado direito é a variável preditora. O argumento `data` permite indicar o objecto onde se encontram as variáveis cujos nomes são referidos na formula. Sem este argumento, e caso as variáveis referidas apenas se encontrem no interior do objecto, não serão identificadas pelo R, que só pesquisa nomes de objectos (e não, por exemplo, variáveis no interior de *data frames*).

Interpretação dos coeficientes:

- Declive estimado: $b_1 = 1.772 \text{ mm/dia}$. Em geral (e como se pode comprovar analisando a fórmula para o declive da recta de regressão), as unidades de b_1 são as unidades da variável resposta y a dividir pelas unidades da variável preditora x . Neste problema, o declive indica que a *variação média no comprimento da raiz (variável resposta), por cada dia que passa (unidade da variável preditora) é de 1.772mm/dia*. Fala-se em “variação média” porque a recta apenas descreve a tendência de fundo, na relação entre x e y .
 - Ordenada na origem estimada: $b_0 = -1.936 \text{ mm}$. Em geral, as unidades de b_0 são as unidades da variável resposta y . A interpretação deste valor é, neste caso, estranha: o comprimento médio das raízes no dia $x = 0$, para o qual um valor negativo é fisicamente impossível. A impossibilidade sugere que, pelo menos na proximidade de $x = 0$, o modelo ajustado tem problemas (não deixando de ser o melhor modelo linear, para os dados em questão).
- (e) Sabe-se que, numa regressão linear simples entre variáveis x e y , o coeficiente de determinação é o quadrado do coeficiente de correlação entre as variáveis, ou seja: $R^2 = r_{xy}^2$. O valor do coeficiente de correlação entre x e y pode ser obtido através do comando `cor`:

```

> cor(raiz$x, raiz$y)
[1] 0.9487426
> cor(raiz$x, raiz$y)^2
[1] 0.9001125

```

No nosso caso $R^2 = 0.9001125$, ou seja, cerca de 90% da variabilidade total observada para a variável resposta y é explicada pela regressão.

-
- (f) O comando `abline(lm(y ~ x, data=raiz))` traça a recta pedida em cima do gráfico anteriormente criado pelo comando `plot`.

Nota: Em geral, o comando `abline(a,b)` traça, num gráfico já criado, a recta de equação $y = a + bx$. No caso do *input* ser o ajustamento dum regressão linear simples (obtido através do comando `lm`, que devolve o par de coeficientes b_0 e b_1), o resultado é o gráfico da recta $y = b_0 + b_1 x$.

- (g) Sabemos que $SQT = (n - 1) s_y^2$, pelo que podemos calcular este valor através do comando:

```
> 6*var(raiz$y)
[1] 8737.714
```

- (h) Sabemos que $R^2 = \frac{SQR}{SQT}$, pelo que $SQR = R^2 \times SQT$:

```
> 0.9001125*8737.714
[1] 7864.926
```

Alternativamente, e uma vez que $SQR = (n - 1) s_{\hat{y}}^2$, pode-se usar o comando `fitted` para obter os valores ajustados de y (\hat{y}_i) e seguidamente obter o valor de SQR :

```
> fitted(lm(y ~ x, data=raiz))
      1          2          3          4          5          6
-0.1644812 10.4655074 21.0954960 33.4971494 45.8988027 58.3004561
      7
107.9070696
> 6*var(fitted(lm(y ~ x, data=raiz)))
[1] 7864.926
```

- (i) O comando `residuals` devolve os resíduos dum modelo ajustado. Logo,

```
> residuals(lm(y ~ x, data=raiz))
      1          2          3          4          5          6          7
 5.1644812 -0.4655074 -9.0954960 -4.4971494 -9.8988027 24.6995439 -5.9070696
> sum(residuals(lm(y ~ x, data=raiz))^2)
[1] 872.7882
```

É fácil de verificar que se tem $SQR + SQRE = SQT$:

```
> 7864.926+872.7882
[1] 8737.714
```

2. Tem-se:

- (a) Seguir as instruções do enunciado, criando o ficheiro de texto `Azeite.txt` na directoria onde temos a sessão de trabalho do R.
- (b) O comando de leitura, a partir da sessão do R, é:

```
> azeite <- read.table("Azeite.txt", header=TRUE)
```

Caso o ficheiro `Azeite.txt` esteja numa directoria diferente da directoria de trabalho do R, o nome do ficheiro deverá incluir a sequência de pastas e subpastas que devem ser percorridas para chegar até ao ficheiro.

