INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2016-17

Resoluções de exercícios de Regressão Linear Simples

1. Admite-se que foi criado o objecto `Cereais`, tal como indicado no enunciado. Para ver *o conteúdo* desse objecto `Cereais`, escrevemos o seu nome, como ilustrado de seguida (tendo sido omitidas várias linhas do conteúdo por razões de espaço):

```
> Cereais
    ano    area
1  1986 8789.69
2  1987 8972.11
3  1988 8388.94
4  1989 9075.35
5  1990 7573.48
(...)
24 2009 3398.99
25 2010 3041.18
26 2011 2830.96
```

NOTA: O comando `read.csv` parte do pressuposto que o ficheiro indicado contém colunas de dados - cada coluna correspondente a uma variável. O objecto `Cereais` criado no comando acima é uma *data frame*, que pode ser encarada como uma tabela de dados em que cada coluna corresponde a uma variável. As variáveis (colunas) individuais da *data frame* podem ser acedidas através duma indexação análoga à utilizada para objectos de tipo matriz, refereciando o número da respectiva coluna:

```
> Cereais[,2]
 [1] 8789.69 8972.11 8388.94 9075.35 7573.48 8276.47 7684.20 7217.93 6773.54
[10] 6756.57 6528.18 6902.34 5065.38 5923.45 5779.21 4927.15 5149.21 4507.98
[19] 4636.46 3893.43 3731.92 3120.99 3653.74 3398.99 3041.18 2830.96
```

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:

```
> Cereais$area
 [1] 8789.69 8972.11 8388.94 9075.35 7573.48 8276.47 7684.20 7217.93 6773.54
[10] 6756.57 6528.18 6902.34 5065.38 5923.45 5779.21 4927.15 5149.21 4507.98
[19] 4636.46 3893.43 3731.92 3120.99 3653.74 3398.99 3041.18 2830.96
```

(a) `> plot(Cereais)`

O gráfico obtido revela uma forte relação linear (decrescente) entre anos e superfície agrícola dedicada à produção de cereais.

Repare-se que o comando funciona correctamente nesta forma muito simples porque: (i) a *data frame* `Cereais` 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. Existe uma forma mais geral do comando que também poderia ser usada neste caso: `plot(x,y)`, onde `x` e `y` indicam os nomes das variáveis que desejamos ocupem, respectivamente o eixo horizontal e o eixo vertical. No nosso exemplo, poderíamos escrever:

```
> plot(Cereais$ano, Cereais$area)
```

(b) O gráfico obtido na alínea anterior apresenta uma tendência linear descrescente, pelo que o coeficiente de correlação será negativo. A tendência linear é bastante acentuada, pelo que é de supor que o coeficiente de correlação seja próximo de -1.

O comando `cor` do `R` calcula coeficientes de correlação. Se os seus argumentos forem dois vectores (necessariamente de igual dimensão), é devolvido o coeficiente de correlação. Se o seu argumento fôr uma *data frame*, é devolvida uma matriz de correlações entre todos os pares de variáveis da *data frame*. No nosso caso, esta segunda alternativa produz:

```
> cor(Cereais)
            ano       area
ano   1.0000000 -0.9826927
area -0.9826927  1.0000000
```

O coeficiente de correlação entre `ano` e `area` é, como previsto, muito próximo de -1, confirmando a existência duma forte relação linear decrescente entre anos e superfície agrícola para a produção de cereais em Portugal, nos anos indicados.

(c) Os parâmetros da recta podem ser calculados, quer a partir da sua definição, quer utilizando o comando do R que ajusta uma regressão linear: o comando `lm` (as iniciais, pela ordem em inglês, de *modelo linear*). Sabemos que:

$$b_1 = \frac{cov_{xy}}{s_x^2} \qquad \text{e} \qquad b_0 = \overline{y} - b_1\,\overline{x}\ .$$

Utilizando o R, é possível calcular os indicadores estatísticos nas definições:

```
> cov(Cereais$ano, Cereais$area)
[1] -15137.48
> var(Cereais$ano)
[1] 58.5
> -15137.48/58.5
[1] -258.7603
> mean(Cereais$area)
[1] 5869.187
> mean(Cereais$ano)
[1] 1998.5
> 5869.187-(-258.7603)*1998.5
[1] 523001.6
```

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

```
> lm(area ~ ano, data=Cereais)
Call:
lm(formula = area ~ ano, data = Cereais)
Coefficients:
(Intercept)          ano
   523001.7       -258.8
```

NOTA: Na fórmula $y \sim x$, 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 fórmula.

O resultado deste ajustamento pode ser guardado como um novo objecto, que poderá ser invocado sempre que se deseje trabalhar com a regressão agora ajustada:

```
> Cereais.lm <- lm(area ~ ano, data=Cereais)
```

Interpretação dos coeficientes:

- Declive: $b_1 = -258.8\,km^2/ano$ indica que, em cada ano que passa, a superficie agrícola dedicada à produção de cereais diminui, em média, $258,8\,km^2$. 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. Fala-se em "variação média" porque a recta apenas descreve a tendência de fundo, na relação entre x e y.
- Ordenada na origem: $b_0 = 523001.7\,km^2$. Em geral, as unidades de b_0 são as unidades da variável resposta y. A interpretação deste valor é, neste caso, estranha: a superfície agrícola utilizada na produção de cereais no ano $x = 0$, seria cerca de 5 vezes superior à área total do país, uma situação claramente impossível. A impossibilidade ilustra a ideia geral de que, *na ausência de mais informação, a validade duma relação linear não poder ser extrapolada para longe da gama de valores de x observada* (neste caso, os anos 1986-2011).

(d) 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(Cereais$ano, Cereais$area)
[1] -0.9826927
> cor(Cereais$ano, Cereais$area)^2
[1] 0.9656849
```

No nosso caso $R^2 = 0.9656849$, ou seja, cerca de $96,6\%$ da variabilidade total observada para a variável resposta y é explicada pela regressão.

O comando `summary`, aplicando ao resultado da regressão ajustada, produz vários resultados de interesse relativos à regressão. O coeficiente de determinação pedido nesta alínea é indicado na penúltima linha da listagem produzida:

```
> summary(Cereais.lm)
(...)
Multiple R-squared: 0.9657
(...)
```

(e) O comando `abline(Cereais.lm)` traça a recta pedida em cima do gráfico anteriormente criado pelo comando `plot`. Confirma-se o bom ajustamento da recta à nuvem de pontos, já indiciado pelo valor muito elevado do R^2.

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 duma regressão linear simples (obtido através do comando `lm` e que devolve o par de coeficientes b_0 e b_1), o resultado é o gráfico da recta $y = b_0 + b_1\,x$.

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

```
> (length(Cereais$area)-1)*var(Cereais$area)
[1] 101404176
```

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

```
>  0.9656849*101404176
[1] 97924482
```

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(Cereais.lm)
       1        2        3        4        5        6        7        8
9103.691 8844.930 8586.170 8327.410 8068.649 7809.889 7551.129 7292.368
       9       10       11       12       13       14       15       16
7033.608 6774.848 6516.087 6257.327 5998.567 5739.806 5481.046 5222.286
(...)
> (length(Cereais$area)-1)*var(fitted(Cereais.lm))
[1] 97924480
```

NOTA: A pequena discrepância nos dois valores obtidos para SQR deve-se a erros de arredondamento.

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

```
> residuals(Cereais.lm)
         1          2          3          4          5          6          7
-314.00068  127.17965 -197.23002  747.94031 -495.16936  466.58097  133.07131
         8          9         10         11         12         13         14
 -74.43836 -260.06803  -18.27770   12.09263  645.01296 -933.18670  183.64363
(...)
> sum(residuals(Cereais.lm)^2)
[1] 3479697
```

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

```
>  97924480+3479697
[1] 101404177
```

(i) Com o auxílio do `R`, podemos efectuar o novo ajustamento. No caso de se efectuar uma transformação duma variável, esta deve ser efectuada, na fórmula do comando `lm`, com a protecção `I()`, como indicado no comando seguinte:

```
> lm(I(area*100) ~ ano, data=Cereais)
Call:
lm(formula = I(area * 100) ~ ano, data = Cereais)
Coefficients:
(Intercept)          ano
   52300171       -25876
```

Comparando estes valores dos parâmetros ajustados com os que haviam sido obtidos incialmente, pode verificar-se que ambos os parâmetros ajustados aparecem multiplicados por 100. Não se trata duma coincidência, o que se pode verificar inspeccionando o efeito da transformação $y \to y^* = c\, y$ (para qualquer constante c) nas fórmulas dos parâmetros da recta ajustada. Indicando por b_1 e b_0 os parâmetros na recta original e por b_1^* e b_0^* os novos parâmetros, obtidos com a transformação indicada, temos (recordando que $cov(x, cy) = c\, cov(x, y)$):

$$b_1^* = \frac{cov_{x\,y^*}}{s_x^2} = \frac{cov(x, c\,y)}{s_x^2} = c\ \frac{cov(x,y)}{s_x^2} = c\, b_1\ ;$$

e (tendo em conta o efeito de constantes multiplicativas sobre a média, ou seja, $\overline{y^*} = c\,\overline{y}$):

$$b_0^* = \overline{y^*} - b_1^*\,\overline{x} = c\overline{y} - c\, b_1\,\overline{x} = c\,(\overline{y} - b_1\overline{x}) = c\, b_0\ .$$

Assim, multiplicar a variável resposta por uma constante c tem por efeito multiplicar os dois parâmetros da recta ajustada por essa mesma constante c. No entanto, o coeficiente de determinação permanece inalterado. Esse facto, que resulta da invariância do valor absoluto do coeficiente de correlação a qualquer transformação linear de uma, ou ambas as variáveis, pode ser confirmado através do `R`:

```
> summary(lm(I(area*100) ~ ano, data=Cereais))
(...)
Multiple R-squared: 0.9657
(...)
```

(j) Nesta alínea é pedida uma translação da variável preditora, da forma $x \to x^* = x + a$, com $a = -1985$. Neste caso, e comparando com o ajustamento inicial, verifica-se que o declive da recta de regressão não se altera, mas a sua ordenada na origem sim:

```
> lm(area ~ I(ano-1985), data=Cereais)
Call:
lm(formula = area ~ I(ano - 1985), data = Cereais)
Coefficients:
  (Intercept)  I(ano - 1985)
       9362.5         -258.8
```

Inspeccionando o efeito duma translação na variável preditora sobre o declive da recta ajustada, temos (tendo em conta que constantes aditivas não alteram, nem a variância, nem a covariância):

$$b_1^* = \frac{cov_{y\,x^*}}{s^2_{x^*}} = \frac{cov(x,y)}{s^2_x} = b_1 \ .$$

Já no que respeita à ordenada na origem, e tendo em conta a forma como os valores médios são afectados por constantes aditivas, tem-se:

$$b_0^* = \overline{y} - b_1^* \, \overline{x^*} = \overline{y} - b_1 \, (\overline{x} + a) = (\overline{y} - b_1 \, \overline{x}) - b_1 \, a = b_0 - a \, b_1 \ .$$

Assim, no nosso caso (e usando os valores com mais casas decimais obtidos acima, para evitar ulteriores erros de arredondamento), tem-se que a nova ordenada na origem é $b_0^* = 523001.6 - (-1985) * (-258.7603) = 9362.405$.

