INSTITUTO SUPERIOR DE AGRONOMIA

MODELOS MATEMÁTICOS e APLICAÇÕES– 2016-17

Resoluções de exercícios de Regressão Linear

1. Escreva, numa sessão do R, o comando indicado no enunciado:

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

Para ver *o conteúdo* do objecto `Cereais` acabado de criar, 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 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 $SQR = (n-1)\, s^2_{\hat{y}}$, 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 $cov(x, cy) = c\, cov(x, y)$):

$$b^*_1 = \frac{cov_{x\,y^*}}{s^2_x} = \frac{cov(x, c\,y)}{s^2_x} = c\, \frac{cov(x, y)}{s^2_x} = c\, b_1 \ ;$$

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. 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 $1.596\,hl$ 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 ($b_0 = -5151.793\,hl$). 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. 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
```

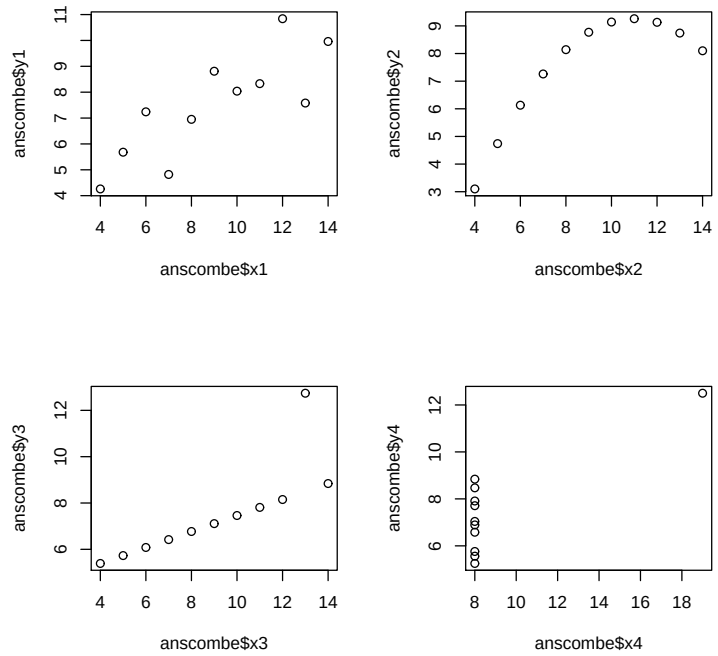
(c) As quatro rectas pedidas têm equação quase idêntica, aproximadamente $y = 3 + 0.5\,x$:

```
> 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.



4. 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, uma relação linear entre $ln(y)$ e $ln(x)$ corresponde a uma relação potência (alométrica) entre as variáveis originais: $y = c\,x^d$. 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).

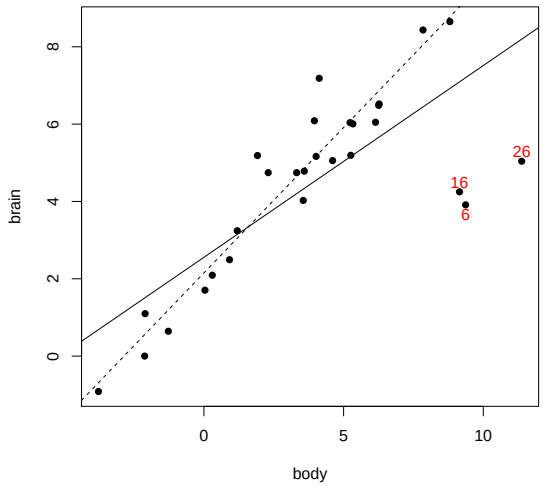
NOTA: É necessário explicitar as coordenadas dos pontos no gráfico que se vai aceder com o comando. No nosso caso, isso significa explicitar as coordenadas dos dados logaritmizados: `log(Animals)`.
O enunciado pede para identificar os pontos que se destacam da relação linear, e que são os pontos 6, 16 e 26. Selecionando as linhas com esses números podemos identificar as espécies em questão, e verificar que se trata de espécies de dinossáurios:

```
> 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:

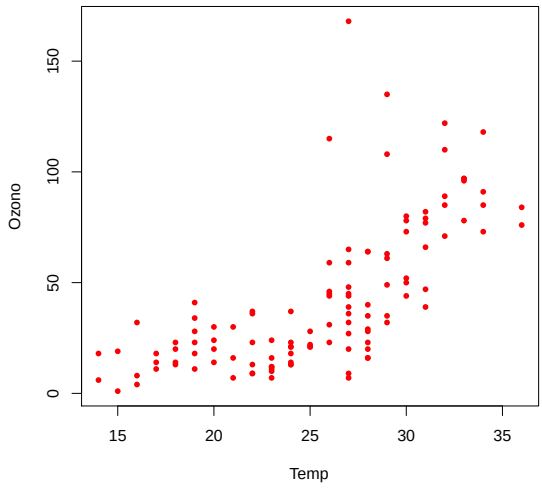
```
> summary(Animals.loglm.sub)
(...)
Multiple R-squared: 0.9217
(...)
```

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 4f), 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.

5. (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
```

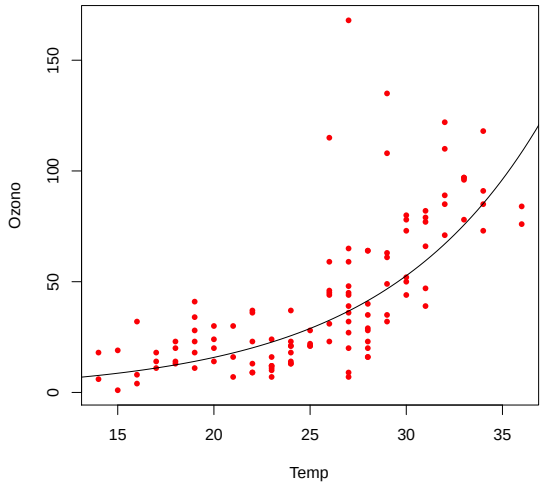
O coeficiente de determinação é de cerca de $R^2 = 0.73^2 = 0.53$ (aplicando o comando `summary` ao modelo agora ajustado verifica-se ser $R^2 = 0.5372$), o que significa que a regressão explica pouco mais de 53% da variabilidade dos log-teores de ozono.

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. a recta relaciona log-ozono com temperatura. Logo, 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: $\mathrm{e}^{3.3633} = 28.8843$.