NOTA: O argumento `header` tem valor lógico que indica se a primeira linha do ficheiro a ser lido contém, ou não, os nomes das variáveis. Por omissão o argumento tem o valor lógico `FALSE`, que considera que na primeira linha do ficheiro já há valores numéricos. Como no

ficheiro `Azeite.txt` a primeira linha contém os nomes das variáveis, foi necessário indicar explicitamente o valor lógico `TRUE`.

O resultado do comando pode ser visto escrevendo o nome do objecto agora lido:

```
> azeite
  Ano Azeitona Azeite
1 1995   311257 477728
2 1996   275143 452038
3 1997   309090 423584
4 1998   225616 360948
5 1999   320865 512264
6 2000   167161 249433
7 2001   218522 349502
8 2002   211574 310474
9 2003   232947 364976
10 2004   300699 500658
11 2005   203909 318174
12 2006   362301 518466
13 2007   203968 352574
14 2008   336479 587422
15 2009   414687 681850
16 2010   435009 686832
```

- (c) Quando aplicado a uma *data frame*, o comando `plot` produz uma “matriz de gráficos” de cada possível par de variáveis (confirme!). Neste caso, não é pedido qualquer gráfico envolvendo a primeira variável da *data frame*. Existem várias maneiras alternativas de pedir apenas o gráfico das segunda e terceira variáveis, uma das quais envolve o conceito de *indexação negativa*, que tanto pode ser utilizado em *data frames* como em matrizes: índices negativos representam linhas ou colunas a serem *omitidas*. Assim, os comandos seguintes produzem todos o gráfico pedido no enunciado:

```
> plot(azeite[,-1])
> plot(azeite[,c(2,3)])
> plot(azeite$Azeitona, azeite$Azeite)
```

- (d) O comando `cor` do R calcula coeficientes de correlação. Pode ser aplicado a dois vectores (de igual comprimento), ou directamente a uma *data frame*. Neste último caso, calcula a *matriz de correlações* entre todos os pares de variáveis da *data frame*. Estas matrizes são sempre necessariamente simétricas:

```
> cor(azeite)
           Ano Azeitona  Azeite
Ano      1.0000000 0.3999257 0.4715217
Azeitona 0.3999257 1.0000000 0.9722528
Azeite   0.4715217 0.9722528 1.0000000
```

O valor da correlação pedido é $r_{xy} = 0.9722528$, um valor muito elevado, que indica uma relação linear crescente muito forte, entre produção de azeitona e produção de azeite.

- (e) Utilizando o comando `lm` do R, tem-se:

```
> lm(Azeite ~ Azeitona, data=azeite)
Call: lm(formula = Azeite ~ Azeitona, data = azeite)
Coefficients:
(Intercept)      Azeitona
-5151.793         1.596
```

Por cada tonelada adicional de produção de azeitona oleificada, há um aumento médio de 1.596hl de produção de azeite. De novo, o valor da ordenada na origem é impossível: indica que, na ausência de produção de azeitona, a produção média de azeite seria negativa ($b_0 = -5151.793hl$). O modelo não deve ser utilizado (nem tal faria sentido) para produções de azeitona próximas de zero. Da mesma forma, deve ser usado com muito cuidado fora da gama de valores observados de x .

- (f) A precisão da recta é uma designação alternativa para o coeficiente de determinação R^2 . Sabe-se que, numa regressão linear simples, $R^2 = r_{xy}^2$. Logo, e tendo em conta os resultados já obtidos, a forma mais fácil de calcular R^2 é $R^2 = 0.9722528^2 = 0.9452755$. Assim, cerca de 94.5% da variabilidade na produção de azeite é explicável pela regressão linear simples sobre a produção de azeitona.