Tal como na alínea anterior, a transformação da variável preditora é linear, pelo que o coeficiente de determinação não se altera: $R^2 = 0.9657$.

2. (a) Seguindo as instruções do enunciado, cria-se o ficheiro de texto `Azeite.txt` na directoria da sessão de trabalho do R, que se recomenda ser uma pasta de nome `AulasED`, num dispositivo de armazenamento de fácil acesso (por exemplo, uma *pen*). Para se saber qual a directoria de trabalho duma sessão do `R`, pode ser dado o seguinte comando:

```
> getwd()
```

(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, qualquer dos seguintes comandos (alternativos) produz 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 a matriz dos coeficientes de correlação entre cada par de variáveis da *data frame*.

```
> 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 positivo 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. Em geral, 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) $\sum_{i=1}^{n} (x_i - \overline{x}) \;=\; \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \overline{x} \;=\; n\,\overline{x} - n\,\overline{x} \;=\; 0.$

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

$$(n-1)cov_{xy} \;=\; \sum_{i=1}^{n}(x_i-\overline{x})y_i - \sum_{i=1}^{n}(x_i-\overline{x})\overline{y} \;=\; \sum_{i=1}^{n}(x_i-\overline{x})y_i - \overline{y}\underbrace{\sum_{i=1}^{n}(x_i-\overline{x})}_{=\,0} \;=\; \sum_{i=1}^{n}(x_i-\overline{x})y_i$$

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

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.

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:

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

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

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

(c) Por definição, a a variância amostral de n observações z_i quaisquer, é dada por $s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \overline{z})^2$. Assim, a variância amostral das n observações y_i da variável resposta

é dada por: $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$. O somatório nesta expressão é, por definição, a Soma de Quadrados Total, SQT. Logo, $SQT = (n-1) \times s_y^2$, como se pedia para justificar. De forma análoga, a variância dos n valores ajustados da variável resposta, os $\hat{y}_i$,é dada por: $s_{\hat{y}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$. Repare-se que, nesta última expressão, a média dos valores $\hat{y}_i$ é igual à média dos n valores observados y_i, como se viu na alínea 5a. Mas o somatório nesta expressão é, por definição, SQR. Logo, $SQR = (n-1) \times s_{\hat{y}}^2$. Finalmente, a variância amostral dos n resíduos e_i é, por definição, $s_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}(e_i - \overline{e})^2$. Mas a média dos n resíduos é nula, como se viu na alínea 5b, e $\overline{e}=0$ implica que o somatório nesta última expressão é apenas a Soma de Quadrados Residual, $SQRE = \sum_{i=1}^{n} e_i^2$. Logo, $SQRE = (n-1) \times s_e^2$.

(d) Ora, recordando a definição dos valores ajustados de y e a expressão da ordenada na origem da recta de regressão, b_0, temos que $\hat{y}_i = b_0 + b_1\, x_i = \overline{y} + b_1(x_i - \overline{x})$. Logo,

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

(e) Na expressão que define SQT vamos introduzir um par de parcelas de soma zero, que nos ajudarão nas contas subsequentes ($-\hat{y}_i + \hat{y}_i = 0$):

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

Para que a igualdade pedida se verifique, é preciso que a última parcela na expressão (1) seja nula. Viu-se acima que $\hat{y}_i = b_0 + b_1\, x_i = \overline{y} + b_1(x_i - \overline{x})$. Logo, o somatório na última parcela da equação (1) pode ser re-escrito como:

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

Tendo em conta que $b_1 = \frac{cov_{xy}}{s_x^2}$, tem-se $b_1\, s_x^2 = cov_{xy}$. Logo, a diferença acima anula-se.

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 podem ser obtidas usando o comando `apply`:

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

```
> 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
```

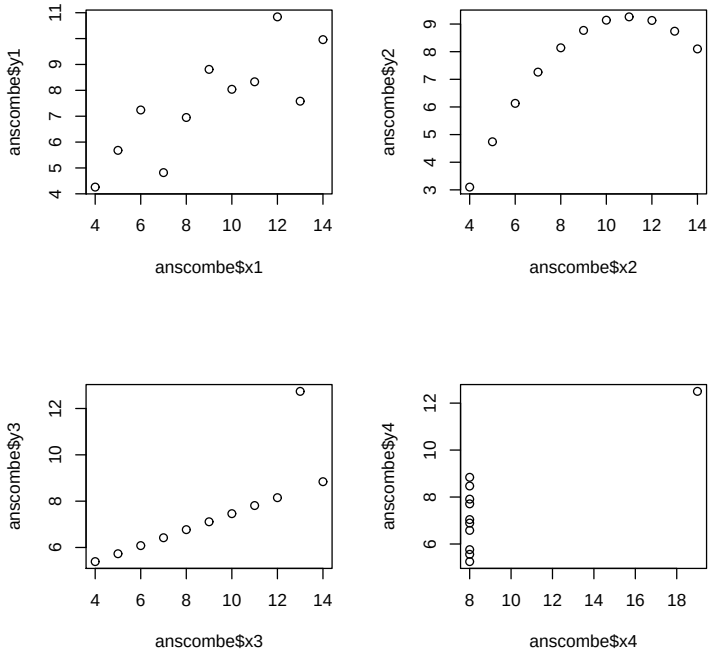
(c) As quatro rectas pedidas têm equação quase idêntica, aproximadamente $y = 3 + 0.5\,x$:

```
> 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
```

(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 seguintes. 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 morfométricas (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 8 linhas de dados podem ser vistas indexando a *data frame* da forma que já conhecemos:

```
> dim(iris)
[1] 150   5
> iris[1:8,]
   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
```

(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 traçar 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 côr vermelha, o que ajuda a identificar cada uma das rectas em questão.

(e) As rectas são diferentes porque resultam de optimizar 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) A nuvem de pontos pedida pode ser obtida através do comando `plot(Animals)`. Quanto ao coeficiente de correlação, tem-se:

```
> cor(Animals)
              body        brain
body   1.000000000 -0.005341163
brain -0.005341163  1.000000000
```

O valor quase nulo do coeficiente de correlação indica ausência de relacionamento linear entre os pesos do corpo e do cérebro, facto que se confirma visualmente no gráfico.

(b) Pedem-se vários gráficos com transformações de uma ou ambas as variáveis. Aproveita-se este exercício para introduzir uma forma alternativa de pedir uma nuvem de pontos, que utiliza uma sintaxe parecida com as usadas para escrever as fórmulas no comando `lm`:

i. O gráfico de log-pesos do cérebro (no eixo vertical) vs. pesos do corpo (eixo horizontal) pode ser obtido através da tradicional forma `plot(x,y)`, que no nosso caso seria

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

Alternativamente, pode dar-se o seguinte comando equivalente:

`> plot(log(brain) ~ body, data=Animals)`

ii. Usando a forma do comando agora introduzida, a nuvem de pontos pedida é dada por:

`> plot(brain ~ log(body), data=Animals)`

iii. Neste caso, e uma vez que a transformação logarítimica se aplica às duas variáveis da *data frame* `Animals`, basta dar o comando

`> plot(log(Animals))`

ou, alternativamente,

`> plot(log(brain) ~ log(body), data=Animals)`

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 (Acetatos 95-97), 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 da alínea anterior indica que é aceitável admitir uma relação potência entre o peso do cérebro e o peso do corpo, nas espécies animais consideradas.

(d) Os coeficientes de correlação e de determinação entre log-pesos do corpo e log-pesos do cérebro podem ser calculados, com o auxílio do R, da seguinte forma:

```
> cor(log(Animals$body), log(Animals$brain))    <-- coeficiente de correlação
[1] 0.7794935
> cor(log(Animals$body), log(Animals$brain))^2  <-- coeficiente de determinação
[1] 0.6076101
```

Dado o valor $R^2 = 0.6076$, a regressão linear entre log-peso do cérebro e log-peso do corpo explica menos de 61% da variabilidade total dos log-pesos do cérebro observados. Este valor, aparentemente contraditório com a relativamente forte relação linear para a maioria das espécies, é reflexo da presença nos dados das três espécies (pontos) que são claramente atípicas face às restantes.

(e) Os comandos pedidos são:

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

(admitindo que o último comando `plot` dado antes deste comando `abline` fosse o do gráfico correspondente à dupla logaritmização).

(f) O declive $b_1^* = 0.496$ 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 $y = c\,x^d$ com $d = b_1^* = 0.496$ e $b_0^* = \ln(c) = 2.555 \Leftrightarrow c = \mathrm{e}^{2.555} = 12.871$. No nosso caso, a tendência de fundo na relação entre peso do corpo (x) e peso do cérebro (y) é $y = 12.871\,x^{0.496}$. O valor de d muito próximo de 0.5 permite simplificar a relação dizendo que o ajustamento indica que o peso do cérebro é aproximadamente proporcional à *raíz quadrada* do peso do corpo.

(g) 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. Selecionando 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:

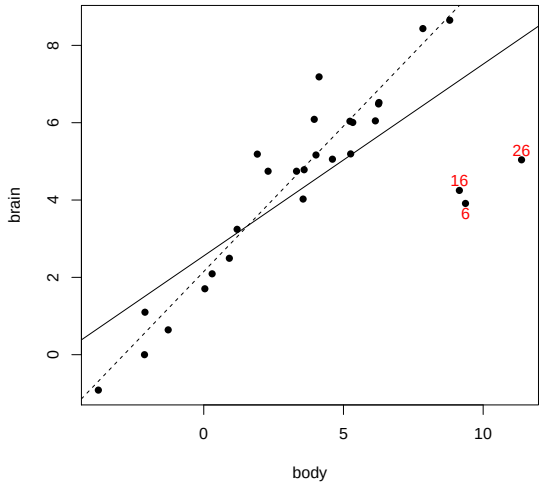
```
> Animals[c(6,16,26),]
              body brain
Dipliodocus   11700  50.0
```

```
Triceratops    9400  70.0
Brachiosaurus 87000 154.5
```

(h) 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 é reproduzido abaixo. 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*. Neste caso, as espécies de dinossáurios "atraem" a recta de regressão, afastando-a da generalidade das restantes espécies.



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

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

Note-se como os parâmetros da recta se alteram: o declive da recta cresce para mais de 0.75 e a ordenada na origem decresce um pouco. Além disso, podemos analisar o efeito sobre o coeficiente de determinação, através da aplicação do comando `summary` à regressão agora ajustada:

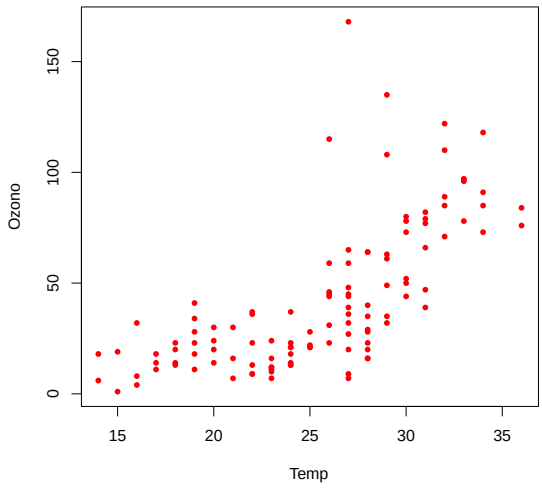
```
> summary(Animals.loglm.sub)
(...)
Multiple R-squared: 0.9217
(...)
```

Com a exclusão das espécies extintas, a recta de regressão passa a explicar mais de 92% da variabilidade total nos restantes log-pesos do cérebro, a partir dos log-pesos do corpo.

(j) O significado biológico dos parâmetros da recta 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 97 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.

10. (a) O comando `plot(ozono)` produz o gráfico pedido. Um gráfico com alguns embelezamentos adicionais é produzido pelo comando:

```
> plot(ozono, col="red", pch=16, cex=0.8)
```



(b) A linearização duma relação exponencial faz-se logaritmizando:

$$y = a\mathrm{e}^{bx} \quad \Leftrightarrow \quad \ln(y) = \ln(a) + b\,x \ ,$$

que é uma relação linear entre x e $y^* = \ln(y)$.

i. O gráfico de log-`Ozono` contra `Temp` pode ser construído pelo comando:

```
> plot(ozono$Temp, log(ozono$Ozono))
```

Uma tendência linear mais ou menos forte neste gráfico indica que a relação exponencial entre as variáveis originais é adequada. Neste caso, o gráfico corresponde a um coeficiente de correlação entre `Temp` e log-`Ozono` de 0.73.

ii. O ajustamento pedido faz-se da seguinte forma:

```
> lm(log(Ozono) ~ Temp, data=ozono)
Call: lm(formula = log(Ozono) ~ Temp, data = ozono)
Coefficients:
(Intercept)         Temp
     0.3558       0.1203
```

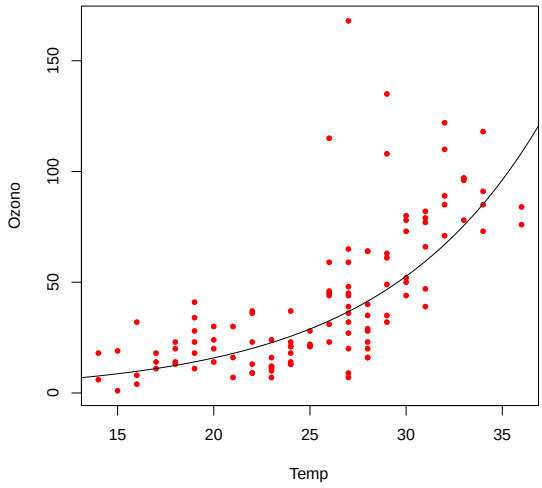
O coeficiente de determinação é de cerca de $R^2 = 0.73^2 = 0.53$ (aplicando o comando `summary` ao modelo agora ajustado verifica-se ser $R^2 = 0.5372$), o que significa que a regressão explica pouco mais de 53% da variabilidade dos log-teores de ozono.

iii. O declive estimado da recta $b_1 = 0.1203$ é o coeficiente do expoente, na relação exponencial original, uma vez que estima o parâmetro b que tem esse significado. Já a ordenada na origem da recta ajustada, $b_0 = 0.3558$ corresponde à estimativa de $\ln(a)$, pelo que a constante multiplicativa a da relação exponencial original é: $a = \mathrm{e}^{0.3558} = 1.4273$.

iv. Para prever o *teor de ozono* y utiliza-se a equação exponencial ajustada nas alíneas anteriores, ou seja, a equação $y = 1.4273\,\mathrm{e}^{0.123\,x}$, onde x indica a temperatura máxima. Assim, o valor ajustado do teor médio de ozono, num dia de temperatura máxima $x = 25$ é dado por $\hat{y} = 1.4273\,\mathrm{e}^{0.123\times 25} = 28.8839$. É igualmente possível chegar a este valor utilizando directamente a recta de regressão, desde que se tenha em atenção que os valores ajustados por essa recta são de log-teor de ozono, e que se torna necessário desfazer a transformação logarítmica. Assim, o valor *de log-ozono* previsto pela recta, para um dia com temperatura máxima de 25º é dado por: $\widehat{y^*} = \widehat{\ln(y)} = 0.3558 + 0.1203 \times 25 = 3.3633$. E o teor estimado *de ozono* (em ppm) é: $\mathrm{e}^{3.3633} = 28.8843$. A pequena diferenças nos valores obtidos por cada uma das vias acima resulta de erros de arredondamento.

(c) Para sobrepôr a curva exponencial à nuvem de pontos de ozono vs. temperaturas, podem usar-se os seguintes comandos:

```
> plot(ozono, col="red", pch=16, cex=0.8)
> curve(1.4273*exp(0.1203*x), from=10, to=40, add=TRUE)
```



11. (a) Com as restrições indicadas no enunciado, y não se anula e pode tomar-se o recíproco de y:

$$\frac{1}{y} = \frac{b+x}{ax} = \frac{b}{a}\cdot\frac{1}{x} + \frac{1}{a} \quad \Leftrightarrow \quad y^* = b_0^* + b_1^* x^* \,,$$

com $y^* = \frac{1}{y}$, $x^* = \frac{1}{x}$, $b_0^* = \frac{1}{a}$ e $b_1^* = \frac{b}{a}$. Assim, uma *relação linear entre os recíprocos de y e de x corresponde a uma relação de Michaelis-Menten entre y e x.*

(b) Tendo em conta os nomes indicados no enunciado, e o facto de os dados do enunciado corresponderem *apenas às 12 primeiras linhas da* ***data frame*** (associadas ao valor `treated` da terceira coluna, de nome `state`) , o modelo linearizado ajusta-se através do comando:

```
> lm(I(1/rate) ~ I(1/conc), data=Puromycin[Puromycin$state=="treated",])
```

sendo os resultados obtidos os seguintes:

```
Coefficients:
(Intercept)    I(1/conc)
0.0051072    0.0002472
```

(c) Tendo em conta as relações vistas na alínea anterior, $b_0^* = \frac{1}{a} = 0.0051072$, tem-se $a = 195.802$. Por outro lado, $b_1^* = \frac{b}{a} = 0.0002472$, logo $b = 0.0002472 \times 195.802 = 0.04840225$. Assim, o modelo de Michaelis-Menten ajustado é: $y = \frac{195.802\,x}{0.04840225+x}$. Repare-se que o limite de y quando x tende para $+\infty$ é 195.802, que é assim a estimativa da assintota superior da relação de Michaelis-Menten. O gráfico da relação original sugere que se pode tratar duma subestimação do verdadeiro valor desta assintota horizontal. Este exemplo ilustra que pode haver inconvenientes associados à utilização de transformações linearizantes, como indicado nos acetatos das aulas teóricas.

Exercícios de inferência estatística na Regressão Linear Simples

12. Comecemos por recordar a definição e propriedades da covariância de variáveis aleatórias, que serão utilizadas na resolução deste exercício:

- $cov[X,Y] \;=\; E\left[(X-E[X])(Y-E[Y])\right] \;=\; E[XY]-E[X]\,E[Y]$.
- $cov[X,X] \;=\; E\left[(X-E[X])^2\right] \;=\; V[X]$.
- $cov[X,Y] \;=\; cov[Y,X]$.
- $cov[a+bX,Y] \;=\; b\,cov[X,Y]$.
- $cov[X+Y,Z] \;=\; cov[X,Z]+cov[Y,Z]$.
- Aplicando repetidamente as propriedades anteriores, vê-se que a covariância de combinações lineares de variáveis aleatórias se pode escrever como uma combinação linear das covariâncias:

$$cov\left[\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right] \;=\; \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j\, cov[X_i,Y_j]\ .$$

(a) Pretende-se calcular a covariância de $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ e $\hat{\beta}_1 = \sum_{j=1}^{n} c_j Y_j$, com $c_j = \frac{(x_j-\overline{x})}{(n-1)s_x^2}$. Ora, pelas propriedades da covariância acima referidas, tem-se:

$$cov[\overline{Y},\hat{\beta}_1] \;=\; cov\left[\frac{1}{n}\sum_{i=1}^{n} Y_i, \sum_{j=1}^{n} c_j Y_j\right] \;=\; \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} c_j\, cov[Y_i,Y_j]\ .$$

Sabemos que as observações Y_i constituem um conjunto de v.a. independentes. Logo, $cov[Y_i,Y_j] = 0$, caso $i \neq j$. Assim, o duplo somatório reduz-se a um único somatório (correspondente a tomar $i = j$). Tendo ainda em conta que $cov[Y_i,Y_i] = V[Y_i] = \sigma^2$, tem-se (ver o Exercício 3a):

$$cov[\overline{Y},\hat{\beta}_1] = \frac{1}{n}\sum_{i=1}^{n} \sigma^2\, c_i = \frac{\sigma^2}{n}\sum_{i=1}^{n} c_i = 0\ ,$$

(b) Tendo em conta que $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\,\overline{x}$, e as propriedades das variâncias e covariâncias, tem-se:

$$cov[\hat{\beta}_0, \hat{\beta}_1] = cov\left[\overline{Y} - \hat{\beta}_1\,\overline{x}, \hat{\beta}_1\right] = \underbrace{cov[\overline{Y}, \hat{\beta}_1]}_{=0\ (alinea\ a)} - cov[\hat{\beta}_1\,\overline{x}, \hat{\beta}_1] = -\overline{x}\,V[\hat{\beta}_1] = \frac{-\overline{x}\ \sigma^2}{(n-1)s_x^2}\ .$$

(c) Sabemos que a independência de duas quantitades aleatórias implica que elas tenham correlação nula. Olhando para a expressão obtida na alínea anterior, é evidente que a correlação entre $\hat{\beta}_0$ e $\hat{\beta}_1$ apenas se anula se $\sigma = 0$ (o que corresponderia a admitir que não há variabilidade estatística na relação entre x e y, contexto que não corresponde a esta disciplina) ou se $\overline{x} = 0$. Apenas nesta última situação poderá existir independência entre $\hat{\beta}_0$ e $\hat{\beta}_1$.

13. Pretendemos determinar a distribuição de probabilidades do estimador $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\,\overline{x} = \sum_{i=1}^{n} d_i Y_i$, com $d_i = \frac{1}{n} - \overline{x}\,c_i$, como se viu nas aulas teóricas. Trata-se duma combinação linear de v.a.s Normais independentes (as observações Y_i), logo de distribuição Normal. Falta determinar os respectivos parâmetros. Recordando os resultados relativos ao estimador $\hat{\beta}_1$, já obtidos nas aulas teóricas, tem-se:

$$E\left[\hat{\beta}_0\right] = E\left[\overline{Y} - \hat{\beta}_1\,\overline{x}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] - \overline{x}\ \underbrace{E\left[\hat{\beta}_1\right]}_{=\ \beta_1} = \frac{1}{n}\sum_{i=1}^{n}\underbrace{(\beta_0 + \beta_1\,x_i)}_{=E[Y_i]} - \beta_1\,\overline{x} = \beta_0 + \beta_1\overline{x} - \beta_1\overline{x} = \beta_0.$$

Tendo em conta as propriedades da variância,

$$\begin{aligned} V[\hat{\beta}_0] &= V\left[\overline{Y} - \hat{\beta}_1\,\overline{x}\right] = V[\overline{Y}] + V[\hat{\beta}_1\,\overline{x}] - 2cov[\overline{Y}, \hat{\beta}_1\,\overline{x}] \\ &= V\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] + \overline{x}^2 V[\hat{\beta}_1] - 2\overline{x}\ \underbrace{cov[\overline{Y}, \hat{\beta}_1]}_{=0\ (Ex.12)} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\underbrace{V[Y_i]}_{=\sigma^2} + \overline{x}^2\,\frac{\sigma^2}{(n-1)s_x^2} = \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)\,s_x^2}\right], \end{aligned}$$

o que completa a demonstração.

14. A informação essencial sobre a regressão pedida pode ser obtida através do comando `summary`:

```
> iris.lm <- lm(Petal.Width ~ Petal.Length, data=iris)
> summary(iris.lm)
Call: lm(formula = Petal.Width ~ Petal.Length, data = iris)
(...)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.363076   0.039762  -9.131  4.7e-16 ***
Petal.Length  0.415755   0.009582  43.387  < 2e-16 ***
(...)
Residual standard error: 0.2065 on 148 degrees of freedom
Multiple R-squared: 0.9271,Adjusted R-squared: 0.9266
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

(a) As estimativas dos desvios padrão associados à estimação de cada um dos parâmetros são indicadas na tabela, na coluna de nome `Std.Error` (ou seja, erro padrão). Assim, o desvio padrão associado à estimação da ordenada na origem é $\hat{\sigma}_{\hat{\beta}_0} = 0.039762$. A variância correspondente é o quadrado deste valor, $\hat{\sigma}^2_{\hat{\beta}_0} = 0.001581$. Seria igualmente possível calcular esta variância estimada a partir da sua fórmula (Acetato 121): $\hat{\sigma}^2_{\hat{\beta}_0} = QMRE \left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)\, s_x^2}\right]$. O valor de $QMRE$ pode ser obtido a partir da listagem acima, uma vez que, sob a designação `Residual standard error`, a listagem indica o valor $\sqrt{QMRE} = 0.2065$. Os outros valores constantes da expressão podem ser calculados como em exercícios anteriores.

De forma análoga, o desvio padrão associado à estimação do declive da recta é $\hat{\sigma}_{\hat{\beta}_1} = 0.009582$, e o seu quadrado é a variância estimada de $\hat{\beta}_1$: $\hat{\sigma}^2_{\hat{\beta}_1} = 9.181472 \times 10^{-5}$. Também aqui, este valor pode ser obtido a partir da expressão $\hat{\sigma}^2_{\hat{\beta}_1} = \frac{QMRE}{(n-1)\, s_x^2}$.

(b) Um intervalo a $(1-\alpha)\times 100\%$ de confiança para β_1 é: $\left]\, b_1 - t_{\frac{\alpha}{2}(n-2)}\hat{\sigma}_{\hat{\beta}_1} \;,\; b_1 + t_{\frac{\alpha}{2}(n-2)}\hat{\sigma}_{\hat{\beta}_1} \,\right[$, sendo neste caso $\alpha = 0.05$, $n = 150$, $b_1 = 0.415755$, $\hat{\sigma}_{\hat{\beta}_1} = 0.009582$ e $t_{0.025(148)} = 1.976122$. Logo, o IC a 95% de confiança para o declive da recta é] 0.39682 , 0.43469 [. Esta é a gama de valores admissíveis (a 95% de confiança) para o declive da recta relacionando largura e comprimento das pétalas dos lírios (das três espécies analisadas). Os intervalos de confiança dos dois parâmetros da recta podem ser obtidos no `R` através do comando:

```
> confint(iris.lm)
                2.5 %     97.5 %
(Intercept)  -0.4416501 -0.2845010
Petal.Length  0.3968193  0.4346915
```

(c) Analogamente, um IC a $(1-\alpha)\times 100\%$ de confiança para β_0 é:

$$\left]\, b_0 - t_{\frac{\alpha}{2}(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_0} \;,\; b_0 + t_{\frac{\alpha}{2}(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_0} \,\right[$$

Neste exemplo, $b_0 = -0.363076$ e $\hat{\sigma}_{\hat{\beta}_0} = 0.039762$. O valor tabelado da distribuição t, para um intervalo a 95% de confiança, é o mesmo que na alínea anterior: $t_{0.025(148)} = 1.976122$. Logo, o intervalo de confiança pedido é $]\, -0.4416501\,,\, -0.2845010\,[$. Repare-se na maior amplitude deste intervalo, em relação ao IC para o declive populacional β_1, o que é consequência directa da maior variabilidade associada à estimação de β_0 (o valor de $\hat{\sigma}_{\hat{\beta}_0}$ é cerca de 4 vezes o valor de $\hat{\sigma}_{\hat{\beta}_1}$). A partir das fórmulas para estes dois erros padrão, é possível verificar que este maior valor de $\hat{\sigma}_{\hat{\beta}_0}$ resulta, não tanto da parcela adicional $\frac{1}{n}$ (como $n = 150$, esta parcela é pequena) mas sobretudo do $\overline{x}^2$ que surge no numerador da segunda parcela. De facto, a média das observações do comprimento de pétalas é aproximadamente $\overline{x} = 3.758$.

(d) A frase do enunciado traduz-se por "$\beta_1 = 0.5$". Assim, faremos um teste de hipóteses desta hipótese nula, contra a hipótese alternativa $H_1 : \beta_1 \neq 0.5$. Os cinco passos do teste são:

Hipóteses: $H_0 : \beta_1 = 0.5$ vs. $H_1 : \beta_1 \neq 0.5$.

Estatística do teste: $T = \frac{\hat{\beta}_1 - \beta_{1|H_0}}{\hat{\sigma}_{\hat{\beta}_1}} \;\cap\; t_{n-2}$

Nível de significância: $\alpha = 0.05$.

Região Crítica (Bilateral): Rejeitar H_0 se $|T_{calc}| > t_{\frac{\alpha}{2}(n-2)} = t_{0.025(148)} = 1.976122$.

Conclusões: O valor calculado da estatística do teste é: $T_{calc} = \frac{0.415755-0.5}{0.009582} = -8.792006$. Logo, rejeita-se claramente a hipótese nula que por cada centímetro a mais no comprimento da pétala, é de esperar meio centímetro a mais na largura da pétala.

(e) A hipótese referida no enunciado é que $\beta_1 < 0.5$. Neste caso, a opção entre colocar esta hipótese em H_0 ou em H_1 corresponde à opção entre dar, ou não, o benefício da dúvida a esta hipótese. Seja como fôr, o valor de fronteira (0.5) terá de pertencer à hipótese nula. Vamos optar por *não* dar o benefício da dúvida à hipótese indicada no enunciado:

Hipóteses: $H_0 : \beta_1 \geq 0.5$ vs. $H_1 : \beta_1 < 0.5$.

Estatística do teste: $T = \frac{\hat{\beta}_1 - 0.5}{\hat{\sigma}_{\hat{\beta}_1}} \cap t_{n-2}$

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral esquerda): Rej. H_0 se $T_{calc} < -t_{\alpha\,(n-2)} = -t_{0.05(148)} = -1.655215$.

Conclusões: O valor calculado da estatística do teste é igual ao da alínea anterior: $T_{calc} = \frac{0.415755-0.5}{0.009582} = -8.792006$. Logo, rejeita-se a hipótese nula, optando-se por H_1. Pode afirmar-se que é estatisticamente significativa a conclusão que, por cada centímetro a mais no comprimento da pétala, em média a respectiva largura cresce menos do que $0.5cm$.

(f) A afirmação do enunciado corresponde à hipótese $\beta_1 = 0$. De facto, se $\beta_1 = 0$, a equação do modelo que relaciona x e Y reduz-se a $Y_i = \beta_0 + \epsilon_i$, não existindo relação linear entre x e Y. O teste às hipóteses $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ pode ser feito como na alínea 14d) acima. No entanto, para o caso particular do valor do parâmetro $\beta_1 = 0$ a informação relativa ao teste já é indicada na listagem produzida pelo comando `summary`, nas terceira e quarta colunas da tabela `Coefficients`. Neste caso, o valor calculado da estatística é $T_{calc} = \frac{0.4157550}{0.009582} = 43.387$. Tendo em conta que a região crítica é igual à da alínea 14d), tem-se uma rejeição clara da hipótese nula $\beta_1 = 0$: o valor estimado $b_1 = 0.415755$ é *significativamente diferente* de zero (ao nível $\alpha = 0.05$), pelo que a recta tem alguma utilidade para prever valores de y (largura da pétala) a partir dos valores de x (comprimento da pétala). Esta conclusão também se pode justificar a partir do valor de prova ($p-value$) do valor calculado da estatística, que é muito pequeno, sendo mesmo inferior à precisão de máquina, $p < 2 \times 10^{-16}$. Mesmo para níveis de significância como $\alpha = 0.01$ ou $\alpha = 0.005$, a conclusão seria a de rejeição de H_0.

(g) Uma abordagem alternativa para a questão estudada na alínea anterior será a de efectuar um *teste de ajustamento global* (teste F) à regressão ajustada. No nosso caso, e definindo $\mathcal{R}^2$ como o coeficiente de determinação populacional, tem-se:

Hipóteses: $H_0 : \mathcal{R}^2 = 0$ vs. $H_1 : \mathcal{R}^2 > 0$

Estatística do teste: $F = \frac{QMR}{QMRE} = (n-2)\,\frac{R^2}{1-R^2} \cap F_{(1,n-2)}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral direita): Rej. H_0 se $F_{calc} > f_{\alpha\,(1,n-2)} = f_{0.05(1,148)} = 3.905$.

Conclusões: O valor calculado da estatística é: $F_{calc} = 148 \times \frac{0.9271}{1-0.9271} = 1882.178$. Logo, rejeita-se claramente a hipótese nula, que corresponde à hipótese dum ajustamento inútil do modelo. A resposta é coerente com a alínea anterior.

NOTA: Repare-se que o comando `summary` do R, quando aplicado ao ajustamento duma regressão, indica na última linha das listagens o valor da estatística calculada F_{calc}, os respectivos graus de liberdade associados, e o valor de prova (*p-value*) correspondente.