(c) O comando que ajusta a curva exponencial à nuvem de pontos de ozono vs. temperaturas (admitindo que este gráfico ainda está activo) pode ser o seguinte:

```
> curve(1.4273*exp(0.1203*x), from=10, to=40, add=TRUE)
```



6. (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, o modelo linearizado ajusta-se através do comando:

```
> lm(I(1/rate) ~ I(1/conc), data=Puromycin)
```

sendo os resultados obtidos os seguintes:

```
Coefficients:
(Intercept)    I(1/conc)
  0.0059734    0.0002329
```

(c) Tendo em conta as relações vistas na alínea anterior, $b_0^* = \frac{1}{a} = 0.0059734$, tem-se $a = 167.4088$. Por outro lado, $b_1^* = \frac{b}{a} = 0.0002329$, logo $b = 0.0002329 \times 167.4088 = 0.03898951$. Assim, o modelo de Michaelis-Menten ajustado é: $y = \frac{167.4088\,x}{0.03898951+x}$. Repare-se que o limite de y quando x tende para $+\infty$ é 167.4088, que é assim a estimativa da assintota superior da relação de Michaelis-Menten. O gráfico da relação original sugere que se trata duma subestimação do verdadeiro valor desta assintota horizontal. Esta subestimação resulta do facto de os recíprocos de números menores (como são as taxas iniciais) serem maiores do que os recíprocos de números maiores, como são as taxas próximas da assíntota horizontal ($\frac{1}{47} = 0.0212766$ é cerca de quatro vezes maior que $\frac{1}{207} = 0.004830918$). Assim, o ajustamento de mínimos quadrados, na escala dos recíprocos, vai dar mais atenção às observações associadas às concentrações mais baixas do que às observações associadas à definição da assíntota. Este exemplo ilustra que pode haver inconvenientes associados à utilização de transformações linearizantes, como indicado nas aulas.

7. (a) A "matriz de nuvens de pontos" produzida pelo comando `plot(vinho.RLM)` tem as nuvens de pontos associadas a cada possível par de entre as $p = 13$ variáveis do conjunto de dados. Na linha indicada pela designação `V8` encontram-se os gráficos em que essa variável surge no eixo vertical. A modelação de `V8` com base num único preditor parece promissor apenas com o preditor `V7` (o que não deixa de ser natural, visto `V7` ser o índice de fenóis totais, sendo `V8` o teor de flavonóides, ou seja, um dos fenóis medidos pela variável `V7`).

(b) O ajustamento pedido é:

```
> summary(lm(V8 ~ V2, data=vinho.RLM))
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.75876    1.17370  -1.498  0.13580
V2           0.29137    0.09011   3.234  0.00146 **
---
Residual standard error: 0.9732 on 176 degrees of freedom
Multiple R-squared: 0.05608,Adjusted R-squared: 0.05072
F-statistic: 10.46 on 1 and 176 DF,  p-value: 0.001459
```

Trata-se dum péssimo ajustamento, o que não surpreende, tendo em conta a nuvem de pontos deste par de variáveis, obtida na alínea anterior. O coeficiente de determinação é quase nulo: $R^2 = 0.05608$ e menos de 6% da variabilidade no teor de flavonóides é explicado pela regressão sobre o teor alcoólico.
Como sempre, a Soma de Quadrados Total é o numerador da variância amostral dos valores observados da variável resposta. Ora,

```
> var(vinho.RLM$V8)
[1] 0.9977187
> dim(vinho.RLM)
[1] 178  13
> 177*var(vinho.RLM$V8)
[1] 176.5962
> 177*var(fitted(lm(V8 ~ V2 , data=vinho.RLM)))
[1] 9.903747
> 177*var(residuals(lm(V8 ~ V2 , data=vinho.RLM)))
```

```
[1] 166.6925
```

logo $SQT = (n–1)\, s_y^2 = 176.5962$; $SQR = (n–1)\, s_{\hat{y}}^2 = 9.903747$; e $SQRE = (n–1)\, s_e^2 = 166.6925$.
NOTA: Há outras maneiras possíveis de determinar estas Somas de Quadrados. Por exemplo, $SQR = R^2 \times SQT = 0.05608 \times 176.5962 = 9.903515$ (com um pequeno erro de arredondamento) e $SQRE = SQT - SQR = 176.5962 - 9.903515 = 166.6927$.

(c) A matriz de correlações (arredondada a duas casas decimais) entre cada par de variáveis é:

```
> round(cor(vinho.RLM), d=2)
V2    V3    V4    V5    V6    V7    V8    V9   V10   V11   V12   V13   V14
V2   1.00  0.09  0.21 -0.31  0.27  0.29  0.24 -0.16  0.14  0.55 -0.07  0.07  0.64
V3   0.09  1.00  0.16  0.29 -0.05 -0.34 -0.41  0.29 -0.22  0.25 -0.56 -0.37 -0.19
V4   0.21  0.16  1.00  0.44  0.29  0.13  0.12  0.19  0.01  0.26 -0.07  0.00  0.22
V5  -0.31  0.29  0.44  1.00 -0.08 -0.32 -0.35  0.36 -0.20  0.02 -0.27 -0.28 -0.44
V6   0.27 -0.05  0.29 -0.08  1.00  0.21  0.20 -0.26  0.24  0.20  0.06  0.07  0.39
V7   0.29 -0.34  0.13 -0.32  0.21  1.00  0.86 -0.45  0.61 -0.06  0.43  0.70  0.50
V8   0.24 -0.41  0.12 -0.35  0.20  0.86  1.00 -0.54  0.65 -0.17  0.54  0.79  0.49
V9  -0.16  0.29  0.19  0.36 -0.26 -0.45 -0.54  1.00 -0.37  0.14 -0.26 -0.50 -0.31
V10  0.14 -0.22  0.01 -0.20  0.24  0.61  0.65 -0.37  1.00 -0.03  0.30  0.52  0.33
V11  0.55  0.25  0.26  0.02  0.20 -0.06 -0.17  0.14 -0.03  1.00 -0.52 -0.43  0.32
V12 -0.07 -0.56 -0.07 -0.27  0.06  0.43  0.54 -0.26  0.30 -0.52  1.00  0.57  0.24
V13  0.07 -0.37  0.00 -0.28  0.07  0.70  0.79 -0.50  0.52 -0.43  0.57  1.00  0.31
V14  0.64 -0.19  0.22 -0.44  0.39  0.50  0.49 -0.31  0.33  0.32  0.24  0.31  1.00
```

Analisando a coluna (ou linha) relativa à variável resposta `V8`, observa-se que a variável com a qual esta se encontra mais correlacionada (em módulo) é `V7` ($r_{7,8} = 0.86$), o que confirma a inspecção visual feita na alínea 7a. Assim, o coeficiente de determinação numa regressão de `V8` sobre `V7` é $R^2 = 0.8645635^2 = 0.74747$, ou seja, o conhecimento do índice de fenóis totais permite, através da regressão ajustada, explicar cerca de 75% da variabilidade total do teor de flavonóides. O valor de $SQT = 176.5962$ é igual ao obtido na alínea anterior, uma vez que diz apenas respeito à variabilidade da variável resposta (não dependendo do modelo de regressão ajustado). Já o valor de SQR vem alterado e é agora: $SQR = R^2 \cdot SQT = 132.0004$, sendo $SQRE = SQT - SQR = 176.5962 - 132.0004 = 44.5958$.

(d) O modelo pedido no enunciado é:

```
> lm(V8 ~ V4 + V5 + V11 + V12 + V13 , data=vinho.RLM)
Coefficients:
(Intercept)           V4           V5          V11          V12          V13
   -2.25196      0.53661     -0.04932      0.09053      0.95720      0.99496

> summary(lm(V8 ~ V4 + V5 + V11 + V12 + V13 , data=vinho.RLM))$r.sq
[1] 0.7144
```

Os cinco preditores referidos permitem obter um coeficiente de determinação quase tão bom, embora ainda inferior, ao obtido utilizando apenas o preditor `V7`.