3. Tem-se:

(a)
$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} = n\bar{x} - n\bar{x} = 0.$$

- (b) Por definição, $(n-1)cov_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$. Distribuindo o primeiro factor de cada parcela pelas parcelas do segundo factor e utilizando o resultado da alínea anterior, temos:

$$(n-1)cov_{xy} = \sum_{i=1}^n (x_i - \bar{x})y_i - \sum_{i=1}^n (x_i - \bar{x})\bar{y} = \sum_{i=1}^n (x_i - \bar{x})y_i - \bar{y} \sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x})y_i$$

Trocando o papel das variáveis x e y , mostra-se que $(n-1)cov_{xy} = \sum_{i=1}^n x_i(y_i - \bar{y})$.

4. Este exercício está resolvido nas pgs. 28-29 das folhas de Estatística Descritiva da Prof. Manuela Neves (<http://www.isa.utl.pt/dm/estat/estat/seb1.pdf>), relativas à disciplina de Estatística dos primeiros ciclos do ISA (*web page* da disciplina em <http://www.isa.utl.pt/dm/estat/estat/estat.html>).

5. (a) Tendo em conta que os valores ajustados de y são dados por $\hat{y}_i = b_0 + b_1 x_i$, tem-se que a média dos valores ajustados é dada por:

$$\frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} \sum_{i=1}^n (b_0 + b_1 x_i) = \frac{1}{n} \sum_{i=1}^n b_0 + \frac{1}{n} \sum_{i=1}^n b_1 x_i = b_0 + b_1 \bar{x}.$$

Mas a ordenada de origem duma recta de regressão é dada por $b_0 = \bar{y} - b_1 \bar{x}$, pelo que a última expressão equivale à média \bar{y} dos valores observados de y .

- (b) Tem-se, por definição, que $e_i = y_i - \hat{y}_i$. Logo (e tendo em conta a alínea anterior),

$$\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i) = \frac{1}{n} \sum_{i=1}^n y_i - \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \bar{y} - \bar{y} = 0.$$

- (c) Na expressão que define SQT vamos introduzir um par de parcelas de valor zero, que nos

ajudarão nas contas subsequentes:

$$\begin{aligned}
 SQT &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\
 &= \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{=SQRE} + \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{=SQR} + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \quad (1)
 \end{aligned}$$

Para que a igualdade pedida se verifique, é preciso que a última parcela na expressão (1) seja nula. Ora, recordando a definição dos valores ajustados de y e a expressão da ordenada na origem da recta de regressão, b_0 , temos que $\hat{y}_i = b_0 + b_1 x_i = \bar{y} + b_1(x_i - \bar{x})$. Logo, a última parcela da equação (1) pode ser re-escrita como:

$$\begin{aligned}
 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= 2 \sum_{i=1}^n [(y_i - \bar{y}) - b_1(x_i - \bar{x})] b_1(x_i - \bar{x}) \\
 &= 2 b_1 \left[\underbrace{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}_{=(n-1) cov_{xy}} - b_1 \underbrace{\sum_{i=1}^n (x_i - \bar{x})^2}_{=(n-1) s_x^2} \right]
 \end{aligned}$$

Pela expressão que define o declive b_1 da recta de regressão, tem-se $b_1 s_x^2 = cov_{xy}$. Logo, a diferença acima indicada anula-se.