(h) A largura esperada duma pétala cujo comprimento seja $x = 4.5cm$ é dada por $\hat{\mu} = b_0 + b_1\,4.5 = -0.363076 + 0.415755 \times 4.5 = 1.507821$. No R, este resultado pode ser obtido através do comando `predict`:

```
> predict(iris.lm, new=data.frame(Petal.Length=4.5))
       1
1.507824
```

O intervalo de confiança para $\mu_{x=4.5} = E[Y|X = 4.5]$ é dado por (Acetato 141):

$$\left]\ (b_0+b_1x) - t_{\frac{\alpha}{2};n-2}\sqrt{QMRE\left[\frac{1}{n}+\frac{(x-\overline{x})^2}{(n-1)\,s_x^2}\right]}\ ,\quad (b_0+b_1x) + t_{\frac{\alpha}{2};n-2}\sqrt{QMRE\left[\frac{1}{n}+\frac{(x-\overline{x})^2}{(n-1)\,s_x^2}\right]}\ \right[$$

em que $\hat{\mu} = b_0 + b_1\,4.5 = 1.507821$, $t_{\frac{\alpha}{2};n-2} = t_{0.025,148} = 1.976122$, $QMRE = 0.2065^2$ (a partir da listagem acima dada). Por outro lado, a média e variância das $n = 150$ observações do preditor `Petal.Length` podem ser calculadas e resultam ser $\overline{x} = 3.758$ e $s_x^2 = 3.116278$. Assim, a 95% de confiança, o verdadeiro valor de $\mu_{x=4.5} = E[Y|X = 4.5]$ faz parte do intervalo $]\,1.47166\,,\ 1.543982\,[$. No R este intervalo de confiança pode ser obtido através do comando

```
> predict(iris.lm, new=data.frame(Petal.Length=4.5), int="conf")
       fit      lwr      upr
1 1.507824 1.471666 1.543982
```

Os extremos do intervalo são dados pelos valores `lwr` (de *lower*) e `upr` (de *upper*).

(i) O intervalo *de predição* para o valor da variável resposta y (largura da pétala) associada a *uma* observação com $x = 4.5$ é dado por:

$$\left]\ (b_0+b_1x) - t_{\frac{\alpha}{2};n-2}\sqrt{QMRE\left[1+\frac{1}{n}+\frac{(x-\overline{x})^2}{(n-1)\,s_x^2}\right]}\ ,\quad (b_0+b_1x) + t_{\frac{\alpha}{2};n-2}\sqrt{QMRE\left[1+\frac{1}{n}+\frac{(x-\overline{x})^2}{(n-1)\,s_x^2}\right]}\ \right[$$

Em relação ao intervalo de confiança pedido na alínea anterior, apenas muda a expressão debaixo da raíz quadrada. No R este tipo de intervalo obtém-se com um comando muito semelhante ao anterior:

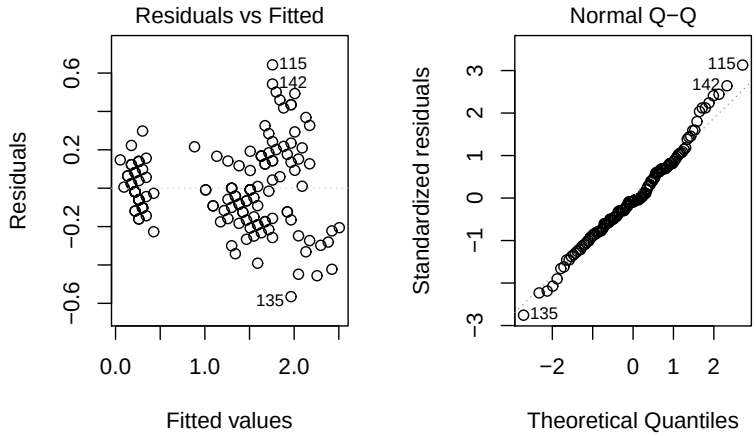
```
> predict(iris.lm, new=data.frame(Petal.Length=4.5), int="pred")
       fit      lwr      upr
1 1.507824 1.098187 1.917461
```

Como seria de esperar, trata-se dum intervalo bastante mais amplo: $]\,1.098187\,,\ 1.917461\,[$.

(j) Dos gráficos de resíduos produzidos pelo comando

```
> plot(lm(Petal.Width ~ Petal.Length, data=iris),which=c(1,2))
```

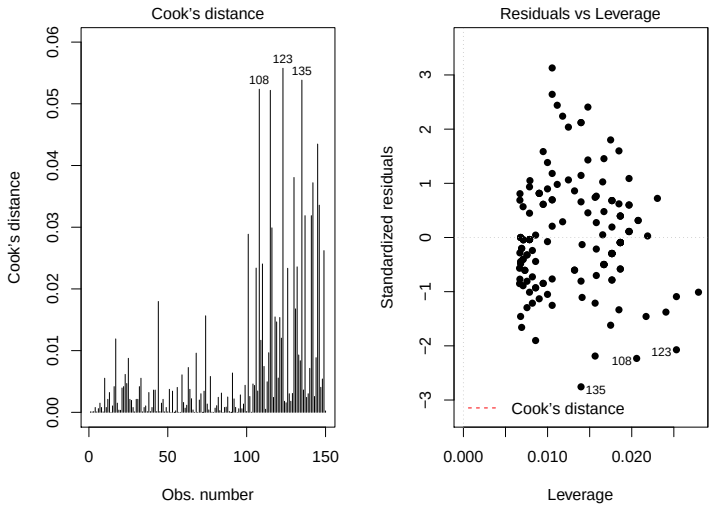
verifica-se que pode existir um problema em relação à hipótese de homogeneidade de variâncias. O gráfico da esquerda sugere que os lírios com comprimento de pétala mais pequeno (do lado esquerdo do gráfico) parecem ter menor variabilidade dos resíduos do que os restantes. Já a linearidade aproximada no *qq-plot* (gráfico da direita) não indicia a existência de problemas com a hipótese de normalidade. Igualmente, não se verificam observações com resíduos muito elevados, não havendo indícios de observações atípicas.



Quanto aos gráficos de diagnóstico produzidos pelo comando

```
> plot(lm(Petal.Width ~ Petal.Length, data=iris),which=c(4,5))
```

observa-se no diagrama de barras das distâncias de Cook que, apesar de haver alguma variabilidade nos valores, em nenhum caso a distância de Cook excede o valor (bastante baixo) de 0.06. Assim, nenhuma observação se deve considerar influente. De igual forma, não há valores elevados do efeito alavanca (*leverage*), sendo o maior valor de h_{ii} inferior a 0.03 (ver o eixo horizontal do gráfico da direita). Assim, nenhuma observação se destaca por ter um efeito alavanca elevado.



(k) Nas três subalíneas, as transformações de uma ou ambas as variáveis são transformações afins (lineares), razão pela qual o quadrado do coeficiente de correlação, ou seja, o coeficiente de determinação R^2 não sofre alteração. O que pode mudar são os parâmetros da recta de regressão ajustada.

i. Neste caso, apenas a variável preditora sofre uma transformação multiplicativa, da forma $x \to x^* = c\,x$ (com $c = 10$). Vejamos qual o efeito deste tipo de transformações

nos parâmetros da recta de regressão. Utilizando a habitual notação dos asteriscos para indicar os valores correspondentes à transformação, temos (tendo em conta que $var(c\,x) = c^2\,var(x)$:

$$b_1^* = \frac{cov_{x^*\,y}}{s^2_{x^*}} = \frac{cov(c\,x, y)}{c^2\,s^2_x} = \frac{1}{c}\,\frac{cov(x,y)}{s^2_x} = \frac{1}{c}\,b_1 \ ;$$

e (tendo em conta o efeito de constantes multiplicativas sobre a média, ou seja, $\overline{x^*} = c\,\overline{x}$):

$$b_0^* \ = \ \overline{y} - b_1^*\,\overline{x^*} \ = \ \overline{y} - \frac{1}{c}\,b_1 \cdot c\,\overline{x} \ = \ \overline{y} - b_1\overline{x} \ = \ b_0 \ .$$

Ou seja, neste caso a ordenada na origem não se altera, enquanto que o declive vem multiplicado por $\frac{1}{10}$. Confirmemos estes resultados com recurso ao R:

```
> lm(formula = Petal.Width ~ I(Petal.Length*10), data = iris)
Call:
lm(formula = Petal.Width ~ I(Petal.Length * 10), data = iris)
Coefficients:
         (Intercept)  I(Petal.Length * 10)
            -0.36308               0.04158
```

ii. Neste caso, estamos perante uma transformação idêntica à usada na alínea 1i), pelo que já sabemos que iremos encontrar, quer a ordenada na origem, quer o declive, multiplicados por $c = 10$. Confirmando no R:

```
> lm(formula = I(Petal.Width*10) ~ Petal.Length, data = iris)
Call:
lm(formula = I(Petal.Width * 10) ~ Petal.Length, data = iris)
Coefficients:
 (Intercept)  Petal.Length
      -3.631         4.158
```

iii. Finalmente, na conjugação das duas transformações discutidas nas subalíneas anteriores, e generalizando para as transformações multiplicativas $x \to c\,x$ e $y \to d\,y$, vem:

$$b_1^* = \frac{cov_{x^*\,y^*}}{s^2_{x^*}} = \frac{cov(c\,x, d\,y)}{c^2\,s^2_x} = \frac{cd}{c^2}\,\frac{cov(x,y)}{s^2_x} = \frac{d}{c}\,b_1 \ ;$$

e:

$$b_0^* = \overline{y^*} - b_1^*\,\overline{x^*} = d\,\overline{y} - \frac{d}{c}\,b_1 \cdot c\,\overline{x} = d\,(\overline{y} - b_1\overline{x}) = d\,b_0 \ .$$

Como no nosso caso $c = d = 10$, o declive não se deve alterar, enquanto a ordenada na origem deverá ser 10 vezes maior do que no caso original dos dados não transformados.

```
> lm(formula = I(Petal.Width*10) ~ I(Petal.Length*10), data = iris)
Call:
lm(formula = I(Petal.Width * 10) ~ I(Petal.Length * 10), data = iris)
Coefficients:
         (Intercept)  I(Petal.Length * 10)
             -3.6308                0.4158
```

15. (a) Tem-se, recordando que $SQRE = SQT - SQR$,

$$F = \frac{QMR}{QMRE} = \frac{SQR/1}{SQRE/(n-2)} = (n-2)\,\frac{SQR}{SQT - SQR} = (n-2)\,\frac{R^2}{1-R^2} \ ,$$

onde a última passagem resulta de dividir numerador e denominador por SQT.

(b) Como R^2 está entre 0 e 1, qualquer aumento de R^2 aumenta o numerador e diminui o denominador, provocando um aumento da fracção. Assim, a maiores valores de R^2 correspondem maiores valores da estatística F. Uma vez que o teste F tem hipótese nula $H_0 : \mathcal{R}^2 = 0$, é natural que se defina uma região crítica unilateral direita.

16. Recordando a expressão para SQR obtida no Exercício 5d), tem-se:

$$T = \frac{\hat{\beta}_1}{\sqrt{\frac{QMRE}{(n-1)\,s_x^2}}} \qquad \Longrightarrow \qquad T^2 = \frac{\hat{\beta}_1^2\,(n-1)\,s_x^2}{QMRE} = \frac{SQR}{QMRE} = \frac{QMR}{QMRE}\ .$$

Nos apontamentos da disciplina de Estatística (dos primeiros ciclos do ISA), foi visto (Apontamentos da Prof. Manuela Neves, p.119, na versão de 2011/12) que, dada uma variável aleatória com distribuição t-Student, $X \cap t_m$, o seu quadrado tem distribuição F, com graus de liberdade como indicado de seguida: $X^2 \cap F_{(1,m)}$. No nosso caso, $m = n-2$. Assim, numa regressão linear simples, usar um teste-t para testar $\beta_1 = 0$, ou um teste F de ajustamento global, é equivalente.