(e) Ajustando a mesma variável resposta `V8` sobre a totalidade das restantes variáveis obtem-se:

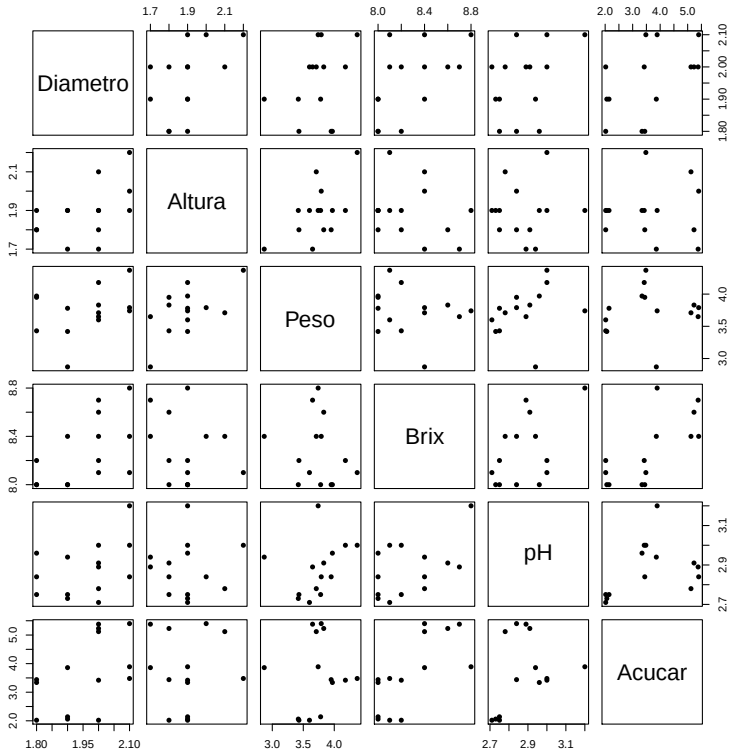
```
> lm(V8 ~ . , data=vinho.RLM)

Coefficients:
(Intercept)           V2           V3           V4           V5           V6           V7
 -1.333e+00    4.835e-03   -4.215e-02    4.931e-01   -2.325e-02   -3.559e-03    7.058e-01
         V9          V10          V11          V12          V13          V14
-1.000e+00    2.840e-01    1.068e-04    4.387e-01    3.208e-01    9.557e-05

> 177*var(fitted(lm(V8 ~ . , data=vinho.RLM)))
```

```
[1] 151.4735
> 177*var(residuals(lm(V8 ~ . , data=vinho.RLM)))
[1] 25.12269
```

i. De novo, o valor da Soma de Quadrados Total já é conhecido das alíneas anteriores: não depende do modelo ajustado, mas apenas da variância dos valores observados de Y (`V8`, neste exercício), que não se alteraram. Logo, $SQT = 176.5962$. Como se pode deduzir da listagem acima, $SQR = (n–1)\cdot s^2_{\hat{y}} = 151.4666$ e $SQRE = (n–1)\cdot s^2_e = 25.12269$. Tem-se agora $R^2 = \frac{151.4735}{176.5962} = 0.8577$. Refira-se que este valor do coeficiente de determinação *nunca poderia ser inferior ao obtido nas alíneas anteriores*, uma vez que os preditores das alíneas anteriores formam um subconjunto dos preditores utilizados aqui. Repare como a diferentes modelos para a variável resposta `V8`, correspondem diferentes formas de decompôr a Soma de Quadrados Total comum, $SQT = 176.5962$. Quanto maior a parcela explicada pelo modelo (SQR), menor a parcela associada aos resíduos ($SQRE$), isto é, menor a parcela do que não é explicado pelo modelo.

ii. Os coeficientes associados a uma mesma variável são diferentes nos diversos modelos ajustados. Assim, *não é possível prever, a partir da equação ajustada num modelo com todos os preditores, qual será a equação ajustada num modelo com menos preditores.*

8. (a) A nuvem de pontos e a matriz de correlações pedidas são:




```
> round(cor(brix),d=3)
         Diametro Altura   Peso   Brix    pH Acucar
Diametro    1.000  0.488  0.302  0.557 0.411  0.492
Altura      0.488  1.000  0.587 -0.247 0.048  0.023
Peso        0.302  0.587  1.000 -0.198 0.308  0.118
```

```
Brix        0.557 -0.247 -0.198  1.000 0.509  0.714
pH          0.411  0.048  0.308  0.509 1.000  0.353
Acucar      0.492  0.023  0.118  0.714 0.353  1.000
```

Das nuvens de pontos conclui-se que não há relações lineares particularmente evidentes, facto que é confirmado pela matriz de correlações, onde a maior correlação é 0.714. Outro aspecto evidente nos gráficos é o de haver relativamente poucas observações.

(b) A equação de base (usando os nomes das variáveis como constam da *data frame*) é:

$$Brix_i \;=\; \beta_0 + \beta_1\, Diametro_i + \beta_2\, Altura_i + \beta_3\, Peso_i + \beta_4\, pH_i + \beta_5\, Acucar_i + \epsilon_i \;,$$

havendo nesta equação seis parâmetros (os cinco coeficientes das variáveis preditoras e ainda a constante aditiva β_0).

(c) Recorrendo ao comando `lm` do `R`, tem-se:

```
> brix.lm <- lm(Brix ~ . , data=brix)
> brix.lm
Call:
lm(formula = Brix ~ Diametro + Altura + Peso + pH + Acucar, data = brix)
Coefficients:
(Intercept)     Diametro       Altura         Peso           pH       Acucar
    6.08878      1.27093     -0.70967     -0.20453      0.51557      0.08971
```

(d) A interpretação dum parâmetro β_j ($j > 0$) obtém-se considerando o valor esperado de Y dado um conjunto de valores dos preditores,

$$\mu = E[Y \,|\, x_1, x_2, x_3, x_4, x_5] \;=\; \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \beta_3\, x_3 + \beta_4\, x_4 + \beta_5\, x_5$$

e o valor esperado obtido aumentando numa unidade apenas o preditor x_j, por exemplo x_3:

$$\mu_* = E[Y \,|\, x_1, x_2, x_3 + 1, x_4, x_5] \;=\; \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \beta_3\, (x_3 + 1) + \beta_4\, x_4 + \beta_5\, x_5 \;.$$

Subtraindo os valores esperados de Y, resulta apenas: $\mu_* - \mu = \beta_3$. Assim, é legítimo falar em β_3 como a *variação no valor esperado de Y, associado a aumentar X_3 em uma unidade (não variando os valores dos restantes preditores)*. No nosso contexto, a estimativa de β_3 é $b_3 = -0.20453$. Corresponde à estimativa da variação esperada no teor brix (variável resposta), associada a aumentar em uma unidade a variável preditora peso, mantendo constantes os valores dos restantes preditores. Ou seja, corresponde a dizer que um aumento de $1g$ no peso dum fruto (mantendo iguais os valores dos restantes preditores) está associado a uma diminuição média do teor brix do fruto de 0.20453 graus. As unidades de medida de b_3 são graus brix/g. Em geral, as unidades de medida de β_j são as unidades da variável resposta Y a dividir pelas unidades do preditor X_j associado a β_j.