(d) Viu-se na alínea anterior que $\hat{y}_i = b_0 + b_1 x_i = \bar{y} + b_1(x_i - \bar{x})$. Logo,

$$SQR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n [b_1(x_i - \bar{x})]^2 = b_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 = b_1^2 (n-1) s_x^2 .$$

6. Pela definição de coeficiente de correlação entre x e y , tem-se:

$$r_{xy} = \frac{COV_{xy}}{s_x \cdot s_y} = \frac{COV_{xy}}{s_x^2} \cdot \frac{s_x}{s_y} = b_1 \cdot \frac{s_x}{s_y}$$

7. Os dados `anscombe` podem ser visualizados escrevendo o nome do objecto:

```

> anscombe
  x1 x2 x3 x4  y1  y2  y3  y4
1 10 10 10  8  8.04 9.14  7.46  6.58
2  8  8  8  8  6.95 8.14  6.77  5.76
3 13 13 13  8  7.58 8.74 12.74  7.71
4  9  9  9  8  8.81 8.77  7.11  8.84
5 11 11 11  8  8.33 9.26  7.81  8.47
6 14 14 14  8  9.96 8.10  8.84  7.04
7  6  6  6  8  7.24 6.13  6.08  5.25
8  4  4  4 19  4.26 3.10  5.39 12.50
9 12 12 12  8 10.84 9.13  8.15  5.56
10 7  7  7  8  4.82 7.26  6.42  7.91
11 5  5  5  8  5.68 4.74  5.73  6.89

```

Os nomes das variáveis indicam quatro variáveis x_i (as primeiras três são idênticas) e quatro variáveis y_i ($i = 1, 2, 3, 4$).

(a) As médias de cada variável são dadas por:

```
> apply(anscombe, 2, mean)
      x1      x2      x3      x4      y1      y2      y3      y4
9.000000 9.000000 9.000000 9.000000 7.500909 7.500909 7.500000 7.500909
```

Repare-se que as quatro variáveis x_i têm a mesma média e as quatro variáveis y_i também (aproximadamente).

(b) As variâncias de cada variável são dadas por:

```
> apply(anscombe, 2, var)
      x1      x2      x3      x4      y1      y2      y3      y4
11.000000 11.000000 11.000000 11.000000 4.127269 4.127629 4.122620 4.123249
```

De novo, as variáveis x_i partilham a mesma variância e as variáveis y_i também (aproximadamente).

(c) Tem-se

```
> lm(y1 ~ x1, data=anscombe)
Call: lm(formula = y1 ~ x1, data = anscombe)
Coefficients:
(Intercept)          x1
      3.0001         0.5001
```

```
> lm(y2 ~ x2, data=anscombe)
Call: lm(formula = y2 ~ x2, data = anscombe)
Coefficients:
(Intercept)          x2
      3.001         0.500
```

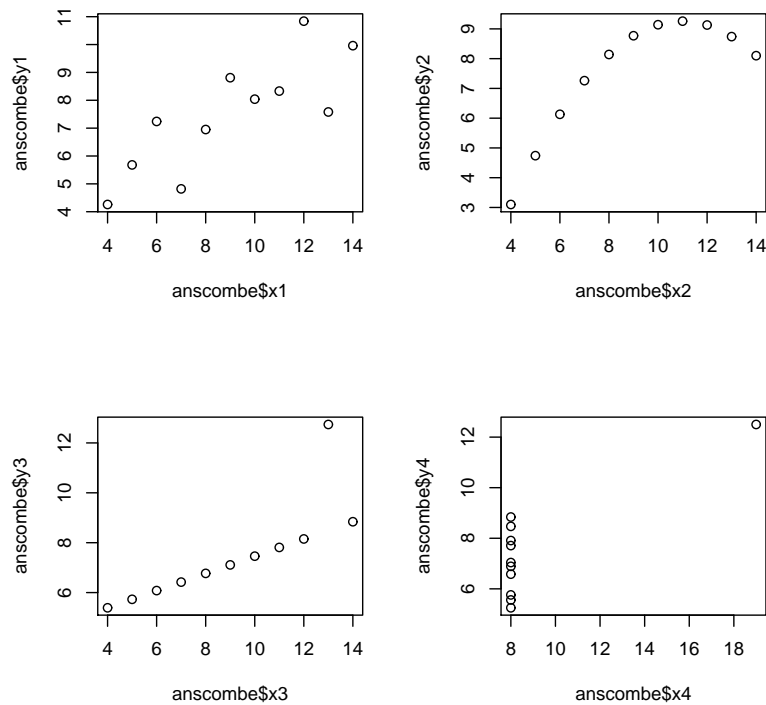
```
> lm(y3 ~ x3, data=anscombe)
Call: lm(formula = y3 ~ x3, data = anscombe)
Coefficients:
(Intercept)          x3
      3.0025         0.4997
```

```
> lm(y4 ~ x4, data=anscombe)
Call: lm(formula = y4 ~ x4, data = anscombe)
Coefficients:
(Intercept)          x4
      3.0017         0.4999
```