17. (a) Admitir que existem erros aleatórios aditivos no modelo linearizado não é a mesma coisa que admitir que existem erros aditivos no modelo original. De facto,

$$\log(Y) = \beta_0 + \beta_1\,\log(x) + \epsilon \quad \Leftrightarrow \quad Y = \mathrm{e}^{\beta_0+\beta_1\,\log(x)+\epsilon} = \mathrm{e}^{\beta_0}\cdot \mathrm{e}^{\log(\beta_1\,x)}\cdot \mathrm{e}^{\epsilon} = \beta_0^*\cdot \mathrm{x}^{\beta_1}\cdot \epsilon^*\ ,$$

pelo que admitir erros aditivos no modelo linearizado corresponde a admitir erros multiplicativos no modelo exponencial original. Além disso, admitir que os erros aditivos ϵ do modelo linearizado têm distribuição Normal significa que $\epsilon^* = \mathrm{e}^{\epsilon}$ **não** tem distribuição Normal (a sua distribuição é a chamada Lognormal, não estudada nesta disciplina). A ideia importante a reter é que *admitir as hipóteses usuais no modelo original é diferente de admitir essas mesmas hipóteses no modelo linearizado.*

(b) Na alínea referida foi ajustado o modelo linearizado, ou seja a regressão linear entre log(`brain`) (variável resposta) e log(`body`) (variável preditora). A parte final do ajustamento produzido no `R` com o comando `summary` é indicada de seguida.

```
> Animals.lm <- lm(log(brain) ~ log(body) , data=Animals)
> summary(Animals.lm)
(...)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.55490    0.41314   6.184 1.53e-06
log(body)    0.49599    0.07817   6.345 1.02e-06
---
Residual standard error: 1.532 on 26 degrees of freedom
Multiple R-squared: 0.6076,Adjusted R-squared: 0.5925
F-statistic: 40.26 on 1 and 26 DF,  p-value: 1.017e-06
```

Utilizar-se-á a informação acima para efectuar o teste global de ajustamento (teste F global). As hipóteses do teste podem ser escritas de formas diferentes, e nesta resolução é usada a que relaciona as hipóteses deste teste com o declive da recta de regressão populacional.

Hipóteses: $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$
Estatística do teste: $F = \frac{QMR}{QMRE} = (n-2)\,\frac{R^2}{1-R^2} \ \cap\ F_{(1,n-2)}$, sob H_0.
Nível de significância: $\alpha = 0.05$.
Região Crítica (Unilateral direita): Rej. H_0 se $F_{calc} > f_{\alpha\,(1,n-2)} = f_{0.05(1,26)} = 4.225201$.

Conclusões: O valor calculado da estatística é: $F_{calc} = 40.26$. Logo, rejeita-se claramente a hipótese nula, que corresponde à hipótese dum ajustamento inútil do modelo. A resposta é coerente com a alínea anterior.

O Coeficiente de Determinação é $R^2 = 0.6076$, um valor relativamente baixo. Tal facto não é contraditório com uma rejeição enfática da hipótese nula do teste F de ajustamento global (o valor de prova é $p = 1.017 \times 10^{-6}$), uma vez que a hipótese nula desse teste pode ser formulada como "na população, o coeficiente de correlação (ao quadrado) entre $\ln(x)$ e $\ln(y)$ é nulo". Esta hipótese nula é muito fraca, indicando a inutilidade do modelo linear. O valor amostral observado de $R^2 = 0.6076$, não sendo elevado, é no entanto suficiente para rejeitar $H_0 : \mathcal{R}^2 = 0$, ou seja, difere significativamente de zero para qualquer dos níveis de significância usuais.

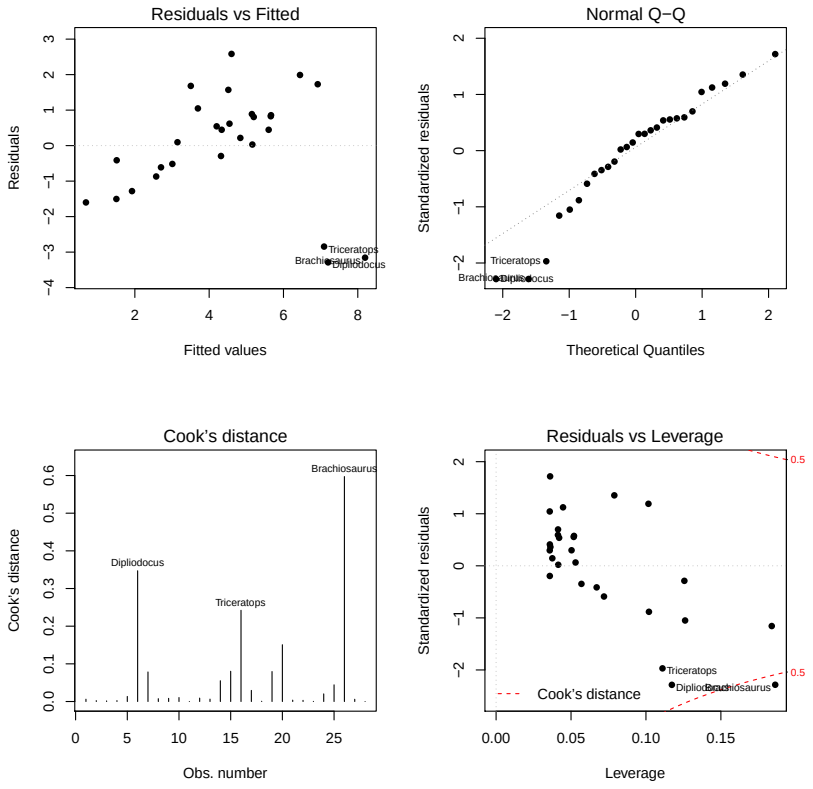
(c) Pretende-se o intervalo a 95% de confiança para β_1, ou seja:

$$\left] \; b_1 - t_{\frac{\alpha}{2}(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1} \;\; , \;\; b_1 + t_{\frac{\alpha}{2}(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1} \; \right[\, ,$$

com $b_1 = 0.49599$, $t_{0.025(26)} = 2.055529$ e $\hat{\sigma}_{\hat{\beta}_1} = 0.07817$. Ou seja, o intervalo é $]\,0.335\,,\,0.657\,[$. Uma relação isométrica corresponde a admitir que o declive da recta populacional é $\beta_1 = 1$, ou seja que as taxas de variação relativas de peso do corpo e peso do cérebro são iguais (ver a resolução do exercício 9). Uma vez que o valor 1 não pertence ao intervalo de confiança, a hipótese de isometria não é admissível (a 95% de confiança).

(d) Os quatro gráficos discutidos nas aulas teóricas resultam do comando

```
> plot(Animals.lm, which=c(1,2,4,5), pch=16)
```



Como se pode constatar, a presença das três observações atípicas (os dinossáurios) é evidente em todos os gráficos. No primeiro (resíduos e_i vs. valores ajustados $\hat{y}_i$) o efeito traduz-se no facto dos restantes resíduos se disporem numa banda inclinada (e não horizontal, como seria adequado). No segundo gráfico, o *qq-plot* indica que os dinossáurios são responsáveis pelo maior afastamento em relação à linearidade aproximada que seria de esperar perante uma distribuição aproximadamente Normal dos resíduos. As distâncias de Cook dessas mesmas observações são claramente grandes, sendo que no caso do *Brachiosaurus* ultrapassam mesmo o nível de guarda 0.5. Recorde-se que as distâncias de Cook procuram medir o efeito sobre o ajustamento que resulta de retirar *uma* observação, sendo de realçar que apesar de haver três observações atípicas próximas umas das outras, basta retirar uma para que haja já diferenças assinaláveis no ajustamento. Finalmente, no quarto gráfico, de resíduos standardizados contra valores do efeito alavanca (*leverage*), verifica-se que o maior efeito alavanca é cerca de 0.2. Tendo em conta que em princípio este valor poderia atingir o valor máximo 1 (aqui não há repetições dos valores de x_i), trata-se dum valor que não parece demasiado elevado. Convém recordar que numa regressão linear simples, as *leverages* h_{ii} são função do afastamento do valor do preditor x em relação à média $\overline{x}$ das observações desse preditor.

(e) Ajustando agora as 25 espécies que não são dinossáurios, obtêm-se os seguintes resultados:

```
> Animals.lm25 <- lm(log(brain) ~ log(body) , data=Animals[-c(6,16,26),])
> summary(Animals.lm25)
(...)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.15041    0.20060   10.72 2.03e-10
log(body)    0.75226    0.04572   16.45 3.24e-14
---
Residual standard error: 0.7258 on 23 degrees of freedom
Multiple R-squared: 0.9217,Adjusted R-squared: 0.9183
F-statistic: 270.7 on 1 and 23 DF,  p-value: 3.243e-14
```

Os parâmetros estimados da recta alteraram-se, e os respectivos erros padrão são agora bastante mais pequenos, factos que estão associados a uma relação linear muito mais forte nas 25 espécies usadas neste ajustamento. Esta relação muito mais forte é confirmada pelo valor muito mais elevado do coeficiente de correlação: $R^2 = 0.9217$, e é visível no gráfico de log-peso do cérebro contra log-peso do corpo, indicado na resolução do exercício 9.

A expressão do intervalo de confiança é a mesma que indicada na alínea 17c), mas agora os valores das quantidades relevantes são: $b_1 = 0.75226$, $t_{0.025(23)} = 2.068658$ (repare-se na mudança dos graus de liberdade, resultante de agora haver apenas $n = 25$ espécies) e $\hat{\sigma}_{\hat{\beta}_1} = 0.04572$. Assim, o IC é agora $]\,0.6577\,,\,0.8468\,[$. Note-se que este intervalo é mais apertado (mais preciso) que o correspondente intervalo obtido na alínea c), o que reflecte o menor erro padrão agora existente. No entanto, e apesar do maior valor do declive estimado, $b_1 = 0.75226$, o intervalo a 95% de confiança continua a não incluir o valor 1 como um valor admissível para β_1, logo a hipótese de isometria continua a não ser admissível.