(e) A interpretação de β_0 é diferente da dos restantes parâmetros, mas igual ao duma ordenada na origem num regressão linear simples: é o *valor esperado de Y associado a todos os preditores terem valor nulo*. No nosso contexto, o valor estimado $b_0 = 6.08878$ não tem grande interesse prático ("frutos" sem peso, nem diâmetro ou altura, com valor pH fora a escala, etc...).

(f) Num contexto descritivo, a discussão da qualidade deste ajustamento faz-se com base no coeficiente de determinação $R^2 = \frac{SQR}{SQT}$. Pode calcular-se a Soma de Quadrados Total como o numerador da variância dos valores observados y_i de teor brix: $SQT = (n-1)\, s_y^2 = 13 \times 0.07565934 = 0.9835714$. A Soma de Quadrados da Regressão é calculada de forma análoga

à anterior, mas com base na variância dos valores ajustados $\hat{y}_i$, obtidos a partir da regressão ajustada: $SQR = (n-1)\, s^2_{\hat{y}} = 13 \times 0.06417822 = 0.8343169$. Logo, $R^2 = \frac{0.8343169}{0.9835714} = 0.848$. Os valores usados aqui são obtidos no `R` com os comandos:

```
> var(brix$Brix)
[1] 0.07565934
> var(fitted(brix.lm))
[1] 0.06417822
```

Assim, esta regressão linear múltipla explica quase 85% da variabilidade do teor *brix*, bastante acima de qualquer das regressões lineares simples, para as quais o maior valor de coeficiente de determinação seria de apenas $R^2 = 0.714^2 = 0.510$ (o maior quadrado de coeficiente de correlação entre `Brix` e qualquer dos preditores).

(g) Tem-se:

```
> X <- model.matrix(brix.lm)
> X
    (Intercept) Diametro Altura Peso   pH Acucar
1             1      2.0    2.1  3.71 2.78   5.12
2             1      2.1    2.0  3.79 2.84   5.40
3             1      2.0    1.7  3.65 2.89   5.38
4             1      2.0    1.8  3.83 2.91   5.23
5             1      1.8    1.8  3.95 2.84   3.44
6             1      2.0    1.9  4.18 3.00   3.42
7             1      2.1    2.2  4.37 3.00   3.48
8             1      1.8    1.9  3.97 2.96   3.34
9             1      1.8    1.8  3.43 2.75   2.02
10            1      1.9    1.9  3.78 2.75   2.14
11            1      1.9    1.9  3.42 2.73   2.06
12            1      2.0    1.9  3.60 2.71   2.02
13            1      1.9    1.7  2.87 2.94   3.86
14            1      2.1    1.9  3.74 3.20   3.89
```

A matriz do modelo é a matriz de dimensões $n \times (p+1)$, cuja primeira coluna é uma coluna de n uns e cujas p colunas seguintes são as colunas dadas pelas n observações de cada uma das variáveis preditoras.

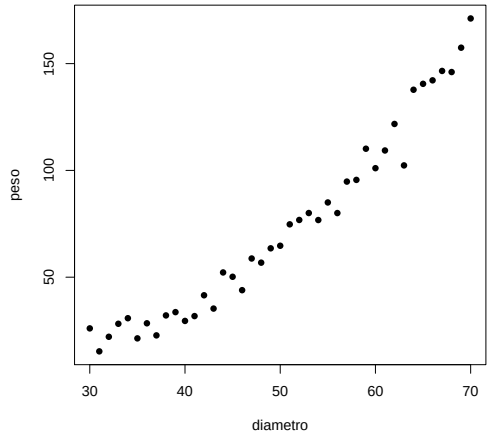
O vector $\mathbf{b}$ dos $p+1$ parâmetros ajustados é dado pelo produto matricial do enunciado: $\mathbf{b} = (\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\mathbf{y})$. Um produto matricial no `R` é indicado pelo operador "`%*%`", enquanto que uma inversa matricial é calculada pelo comando `solve`. A transposta duma matriz é dada pelo comando `t`. Logo, o vector $\mathbf{b}$ obtém-se da seguinte forma:

```
> solve(t(X) %*% X) %*% t(X) %*% brix$Brix
                [,1]
(Intercept)  6.08877506
Diametro     1.27092840
Altura      -0.70967465
Peso        -0.20452522
pH           0.51556821
Acucar       0.08971091
```

Como se pode confirmar, trata-se dos valores já obtidos através do comando `lm`.

9. (a) O gráfico pedido pode ser obtido da forma usual:

```
> plot(ameixas, pch=16)
```



Embora uma relação linear não seja uma opção disparatada, o gráfico sugere a existência de curvilinearidade na relação entre diâmetro e peso.

(b) É pedida uma *regressão polinomial* entre diâmetro e peso (mais concretamente uma relação quadrática), que pode ser ajustada como um caso especial de regressão múltipla, apesar de haver um único preditor (`diametro`). De facto, e como foi visto nas aulas, a equação polinomial de segundo grau $Y = \beta_0 + \beta_1 X + \beta_2 X^2$ pode ser vista como uma relação linear de fundo entre a variável resposta Y e dois preditores: $X_1 = X$ e $X_2 = X^2$. Para ajustar este modelo, procedemos da seguinte forma:

```
> ameixas2.lm <- lm(peso ~ diametro + I(diametro^2), data=ameixas)
> summary(ameixas2.lm)
(...)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   63.763698  18.286767   3.487  0.00125 **
diametro      -3.604849   0.759323  -4.747 2.91e-05 ***
I(diametro^2)  0.072196   0.007551   9.561 1.17e-11 ***
---
Residual standard error: 6.049 on 38 degrees of freedom
Multiple R-squared: 0.9826,Adjusted R-squared: 0.9816
F-statistic:  1071 on 2 and 38 DF,  p-value: < 2.2e-16
```

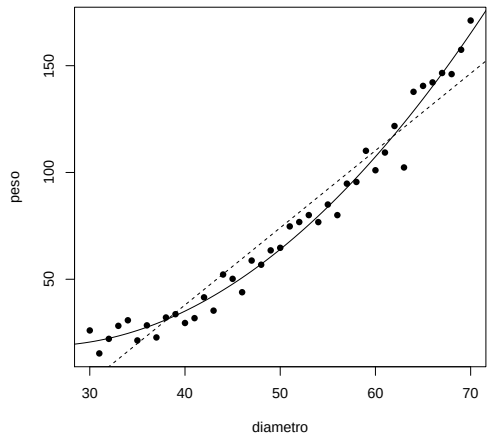
O ajustamento global deste modelo é muito bom. É possível interpretar o valor $R^2 = 0.9826$ da mesma forma que para qualquer outro modelo de regressão linear múltipla: este modelo explica cerca de $98,26\%$ da variabilidade dos pesos das ameixas.