As quatro rectas de regressão pedidas são quase idênticas, de equação $y = 3 + 0.5x$.

(d) Os quatro coeficientes de correlação $r_{x_i y_i}$ ($i = 1, 2, 3, 4$) são quase iguais, de valor aproximado $r_{x_i y_i} = 0.816$, pelo que os quatro coeficientes de determinação das quatro rectas de regressão pedidas são quase iguais, de valores muito próximos de $R^2 = 0.667$.

Apesar de tudo indicar que os quatro pares de variáveis x_i e y_i são análogos, trata-se de conjuntos de dados muito diferentes como revelam as quatro nuvens de pontos:



Este exercício visa frisar que, por muito valor que tenham indicadores descritivos e de síntese das relações entre variáveis, é sempre aconselhável utilizar todas as ferramentas de análise dos dados disponíveis.

8. A *data frame* *iris* tem observações de quatro variáveis morfológicas (comprimento e largura de pétalas e sépalas) em $n = 150$ lírios de cada uma de três diferentes espécies. O tamanho da *data frame* pode ser vista através do comando `dim`, enquanto que as primeiras 10 linhas de dados podem ser vistas indexando a *data frame* da forma que já conhecemos:

```
> dim(iris)
[1] 150  5
> iris[1:10,]
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1           5.1          3.5          1.4          0.2  setosa
2           4.9          3.0          1.4          0.2  setosa
3           4.7          3.2          1.3          0.2  setosa
4           4.6          3.1          1.5          0.2  setosa
5           5.0          3.6          1.4          0.2  setosa
6           5.4          3.9          1.7          0.4  setosa
7           4.6          3.4          1.4          0.3  setosa
8           5.0          3.4          1.5          0.2  setosa
9           4.4          2.9          1.4          0.2  setosa
10          4.9          3.1          1.5          0.1  setosa
```

- (a) A nuvem de pontos pedida envolve as variáveis correspondentes às colunas 3 (x) e 4 (y). Logo, a nuvem de pontos pedida obtém-se através do comando:

```
> plot(iris[,c(3,4)])
```

- (b) Os comandos para responder ao que se pede no enunciado são:

```
> lm(Petal.Width ~ Petal.Length, data=iris)
> abline(lm(Petal.Width ~ Petal.Length, data=iris))
```

Os coeficientes da recta de regressão ajustada são $b_0 = -0.3631$ e $b_1 = 0.4158$.

- (c) Pede-se para trocar o papel das variáveis preditora e resposta. A recta de regressão “de x sobre y ” é dada pelo comando:

```
> lm(Petal.Length ~ Petal.Width, data=iris)
```

que indica que os valores dos parâmetros da recta são $b_0^* = 1.084$ e $b_1^* = 2.230$.

- (d) Para traçar a recta obtida *no sistema de eixos original* (isto é, com a variável `Petal.Width` no eixo vertical e a variável `Petal.Length` no eixo horizontal), é necessário ter em conta o facto indicado no enunciado: uma recta de equação $x = b_0^* + b_1^* y$, expressa na forma usual (isolando a variável y que vai para o eixo vertical) tem equação $y = -\frac{b_0^*}{b_1^*} + \frac{1}{b_1^*} x$. Logo, o comando necessário para dizer esta nova recta em cima dos eixos originais é:

```
> abline(-1.084/2.230, 1/2.230, col="red")
```

NOTA: O parâmetro `col` indica que a recta será traçada com a cor vermelha, o que ajuda a identificar cada uma das rectas em questão.