(f) O valor esperado para log-peso do cérebro, numa espécie com peso do corpo igual a 250, e portanto *log*-peso do corpo $x^* = log(250) = 5.521461$ será: $\hat{\mu}_{Y^*|X^*=log(250)} = b_0 + b_1 \cdot log(250) = 2.15041 + 0.75226 \cdot 5.521461 = 6.303984$. Um intervalo a $(1-\alpha) \times 100\%$ de

confiança para o verdadeiro valor de $E[Y^*|X^* = log(250)]$ será:

$$\left]\ (b_0+b_1x^*) - t_{\frac{\alpha}{2};n-2}\sqrt{QMRE\ \left[\frac{1}{n}+\frac{(x^*-\overline{x^*})^2}{(n-1)\,s^2_{x*}}\right]}\ ,\quad (b_0+b_1x^*) + t_{\frac{\alpha}{2};n-2}\sqrt{QMRE\ \left[\frac{1}{n}+\frac{(x^*-\overline{x^*})^2}{(n-1)\,s^2_{x*}}\right]}\ \right[$$

Os valores de b_0 e b_1 já foram indicados, tal como o número de observações $n = 25$ e $t_{0.025(23)} = 2.068658$. Por outro lado, e tendo em conta que sob a designação *residual Standard error*, a listagem produzida pelo `R` dá o valor da *raíz quadrada* do $QMRE$, tem-se: $QMRE = 0.7258^2 = 0.5267856$. Finalmente, o valor da média e a variância das observações do preditor dizem agora respeito aos *log*-pesos do cérebro, sendo, respectivamente:

```
> mean(log(Animals$body[-c(6,16,26)]))
[1] 3.028283
> var(log(Animals$body[-c(6,16,26)]))
[1] 10.50226
```

Com base neste valores, a raíz quadrada acima indicada tem valor

$$\sqrt{0.5267856 \cdot \left[\frac{1}{25} + \frac{(5.521461 - 3.028283)^2}{24 * 10.50226}\right]} = 0.1845604\ .$$

Assim, o intervalo a 95% de confiança para o log-peso do cérebro esperado em espécies com peso do corpo 250 é $]\,5.922\,,\,6.686\,[$. No `R`, este intervalo de confiança poderia ser obtido através do comando

```
> predict(Animals.lm25, new=data.frame(body=250), int="conf")
      fit      lwr      upr
1 6.30399 5.922178 6.685803
```

Repare-se que, sendo necessário dar o novo valor da variável preditora com o nome da variável preditora original, foi dado o valor $x = 250$. O `R` tem em conta a transformação logarítmica usada no ajustamento da regressão linear em `Animals.lm25`.

(g) Agora, pretende-se um intervalo de predição para o log-peso do cérebro, Y^*, *duma única espécie* cujo peso do corpo seja $x = 250kg$ (e log-peso do corpo $x^* = \log(250)$). A expressão para este intervalo de predição a $(1-\alpha) \times 100\%$ é:

$$\left]\ (b_0+b_1x^*) - t_{\frac{\alpha}{2};n-2}\sqrt{QMRE\ \left[1+\frac{1}{n}+\frac{(x^*-\overline{x^*})^2}{(n-1)\,s^2_{x*}}\right]}\ ,\quad (b_0+b_1x^*) + t_{\frac{\alpha}{2};n-2}\sqrt{QMRE\ \left[1+\frac{1}{n}+\frac{(x^*-\overline{x^*})^2}{(n-1)\,s^2_{x*}}\right]}\ \right[$$

O valor da raíz quadrada é agora:

$$\sqrt{0.5267856 \cdot \left[1 + \frac{1}{25} + \frac{(5.521461 - 3.028283)^2}{24 * 10.50226}\right]} = 0.748898\ ,$$

pelo que o referido intervalo de predição é $]\,4.755\,,\,7.853\,[$. Como seria de esperar, trata-se dum intervalo bastante mais amplo que o anterior, uma vez que tem em conta a variabilidade adicional associada a observações individuais. No `R`, utilizar-se-ia o comando

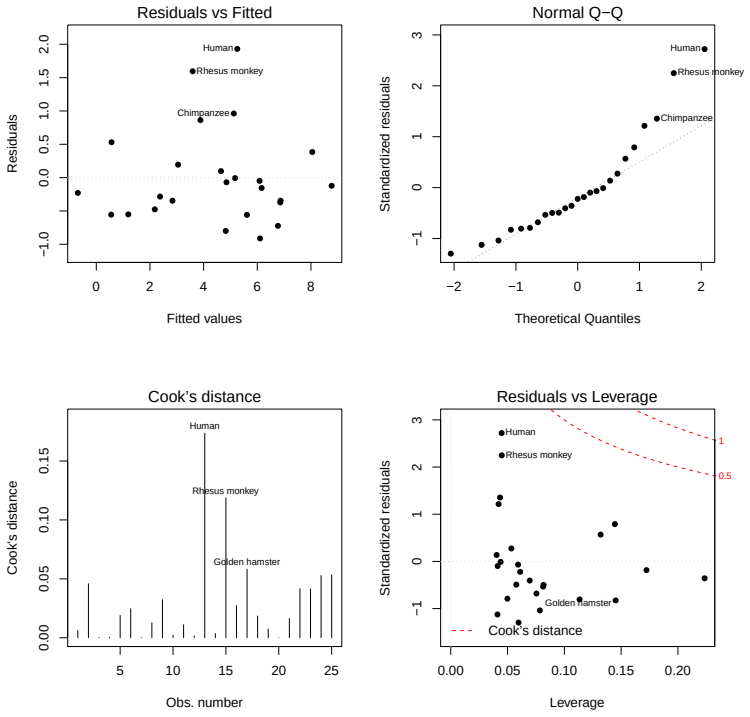
```
> predict(Animals.lm25, new=data.frame(body=250), int="pred")
      fit      lwr      upr
1 6.30399 4.754694 7.853287
```

Para obter o intervalo de predição para os valores do *peso do cérebro* (sem logaritmização), basta tomar as exponenciais dos extremos do intervalo acima referido. De facto, se (ao nível 95% e para $x = 250kg$) o intervalo de predição para $Y^* = \log(Y)$ é: $4.755 < log(Y) < 7.853$, então a dupla desigualdade equivalente $e^{4.755} = 116.16 < Y < 2573.443 = e^{7.853}$ será um intervalo de predição a 95% para uma observação individual de Y. Trata-se dum intervalo de grande amplitude, associado quer ao facto de ser um intervalo de predição para valores individuais de Y, quer à exponenciação.

NOTA: Na alínea anterior não se pode efectuar uma transformação análoga, uma vez que valor esperado e logaritmização não são operações intercambiáveis. Ou seja, $E[\log(Y)] \neq \log(E[Y])$, pelo que não sabemos como transformar a dupla desigualdade $a < E[\log(Y)] < b$ numa dupla desigualdade equivalente apenas com $E[Y]$ no meio.

(h) Os gráficos de resíduos e diagnósticos são dados pelo seguinte comando e são reproduzidos de seguida.

```
> plot(Animals.lm25, which=c(1,2,4,5), pch=16)
```

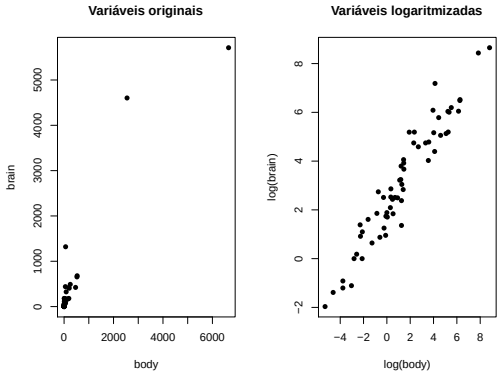


A exclusão dos dinossáurios do conjunto das espécies analisadas tornou saliente que, entre as 25 espécies restantes, duas se destacam por terem resíduos positivos um pouco maiores: o ser humano e o macaco *Rhesus*. Esse facto indica que o log-peso do cérebro destas espécies é razoavelmente maior do que seria de esperar dado o log-peso dos seus corpos. As duas espécies são igualmente salientes no *qq-plot* e têm distância de Cook elevada, embora longe dos níveis de guarda. No entanto, repare-se que os valores do efeito alavanca destas espécies com resíduos e distância de Cook mais elevados são muito baixos. Tal facto (que reflecte o facto de os log-pesos dos corpos destas espécies estarem próximos da média de log-pesos do corpo das espécies observadas) ilustra que os conceitos de influência, atipicidade e valor do efeito alavanca são diferentes. Uma eventual exclusão destas espécies (sobretudo no

caso do macaco *Rhesus*) já é mais problemática que no caso dos dinossáurios, uma vez que obrigaria a redefinir a população de interesse num sentido mais discutível. Nem tal deve ser feito apenas para "melhorar" o aspecto de gráficos de diagnóstico. Aliás, o que aconteceu acima ilustra que uma exclusão pode até fazer surgir novas espécies atípicas, influentes ou de elevado valor alavanca.

18. Para resolver este exercício, onde se considera um grupo de $n = 62$ espécies de mamíferos, é necessário ter previamente carregado o módulo `MASS`, o que se pode fazer através do comando `library(MASS)`.

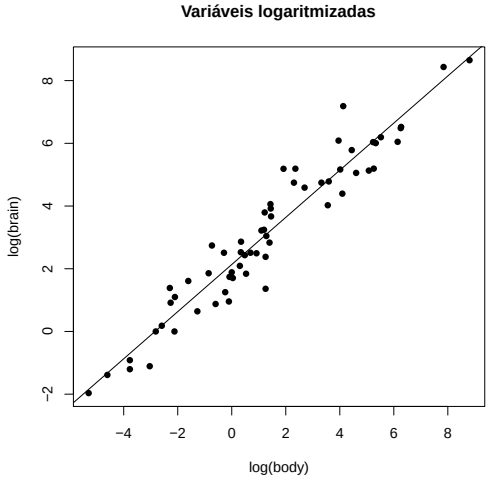
As nuvens de pontos pedidos nas duas alíneas iniciais são indicadas à direita. É evidente o efeito de linearização obtido através da logaritmização, quer do peso do corpo, quer do peso do cérebro. Tal linearização sugere que um modelo potência (alométrico) é adequado para descrever a relação entre peso do corpo e peso do cérebro, nos mamíferos.



(c) A resposta é idêntica à que foi dada no exercício 9.

(d) Os comandos para responder, no `R` são:

```
> mammals.lm <- lm(log(brain) ~ log(body), data=mammals)
> mammals.lm
Call: lm(formula = log(brain) ~ log(body), data = mammals)
Coefficients:
(Intercept)    log(body)
     2.1348       0.7517

> plot(log(brain) ~ log(body), data=mammals, pch=16, main="Variáveis logaritmizadas")
> abline(mammals.lm)
```



Note-se como os parâmetros da recta ajustada utilizando 62 espécies são muito próximos dos parâmetros obtidos utilizando apenas as 25 espécies (não dinossáurios) no Exercício 17, facto que sugere uma boa robustez do resultado obtido. A recta de regressão ajustada é uma boa síntese da nuvem de pontos.