Os parâmetros do modelo (β_0, β_1 e β_2) são estimados, respectivamente, por: $b_0 = 63.763698$, $b_1 = -3.604849$ e $b_2 = 0.072196$. Logo, a parábola ajustada tem a seguinte equação:

$$peso = 63.763698 - 3.604849\, diametro + 0.072196\, diametro^2 \ .$$

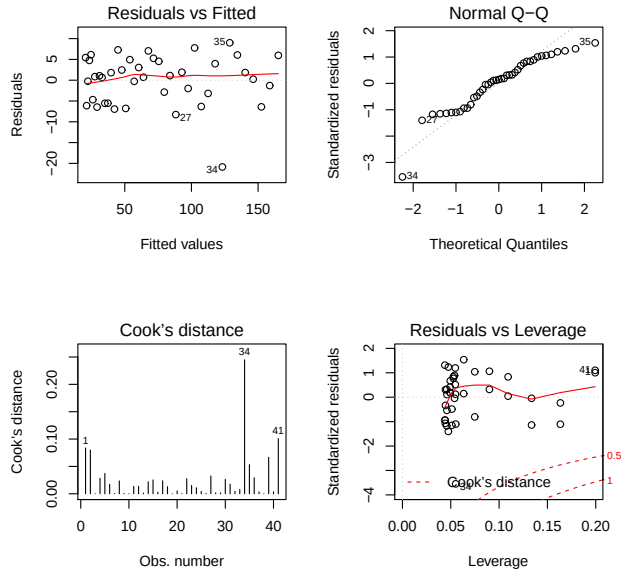
Para desenhar esta parábola em cima da nuvem de pontos criada acima, já não é possível usar o comando `abline` (que apenas serve para traçar rectas). Podemos, no entanto, usar o comando `curve`, como se ilustra seguidamente. O argumento `add=TRUE` usado nesse comando serve para que o gráfico da função cuja expressão é dada no comando, seja traçado em cima da janela gráfica já aberta (e não criando uma nova janela gráfica). Embora não seja pedido no enunciado, ajusta-se também uma recta de regressão de peso sobre diâmetro, recta igualmente indicada no gráfico a tracejado, a fim de visualizar a melhoria do ajustamento ao passar dum polinómio de grau 1 (associado à recta) para um polinómio de grau 2 (associado à parábola).

```
> curve(63.763698 - 3.604849*x + 0.072196*x^2, from=25, to=75, add=TRUE)
> abline(lm(peso ~ diametro, data=ameixas), lty="dashed")
```



(c) Vejamos os principais gráficos dos resíduos e diagnósticos:

```
> plot(ameixas2.lm, which=c(1,2,4,5))
```



Todos os gráficos parecem corresponder ao que seria de desejar, com excepção da existência duma observação (a número 34) que, sob vários aspectos é invulgar: tem um resíduo elevado (em módulo), sai fora da linearidade no *qq-plot* (que parece adequado para as restantes observações) e tem a maior distância de Cook (cerca de 0.25 e bastante maior que qualquer das restantes). Trata-se evidentemente duma observação anómala (qualquer que seja a razão), mas tratando-se duma observação isolada não é motivo para questionar o bom ajustamento geral do modelo.

(d) Para responder a esta questão, será necessário ajustar um polinómio de terceiro grau aos dados. O ajustamento correspondente é dado por:

```
> ameixas3.lm <- lm(formula = peso ~ diametro + I(diametro^2) + I(diametro^3), data = ameixas)
> summary(ameixas3.lm)
(...)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.127e+01  8.501e+01   0.838    0.407
diametro      -4.089e+00  5.405e+00  -0.757    0.454
I(diametro^2)  8.222e-02  1.110e-01   0.741    0.463
I(diametro^3) -6.682e-05  7.380e-04  -0.091    0.928

Residual standard error: 6.13 on 37 degrees of freedom
Multiple R-squared: 0.9826,Adjusted R-squared: 0.9812
F-statistic: 695.1 on 3 and 37 DF,  p-value: < 2.2e-16
```

O polinómio de terceiro grau ajustado tem equação

$$peso = 71.27 - 4.089\, diametro + 0.08222\, diametro^2 - 0.0006682\, diametro^3 \ .$$

No entanto, o acréscimo no valor do valor de R^2 não se faz sentir nas quatro casas decimais mostradas, indicando que o ganho na qualidade de ajustamento com a passagem dum modelo quadrático para um modelo cúbico é quase inexistente.

Refira-se ainda que, como para qualquer outra regressão linear múltipla, também aqui se verifica que não é possível identificar o modelo quadrático a partir do modelo cúbico: a equação da parábola obtida na alínea 9b não é igual à que se obteria ignorando a última parcela do ajustamento cúbico agora efectuado.

Admitindo já um *contexto inferencial* (isto é, admitindo os pressupostos adicionais do modelo linear), será possível efectuar um teste de hipóteses bilateral a que o coeficiente do termo cúbico seja nulo, $H_0 : \beta_3 = 0$ (em cujo caso o modelo cúbico e quadrático coincidem) vs. $H_1 : \beta_3 \neq 0$, não permite rejeitar a hipótese nula (o valor de prova é um elevadíssimo $p = 0.928$). Logo, os modelos quadrático e cúbico não diferem significativamente, preferindo-se nesse caso o mais parcimonioso modelo quadrático (a parábola). Repare-se ainda que, na tabela do ajustamento deste modelo cúbico, nenhum dos coeficientes das variáveis preditoras tem valor significativamente diferente de zero, sendo o menor dos valores de prova (*p-values*) nos testes às hipótese $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$, um elevado $p = 0.454$. No entanto, esse facto não legitima a conclusão de que se poderiam excluir, simultaneamente e sem perdas significativas na qualidade do ajustamento, *todas* as parcelas do modelo correspondentes a estes coeficientes β_j. Aliás, se assim se fizesse, deitar-se-ia fora qualquer relação entre peso e diâmetro das ameixas, quando sabemos que o modelo acima referido explica 98.26% da variabilidade dos pesos com base na relação destes com os diâmetros. Este exemplo ilustra bem que os testes t aos coeficientes β_j não devem ser usados para justificar exclusões simultâneas de mais do que um preditor.