- (e) As rectas são diferentes porque resultam de otimizar critérios diferentes. Fixando o sistema de eixos de tal forma que o Comprimento das Pétalas esteja no eixo horizontal (x) e a Largura das Pétalas esteja no eixo vertical (y), a recta de regressão tradicional (de y sobre x) resulta de minimizar a soma dos quadrados das distâncias na vertical entre os pontos e a recta, enquanto que a “recta de regressão de x sobre y ” resulta de minimizar a soma dos quadrados das distâncias *na horizontal* entre pontos e recta.

9. Os dados referidos no enunciado são obtidos como se indica a seguir:

```
> library(MASS)
> Animals
      body  brain
Mountain beaver  1.350   8.1
Cow              465.000 423.0
Grey wolf        36.330 119.5
Goat             27.660 115.0
Guinea pig      1.040   5.5
Dipliodocus     11700.000  50.0
Asian elephant  2547.000 4603.0
Donkey          187.100 419.0
Horse           521.000 655.0
Potar monkey    10.000 115.0
Cat             3.300  25.6
Giraffe         529.000 680.0
Gorilla         207.000 406.0
Human           62.000 1320.0
African elephant 6654.000 5712.0
Triceratops     9400.000  70.0
```

Rhesus monkey	6.800	179.0
Kangaroo	35.000	56.0
Golden hamster	0.120	1.0
Mouse	0.023	0.4
Rabbit	2.500	12.1
Sheep	55.500	175.0
Jaguar	100.000	157.0
Chimpanzee	52.160	440.0
Rat	0.280	1.9
Brachiosaurus	87000.000	154.5
Mole	0.122	3.0
Pig	192.000	180.0

(a) `> plot(Animals)`.

(b) Pedem-se vários gráficos com transformações de uma ou ambas as variáveis:

i. `> plot(Animals$body, sqrt(Animals$brain))`.

ii. `> plot(Animals$body, log(Animals$brain))`.

iii. `> plot(log(Animals$body), Animals$brain)`.

iv. `> plot(log(Animals))`

(ou, alternativamente, `plot(log(Animals$body), log(Animals$brain))`).

O resultado do comando `log(Animals)` é o de calcular os logaritmos de *todos* os valores na *data frame*, ou seja, logaritmizar as duas variáveis simultaneamente.

NOTA: Os logaritmos aqui referidos são os logaritmos naturais, \ln . Por omissão, o comando `log` do R calcula logaritmos naturais.

(c) Como se viu nas aulas teóricas (Acetato 78), uma relação linear entre $\ln(y)$ e $\ln(x)$ corresponde a uma relação potência (alométrica) entre as variáveis originais: $y = cx^d$. Neste caso, tem-se uma relação de tipo alométrico entre pesos duma parte do organismo (cérebro) e do todo (corpo). O último gráfico indica que é aceitável admitir uma relação potência para estes dados.

(d) O comando

```
> identify(log(Animals))
```

permite, com o auxílio do rato, identificar pontos seleccionados pelo utilizador. (Para sair do modo interactivo, clicar no botão direito do rato).

NOTA: É necessário explicitar as coordenadas dos pontos no gráfico que se vai aceder com o comando. No nosso caso, isso significa explicitar as coordenadas dos dados logaritmizados: `log(Animals)`.

O enunciado pede para identificar os pontos que se destacam da relação linear, e que são os pontos 6, 16 e 26. Seleccionando as linhas com esses números podemos identificar as espécies em questão, e verificar que se trata de espécies de dinossáurios, as únicas espécies de animais extintos presentes no conjunto de dados:

```
> Animals[c(6,16,26),]
      body brain
Dipliodocus 11700 50.0
Triceratops  9400 70.0
Brachiosaurus 87000 154.5
```

(e) os comandos pedidos são:

```

> lm(log(brain) ~ log(body), data=Animals)
Call: lm(formula = log(brain) ~ log(body), data = Animals)
Coefficients:
(Intercept)    log(body)
      2.555         0.496
> abline(lm(log(brain) ~ log(body), data=Animals))