(e) Como se pode constatar, o coeficiente de determinação é muito elevado ($R^2 = 0.9208$ e naturalmente muito significativamente diferente de zero, com *p-value* inferior a 2.2×10^{-16}, ou seja, inferior à precisão do computador), o que indica uma muito boa relação linear entre as variáveis logaritmizadas, logo uma boa relação potência do peso do cérebro e do peso do corpo.

```
> summary(mammals.lm)
Call: lm(formula = log(brain) ~ log(body), data = mammals)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.13479    0.09604   22.23   <2e-16 ***
log(body)    0.75169    0.02846   26.41   <2e-16 ***
---
Residual standard error: 0.6943 on 60 degrees of freedom
Multiple R-squared: 0.9208,Adjusted R-squared: 0.9195
F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

(f) Como em qualquer linearização dum modelo potência, o declive da recta é a potência estimada na relação $y = c\,x^d$, ou seja, o valor de d. No caso desta relação, esse valor estimado é aproximadamente $d = 0.75$, valor que confirma a relação das espécies não dinossáurios do exercício 17. Como foi visto nas aulas teóricas, esse valor corresponde a que a taxa de variação relativa do peso do cérebro seja 3/4 da taxa de variação relativa no peso do corpo.

(g) Os intervalos de confiança a 95% para ambos os parâmetros da recta são:

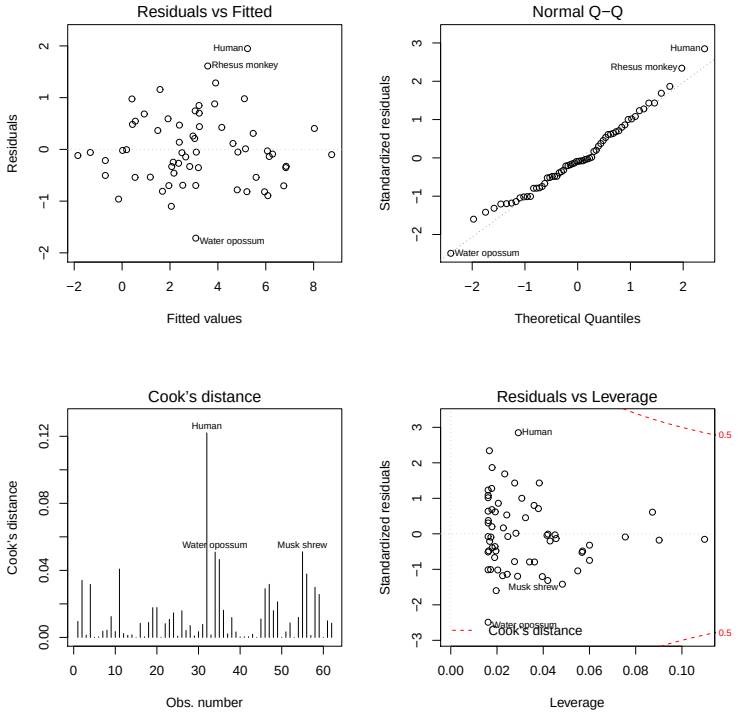
```
> confint(mammals.lm)
                2.5 %    97.5 %
(Intercept) 1.9426733 2.3269041
log(body)   0.6947503 0.8086215
```

Assim, o intervalo de confiança para o declive da recta populacional entre log-peso do corpo e log-peso do cérebro é $]\,0.695\,,\,0.807\,[$. O intervalo não inclui o valor 1 que corresponderia à isometria, ou seja a uma taxa de variação relativa igual entre peso do corpo e peso do cérebro.

(h) Os gráficos de resíduos e diagnósticos obtêm-se com o comando

```
> plot(mammals.lm, which=c(1,2,4,5), add.smooth=FALSE)
```

e são indicados a seguir. Nenhum dos gráficos indicia problemas com os pressupostos do modelo linear, nem observações dignas de especial destaque. Apesar do ser humano surgir com algum destaque em vários gráficos, não se distingue de forma que justifique qualquer reparo especial.



19. Tem-se

$$\begin{aligned} V[\hat{\mu}_{Y|_x}] &= V[\hat{\beta}_0 + \hat{\beta}_1\, x] = V[\hat{\beta}_0] + V[\hat{\beta}_1\, x] + 2cov(\hat{\beta}_0, \hat{\beta}_1\, x) \\ &= \underbrace{\sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right]}_{=V[\hat{\beta}_0]} + x^2 \cdot \underbrace{\frac{\sigma^2}{(n-1)s_x^2}}_{=V[\hat{\beta}_1]} + 2x \cdot \underbrace{\frac{-\sigma^2 \overline{x}}{(n-1)s_x^2}}_{=Cov[\hat{\beta}_0,\hat{\beta}_1]\ (Ex.12)} \\ &= \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2 + x^2 - 2\overline{x}\, x}{(n-1)\, s_x^2}\right] = \sigma^2 \left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\, s_x^2}\right] . \end{aligned}$$

20. (a) Pretende-se calcular $Cov[Y_i, \hat{Y}_i]$. Relembrando que $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ e $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$, pelas propriedades da covariância, tem-se:

$$\begin{aligned} Cov[Y_i, \hat{Y}_i] &= Cov[Y_i, \hat{\beta}_0 + \hat{\beta}_1 x_i] = Cov[Y_i, \hat{\beta}_0] + Cov[Y_i, \hat{\beta}_1 x_i] \\ &= Cov[Y_i, \overline{Y} - \hat{\beta}_1 \overline{x}] + x_i Cov[Y_i, \hat{\beta}_1] = Cov[Y_i, \overline{Y}] - Cov[Y_i, \hat{\beta}_1 \overline{x}] + x_i Cov[Y_i, \hat{\beta}_1] \\ &= Cov[Y_i, \overline{Y}] - \overline{x} Cov[Y_i, \hat{\beta}_1] + x_i Cov[Y_i, \hat{\beta}_1] = Cov[Y_i, \overline{Y}] + (x_i - \overline{x}) Cov[Y_i, \hat{\beta}_1] \end{aligned}$$

Como $\overline{Y} = \frac{1}{n} \sum_{j=1}^{n} Y_j$ e $\hat{\beta}_1 = \sum_{j=1}^{n} c_j Y_j$, com $c_j = \frac{(x_j - \overline{x})}{(n-1)s_x^2}$,

$$\begin{aligned} Cov[Y_i, \hat{Y}_i] &= Cov[Y_i, \tfrac{1}{n} \sum_{j=1}^{n} Y_j] + (x_i - \overline{x}) Cov[Y_i, \sum_{j=1}^{n} c_j Y_j] \\ &= \tfrac{1}{n} \sum_{j=1}^{n} Cov[Y_i, Y_j] + (x_i - \overline{x}) \sum_{j=1}^{n} c_j Cov[Y_i, Y_j] \end{aligned}$$

Dado as observações $\{Y_i\}_{i=1}^n$ serem v.a. independentes, $Cov[Y_i, Y_j] = 0$, se $i \neq j$.
Além disso, $Cov[Y_i, Y_i] = var[Y_i] = \sigma^2$, pelo que

$$Cov[Y_i, \hat{Y}_i] = \frac{\sigma^2}{n} + (x_i - \overline{x})c_i\sigma^2 = \frac{\sigma^2}{n} + \frac{(x_i - \overline{x})^2\sigma^2}{(n-1)s_x^2} = \sigma^2 \left[\frac{1}{n} + \frac{(x_i - \overline{x})^2}{(n-1)s_x^2}\right] = \sigma^2\, h_{ii}.$$

(b) Sabemos que $E_i = Y_i - \hat{Y}_i$. Pelas propriedades da covariância e a alínea anterior, temos:

$$\begin{aligned} Cov[Y_i, E_i] = \; & Cov[Y_i, Y_i - \hat{Y}_i] \;=\; Cov[Y_i, Y_i] - Cov[Y_i, \hat{Y}_i] \\ = \; & \sigma^2 - \sigma^2\, h_{ii} \;=\; \sigma^2\,[1 - h_{ii}] \end{aligned}$$

(c) De acordo com o resultado da alínea a) e como

$$Cov[\hat{Y}_i, \hat{Y}_i] = V[\hat{Y}_i] = V[\hat{\beta}_0 + \hat{\beta}_1 x_i] = V[\hat{\mu}_{Y|x_i}] \underset{\text{(ex. 19)}}{=} \sigma^2 \left[\frac{1}{n} + \frac{(x_i - \overline{x})^2}{(n-1)s_x^2}\right] \;=\; \sigma^2\, h_{ii},$$

tem-se:

$$\begin{aligned} Cov[\hat{Y}_i, E_i] = Cov[\hat{Y}_i, Y_i - \hat{Y}_i] \;=\; & Cov[\hat{Y}_i, Y_i] - Cov[\hat{Y}_i, \hat{Y}_i] \\ = \; & \sigma^2\, h_{ii} - \sigma^2\, h_{ii} \;=\; 0. \end{aligned}$$

Deste modo, se o modelo de RLS for válido, não deverá haver nenhum padrão no gráfico de resíduos vs. valores ajustados de Y já que o valor da covariância entre estas duas variáveis é zero. O mesmo não acontece no gráfico de resíduos vs. valores observados de Y pois, como mostrámos na alínea anterior, a covariância entre E_i e Y_i é, em geral, diferente de zero.

(d) Como se viu nas aulas teóricas, cada resíduo pode escrever-se como combinação linear das observações Y_i,

$$E_i = \sum_{j=1}^{n} k_j Y_j, \text{ com } k_j = \begin{cases} -(d_j + x_i c_j) & \text{se} \quad j \neq i \\ 1 - (d_j + x_i c_j) & \text{se} \quad j = i \end{cases}$$

E_i é então combinação linear de v.a.s Normais independentes, logo tem ainda distribuição Normal. Relativamente aos parâmetros, temos que

$$E[E_i] = E[Y_i - \hat{Y}_i] = E[Y_i] - E[\hat{Y}_i] = \underbrace{(\beta_0 + \beta_1 x_i)}_{E[Y_i]} - \underbrace{(\beta_0 + \beta_1 x_i)}_{E[\hat{Y}_i]} = 0$$

$$\begin{aligned} V[E_i] = V[Y_i - \hat{Y}_i] &= V[Y_i] + V[\hat{Y}_i] - 2\, Cov[Y_i, \hat{Y}_i] \\ &= \sigma^2 + \underbrace{\sigma^2\, h_{ii}}_{\text{(ex. 19)}} - 2 \underbrace{\sigma^2\, h_{ii}}_{\text{(alínea a)}} \\ &= \; \sigma^2(1 - h_{ii}), \end{aligned}$$

com $h_{ii} = \dfrac{1}{n} + \dfrac{(x_i - \overline{x})^2}{(n-1)s_x^2}$.

21. (a) Com base na expressão da alínea 5d), temos:

$$E[SQR] \;=\; E[\hat{\beta}_1^2 \cdot (n-1)s_x^2] \;=\; (n-1)s_x^2 \cdot E[\hat{\beta}_1^2] \;,$$

Ora, sabemos que, para qualquer variável aleatória X,

$$V[X] = E[X^2] - E^2[X] \qquad \Longleftrightarrow \qquad E[X^2] = V[X] + E^2[X] \;.$$

Tomando $X = \hat{\beta}_1$, temos

$$E[\hat{\beta}_1^2] = V[\hat{\beta}_1] + E^2[\hat{\beta}_1] = \frac{\sigma^2}{(n-1)s_x^2} + \beta_1^2 \qquad \Longrightarrow \qquad E[SQR] = \sigma^2 + \beta_1^2 \cdot (n-1)s_x^2 \;.$$

(b) Já vimos que, em qualquer regressão, $E[QMRE] = \sigma^2$. Vimos agora que, numa regressão linear simples, $E[QMR] = E[SQR/1] = E[SQR] = \sigma^2 + \beta_1^2 \cdot (n-1)s_x^2$. Assim,

$$\begin{array}{ll} \text{se } \beta_1 = 0 & \Longrightarrow \quad E[QMR] = E[QMRE] \\ \text{se } \beta_1 \neq 0 & \Longrightarrow \quad E[QMR] > E[QMRE] \end{array}$$

Logo, é natural que a estatística $F = \frac{QMR}{QMRE}$ tome valores próximos de 1 caso seja verdade $H_0 : \beta_1 = 0$. Valores muito grandes de F_{calc} fazem suspeitar que H_0 não seja verdadeira, devendo portanto a região crítica do teste ser unilateral direita.