10. Comecemos por recordar alguns resultados já previamente discutidos:

- Sabemos que, para qualquer conjunto de n pares de observações, se tem: $(n-1)\,\mathrm{cov}_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})y_i$. Distribuindo y_i e o somatório pela diferença, tem-se:

$$(n\text{–}1)\,\mathrm{cov}_{xy} = \sum_{i=1}^{n} x_i y_i - \overline{x}\underbrace{\sum_{i=1}^{n} y_i}_{=n\,\overline{y}} = \sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y} \quad \Leftrightarrow \quad \sum_{i=1}^{n} x_i y_i = (n\text{–}1)\,\mathrm{cov}_{xy} + n\,\overline{x}\,\overline{y}. \quad (1)$$

- Tomando $y_i = x_i$, para todo o i, na fórmula anterior, obtém-se:

$$(n-1)\,s_x^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\,\overline{x}^2 \quad \Leftrightarrow \quad \sum_{i=1}^{n} x_i^2 = (n-1)\,s_x^2 + n\,\overline{x}^2 \tag{2}$$

- O *produto de matrizes* AB só é possível quando o número de colunas da matriz A fôr igual ao número de linhas da matriz B (matrizes *compatíveis* para a multiplicação). Se A é de dimensão $p \times q$ e B de dimensão $q \times r$, o produto AB é de dimensão $p \times r$.
- O elemento na linha i, coluna j, dum produto matricial AB, é dado pelo *produto interno* da linha i de A com a coluna j de B: $(AB)_{ij} = (a_{i1}\ a_{i2}\ ....\ a_{iq}) \begin{pmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{qj} \end{pmatrix} = \sum_{k=1}^{q} a_{ik} b_{kj}$.
- O produto interno de dois vectores n-dimensionais $\mathbf{x}$ e $\mathbf{y}$ é dado por $\mathbf{x}^t\mathbf{y} = \sum_{i=1}^{n} x_i\,y_i$. No caso de um dos vectores ser o vector de n uns, $\mathbf{1}_n$, o produto interno resulta na soma dos elementos do outro vector, ou seja, em n vezes a média dos elementos do outro vector: $\mathbf{1}_n^t\mathbf{x} = \sum_{i=1}^{n} x_i = n\,\overline{x}$.
- A *matriz inversa* duma matriz $n \times n$ A é definida (caso exista) como a matriz (única) A^{-1}, também de dimensão $n \times n$, tal que $AA^{-1} = \mathbf{I}_n$, onde $\mathbf{I}_n$ é a matriz identidade de dimensão $n \times n$ (recorde-se que uma matriz identidade é uma matriz quadrada com todos os elementos diagonais iguais a 1 e todos os elementos não diagonais iguais a zero).
- No caso de A ser uma matriz 2×2, de elementos $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, a matriz inversa é dada (verifique!) por:

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \tag{3}$$

esta matriz inversa existe *se e só se* o *determinante* $ad - bc \neq 0$.

Com estes resultados prévios, as contas do exercício resultam de forma simples:

(a) A matriz do modelo $\mathbf{X}$ é de dimensão $n \times (p+1)$, que no caso duma regressão linear simples ($p=1$), significa $n \times 2$. Tem uma primeira coluna de uns (o vector $\mathbf{1}_n$) e uma segunda coluna com os n valores observados da variável preditora x, coluna essa que designamos pelo vector $\mathbf{x}$. Logo, a sua transposta $\mathbf{X}^t$ é de dimensão $2 \times n$. Como o vector $\mathbf{y}$ é de dimensão $n \times 1$, o produto $\mathbf{Xy}$ é possível e o resultado é um vector de dimensão 2×1. O primeiro elemento (na posição (1,1)) desse produto é dada pelo produto interno da primeira linha de $\mathbf{X}^t$ com a primeira e única coluna de $\mathbf{y}$, ou seja, por $\mathbf{1}_n^t\mathbf{y} = \sum_{i=1}^{n} y_i = n\,\overline{y}$. O segundo elemento (posição (2,1)) desse vector é dado pelo produto interno da segunda linha de $\mathbf{X}^t$ e a única coluna de $\mathbf{y}$, ou seja, por $\mathbf{x}^t\mathbf{y} = \sum_{i=1}^{n} x_i\,y_i = (n-1)\,\mathrm{cov}_{xy} + n\,\overline{x}\,\overline{y}$, tendo em conta a equação (1).

(b) Tendo em conta que $\mathbf{X}^t$ é de dimensão $2 \times n$ e $\mathbf{X}$ é de dimensão $n \times 2$, o produto $\mathbf{X}^t\mathbf{X}$ é possível e de dimensão 2×2. O elemento na posição $(1,1)$ é o produto interno da primeira linha de $\mathbf{X}^t$ ($\mathbf{1}_n$) com a primeira coluna de $\mathbf{X}$ (igualmente $\mathbf{1}_n$), logo é: $\mathbf{1}_n^t\mathbf{1}_n = n$. O elemento

na posição (1,2) é o produto interno da primeira linha de $\mathbf{X}^t$ ($\mathbf{1}_n$) e segunda coluna de $\mathbf{X}$ ($\mathbf{x}$), logo é $\mathbf{1}_n^t\mathbf{x} = \sum_{i=1}^{n} x_i = n\,\overline{x}$. O elemento na posição (2,1) é o produto interno da segunda linha de $\mathbf{X}^t$ ($\mathbf{x}$) com a primeira coluna de $\mathbf{X}$ ($\mathbf{1}_n$), logo é também $n\,\overline{x}$. Finalmente, o elemento na posição (2,2) é o produto interno da segunda linha de $\mathbf{X}^t$ ($\mathbf{x}$) com a segunda coluna de $\mathbf{X}$ ($\mathbf{x}$), ou seja, $\mathbf{x}^t\mathbf{x} = \sum_{i=1}^{n} x_i^2$. Fica assim provado o resultado do enunciado.

(c) A primeira expressão da inversa dada no enunciado vem directamente de aplicar a fórmula (3) à matriz $(\mathbf{X}^t\mathbf{X})$ obtida na alínea anterior. Apenas há que confirmar a expressão do determinante $ad–bc = n\sum_{i=1}^{n} x_i^2 – \left(\sum_{i=1}^{n} x_i\right)^2 = n\sum_{i=1}^{n} x_i^2 – (n\,\overline{x})^2 = n\left(\sum_{i=1}^{n} x_i^2 – n\overline{x}^2\right) = n(n–1)\,s_x^2$, tendo em conta a fórmula (2). Igualmente a partir da fórmula (2) obtém-se a expressão alternativa do elemento na posição (1,1), que surge na segunda expressão para $(\mathbf{X}^t\mathbf{X})^{-1}$. *Admitindo um contexto inferencial*, ao multiplicar a matriz $(\mathbf{X}^t\mathbf{X})^{-1}$ pela variância σ^2 dos erros aleatórios obtém-se a matriz

$$\sigma^2(\mathbf{X}^t\mathbf{X})^{-1} = \begin{bmatrix} \sigma^2\frac{(n–1)s_x^2+n\,\overline{x}^2}{n(n–1)s_x^2} & \frac{-\cancel{n}\overline{x}\sigma^2}{\cancel{n}(n–1)s_x^2} \\ \frac{-\cancel{n}\overline{x}\sigma^2}{\cancel{n}(n–1)s_x^2} & \frac{\cancel{n}\sigma^2}{\cancel{n}(n–1)s_x^2} \end{bmatrix} = \begin{bmatrix} \sigma^2[\frac{1}{n}+\frac{\overline{x}^2}{(n–1)s_x^2}] & \frac{-\overline{x}\sigma^2}{(n–1)s_x^2} \\ \frac{-\overline{x}\sigma^2}{(n–1)s_x^2} & \frac{\sigma^2}{(n–1)s_x^2} \end{bmatrix}$$

No canto superior esquerdo tem-se a expressão de $V[\hat{\beta}_0]$. No canto inferior direito a expressão de $V[\hat{\beta}_1]$. O elemento comum às duas posições não diagonais é $Cov[\hat{\beta}_0, \hat{\beta}_1] = Cov[\hat{\beta}_1, \hat{\beta}_0]$.