```

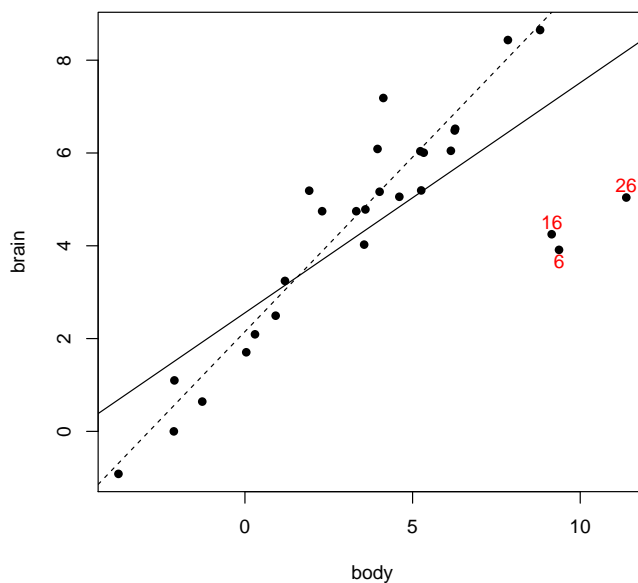
- (f) O parâmetro $b_1 = 0.496$ da recta ajustada tem duas leituras possíveis. Na relação entre as variáveis logaritmizadas tem a habitual leitura de qualquer declive duma recta de regressão: o log-peso do cérebro aumenta em média 0.496 log-gramas, por cada aumento de 1 log-kg no peso do corpo. Mais compreensível é a interpretação na relação potência entre as variáveis originais. Como se viu nas aulas teóricas, a relação original entre y e x é da forma $y = b_0 x^{b_1}$ com $b_1 = 0.496$ e $\ln(b_0) = 2.555 \Leftrightarrow b_0 = e^{2.555} = 12.871$. O valor de b_1 muito próximo de 0.5 permite simplificar a relação dizendo que o ajustamento indica que o peso do cérebro é aproximadamente proporcional à *raiz quadrada* do peso do corpo.
- (g) Utilizando a indexação negativa para eliminar as três espécies de dinossáurios pode proceder-se ao reajustamento da regressão, modificando o argumento `data` do comando `lm`. Pode juntar-se a nova recta ao gráfico obtido antes, através do comando `abline`. Este comando será invocado com um argumento pedindo que a recta seja desenhada a tracejado, a fim de melhor a distinguir da recta originalmente obtida:

```

> abline(lm(log(brain) ~ log(body), data=Animals[-c(6,16,26),]), lty="dashed")

```

O gráfico resultante tem o seguinte aspecto:



A exclusão das três espécies de dinossáurios (as observações atípicas) permitiu que a recta ajustada acompanhe melhor a relação linear existente entre a generalidade das espécies do conjunto de dados. Este exemplo ilustra que *as rectas de regressão são sensíveis à presença de observações atípicas*.

- (h) O ajustamento sem as espécies extintas produz os seguintes parâmetros da recta:

```
> lm(log(brain) ~ log(body), data=Animals[-c(6,16,26),])
Call: lm(formula = log(brain) ~ log(body), data = Animals[-c(6,16,26),])
Coefficients:
(Intercept)      log(body)
      2.1504         0.7523
```

O significado biológico destes valores é semelhante ao que foi visto na alínea 9f), com as diferenças resultantes dos novos valores . Assim, na relação alométrica entre peso do cérebro e peso do corpo (variáveis não transformadas), o expoente será aproximadamente 0.75, o que significa que o peso do cérebro é proporcional à potência $3/4$ do peso do corpo. Tendo em conta a relação na origem das relações potência (Acetato 78 das aulas teóricas), pode afirmar-se que *a taxa de variação relativa do peso do cérebro é aproximadamente três quartos da taxa de variação relativa do peso do corpo*, para o conjunto das espécies (não extintas) analisadas.