(d) Usando as expressões finais obtidas nas alíneas (c) e (a), obtém-se

$$\begin{aligned} (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} &= \frac{1}{n(n–1)s_x^2}\begin{bmatrix} (n–1)s_x^2+n\overline{x}^2 & -n\overline{x} \\ -n\overline{x} & n \end{bmatrix}\begin{bmatrix} n\overline{y} \\ (n–1)cov_{xy}+n\,\overline{x}\,\overline{y} \end{bmatrix} \\ &= \frac{1}{n(n–1)\,s_x^2}\begin{bmatrix} (n–1)\,s_x^2\,n\,\overline{y}+\cancel{n^2\overline{x}^2\overline{y}}-n\,\overline{x}\,(n–1)\,cov_{xy}-\cancel{n^2\overline{x}^2\overline{y}} \\ -\cancel{n^2\overline{x}\,\overline{y}}+n(n–1)\,cov_{xy}+\cancel{n^2\overline{x}\,\overline{y}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\cancel{n(n–1)}\,\cancel{s_x^2}\,\overline{y}}{\cancel{n(n–1)}\,\cancel{s_x^2}} - \frac{n(n–1)cov_{xy}\,\overline{x}}{n(n–1)\,s_x^2} \\ \frac{\cancel{n(n–1)}\,cov_{xy}}{\cancel{n(n–1)}\,\cancel{s_x^2}} \end{bmatrix} = \begin{bmatrix} \overline{y}-b_1\overline{x} \\ b_1 \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}. \end{aligned}$$

11. Sabemos que a matriz de projecção ortogonal referida é dada por $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, onde $\mathbf{X}$ é a matriz do modelo, ou seja, a matriz de dimensões $n \times (p+1)$ que tem na primeira coluna, n uns, e em cada uma das p restantes colunas, as n observações de cada variável preditora. Ora,

(a) A idempotência é fácil de verificar, tendo em conta que $(\mathbf{X}^t\mathbf{X})^{-1}$ é a matriz inversa de $\mathbf{X}^t\mathbf{X}$:

$$\mathbf{H}\,\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\,\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\cancel{\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}}\mathbf{X}^t = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{H}\,.$$

A simetria resulta de três propriedade conhecidas de matrizes: a transposta duma matriz transposta é a matriz original ($(\mathbf{A}^t)^t = \mathbf{A}$); a transposta dum produto de matrizes é o produto das correspondentes transpostas, pela ordem inversa ($(\mathbf{AB})^t = \mathbf{B}^t\mathbf{A}^t$); e a transposta duma matriz inversa é a inversa da transposta ($(\mathbf{A}^{-1})^t = (\mathbf{A}^t)^{-1}$). De facto, tem-se:

$$\mathbf{H}^t = [\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t]^t = \mathbf{X}[(\mathbf{X}^t\mathbf{X})^{-1}]^t\mathbf{X}^t = \mathbf{X}[(\mathbf{X}^t\mathbf{X})^t]^{-1}\mathbf{X}^t = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{H}\,.$$

(b) Como foi visto nas aulas, qualquer vector do subespaço das colunas da matriz $\mathbf{X}$, ou seja, do subespaço $\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^n$, se pode escrever como $\mathbf{Xa}$, onde $\mathbf{a} \in \mathbb{R}^{p+1}$ é o vector dos $p+1$

coeficientes na combinação linear das colunas de $\mathbf{X}$. Ora, a projecção ortogonal deste vector sobre o subespaço $\mathcal{C}(\mathbf{X})$ (que já o contém) é dada por

$$\mathbf{HXa} \;=\; \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{Xa}) \;=\; \mathbf{X}\cancel{(\mathbf{X}^t\mathbf{X})^{-1}}\cancel{(\mathbf{X}^t\mathbf{X})}\mathbf{a} \;=\; \mathbf{Xa}\,.$$

Assim, o vector $\mathbf{Xa}$ fica igual após a projecção.

(c) Por definição, o vector dos valores ajustados é dado por $\hat{\mathbf{y}} = \mathbf{Hy}$. Ora, a média desses valores ajustados, que podemos representar por $\overline{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i$, pode ser calculado tomando o produto interno do vector $\mathbf{1}_n$ de n uns com o vector $\hat{\mathbf{y}}$, uma vez que esse produto interno devolve a soma dos elementos de $\hat{\mathbf{y}}$. Assim, a média dos valores ajustados é $\overline{\hat{y}} = \frac{1}{n}\mathbf{1}_n^t\hat{\mathbf{y}} = \frac{1}{n}\mathbf{1}_n^t\mathbf{Hy} = \frac{1}{n}(\mathbf{H1}_n)^t\mathbf{y} = \frac{1}{n}\mathbf{1}_n^t\mathbf{y}$, uma vez que $\mathbf{H1}_n = \mathbf{1}_n$, já que a projecção ortogonal dum vector num subespaço onde ele já está contido deixa esse vector invariante, e o vector $\mathbf{1}_n$ pertence ao subespaço $\mathcal{C}(\mathbf{X})$ sobre o qual $\mathbf{H}$ projecta, já que é a primeira das colunas da matriz $\mathbf{X}$. Mas a expressão final obtida, $\frac{1}{n}\mathbf{1}_n^t\mathbf{y}$ é a média $\overline{y}$ dos valores observados de Y (já que $\mathbf{1}_n^t\mathbf{y}$ devolve a soma dos elementos do vector dessas observações, $\mathbf{y}$). Assim, na regressão linear múltipla, valores observados de Y e correspondentes valores ajustados partilham o mesmo valor médio.

(d) O vector dos resíduos é dado por $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{Hy}$. A soma dos resíduos resulta do produto interno do vector $\mathbf{e}$ e o vector $\mathbf{1}_n$. Assim, tem-se (tendo também em conta a discussão das alíneas anteriores) $\mathbf{1}_n^t\mathbf{e} = \mathbf{1}_n^t(\mathbf{y} - \mathbf{Hy}) = \mathbf{1}_n^t\mathbf{y} - \mathbf{1}_n^t\mathbf{Hy} = \mathbf{1}_n^t\mathbf{y} - \mathbf{1}_n^t\mathbf{y} = 0$.

Exercícios de inferência estatística na Regressão Linear

12. 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: $\hat{\sigma}^2_{\hat{\beta}_0} = QMRE \cdot (\mathbf{X}^t\mathbf{X})^{-1}_{(1,1)}$. Acrescente-se que, tratando-se duma regressão linear *simples*, é possível provar a seguinte fórmula alternativa: $\hat{\sigma}^2_{\hat{\beta}_0} = QMRE\left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)\,s_x^2}\right]$, onde $\overline{x}$ e s_x^2 indicam, respectivamente, a média e a variância amostral dos n valores de X observados. 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 $\hat{\beta}_1$: $\hat{\sigma}^2_{\hat{\beta}_1} = 9.181472 \times 10^{-5}$. Este valor pode ser obtido a partir da expressão $\hat{\sigma}^2_{\hat{\beta}_1} = QMRE \cdot (\mathbf{X}^t\mathbf{X})^{-1}_{(2,2)}$. Também neste caso, e tratando-se duma regressão linear simples, se prova a seguinte expressão alternativa: $\hat{\sigma}^2_{\hat{\beta}_1} = \frac{QMRE}{(n-1)\, s_x^2}$.

(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 12d) 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 12d), 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:

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

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:

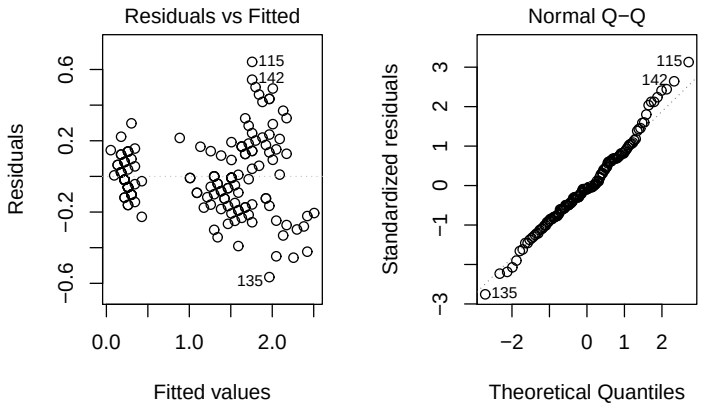
```
> predict(iris.lm, new=data.frame(Petal.Length=4.5), int="pred")
       fit      lwr      upr
1 1.507824 1.098187 1.917461
```

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))
```

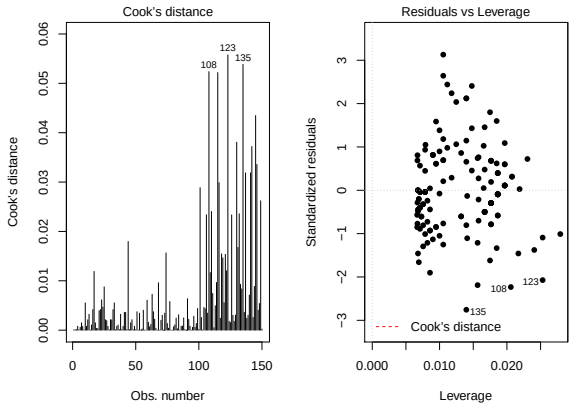
verifica-se que pode existir um problema em relação à hipótese de homogeneidade de variâncias. O gráfico da esquerda sugere que os lírios com comprimento de pétala mais pequeno (do lado esquerdo do gráfico) parecem ter menor variabilidade dos resíduos do que os restantes. Já a linearidade aproximada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



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
```

13. Seja $\mathbf{Z} = (Z_1, Z_2, ..., Z_k)^t$. Tendo em conta a definição de vector esperado e de matriz de variâncias-covariâncias, bem como as propriedades dos valores esperados, variâncias e covariâncias de variáveis aleatórias (unidimensionais) tem-se:

(a)

$$E[\alpha\mathbf{Z}] = \begin{bmatrix} E[\alpha\,Z_1] \\ E[\alpha\,Z_2] \\ \vdots \\ E[\alpha\,Z_k] \end{bmatrix} = \begin{bmatrix} \alpha\,E[Z_1] \\ \alpha\,E[Z_2] \\ \vdots \\ \alpha\,E[Z_k] \end{bmatrix} = \alpha\,E[\mathbf{Z}] \ .$$

(b)

$$E[\mathbf{Z}+\mathbf{a}] = \begin{bmatrix} E[Z_1+a_1] \\ E[Z_2+a_2] \\ \vdots \\ E[Z_k+a_k] \end{bmatrix} = \begin{bmatrix} E[Z_1]+a_1 \\ E[Z_2]+a_2 \\ \vdots \\ E[Z_k]+a_k \end{bmatrix} = \begin{bmatrix} E[Z_1] \\ E[Z_2] \\ \vdots \\ E[Z_k] \end{bmatrix} + \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix} = E[\mathbf{Z}] + \mathbf{a} \ .$$

(c)

$$\begin{aligned} V[\alpha\mathbf{Z}] &= \begin{bmatrix} V[\alpha\, Z_1] & Cov[\alpha\, Z_1, \alpha\, Z_2] & \cdots & Cov[\alpha\, Z_1, \alpha\, Z_k] \\ Cov[\alpha\, Z_2, \alpha\, Z_1] & V[\alpha\, Z_2] & \cdots & Cov[\alpha\, Z_2, \alpha\, Z_k] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[\alpha\, Z_k, \alpha\, Z_1] & Cov[\alpha\, Z_k, \alpha\, Z_2] & \cdots & V[\alpha\, Z_k] \end{bmatrix} \\ &= \begin{bmatrix} \alpha^2\, V[Z_1] & \alpha^2\, Cov[Z_1, Z_2] & \cdots & \alpha^2\, Cov[Z_1, Z_k] \\ \alpha^2\, Cov[Z_2, Z_1] & \alpha^2\, V[Z_2] & \cdots & \alpha^2\, Cov[Z_2, Z_k] \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^2\, Cov[Z_k, Z_1] & \alpha^2\, Cov[Z_k, Z_2] & \cdots & \alpha^2\, V[Z_k] \end{bmatrix} = \alpha^2\, V[\mathbf{Z}] \end{aligned}$$

(d)

$$\begin{aligned} V[\mathbf{Z}+\mathbf{a}] &= \begin{bmatrix} V[Z_1+a_1] & Cov[Z_1+a_1, Z_2+a_2] & \cdots & Cov[Z_1+a_1, Z_k+a_k] \\ Cov[Z_2+a_2, Z_1+a_1] & V[Z_2+a_2] & \cdots & Cov[Z_2+a_2, Z_k+a_k] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[Z_k+a_k, Z_1+a_1] & Cov[Z_k+a_k, Z_2+a_2] & \cdots & V[Z_k+a_k] \end{bmatrix} \\ &= \begin{bmatrix} V[Z_1] & Cov[Z_1, Z_2] & \cdots & Cov[Z_1, Z_k] \\ Cov[Z_2, Z_1] & V[Z_2] & \cdots & Cov[Z_2, Z_k] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[Z_k, Z_1] & Cov[Z_k, Z_2] & \cdots & V[Z_k] \end{bmatrix} = V[\mathbf{Z}] \end{aligned}$$

(e)

$$E[\mathbf{Z}+\mathbf{U}] = \begin{bmatrix} E[Z_1+U_1] \\ E[Z_2+U_2] \\ \vdots \\ E[Z_k+U_k] \end{bmatrix} = \begin{bmatrix} E[Z_1]+E[U_1] \\ E[Z_2]+E[U_2] \\ \vdots \\ E[Z_k]+E[U_k] \end{bmatrix} = \begin{bmatrix} E[Z_1] \\ E[Z_2] \\ \vdots \\ E[Z_k] \end{bmatrix} + \begin{bmatrix} E[U_1] \\ E[U_2] \\ \vdots \\ E[U_k] \end{bmatrix} = E[\mathbf{Z}]+E[\mathbf{U}] \ .$$

14. (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.

15. (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.

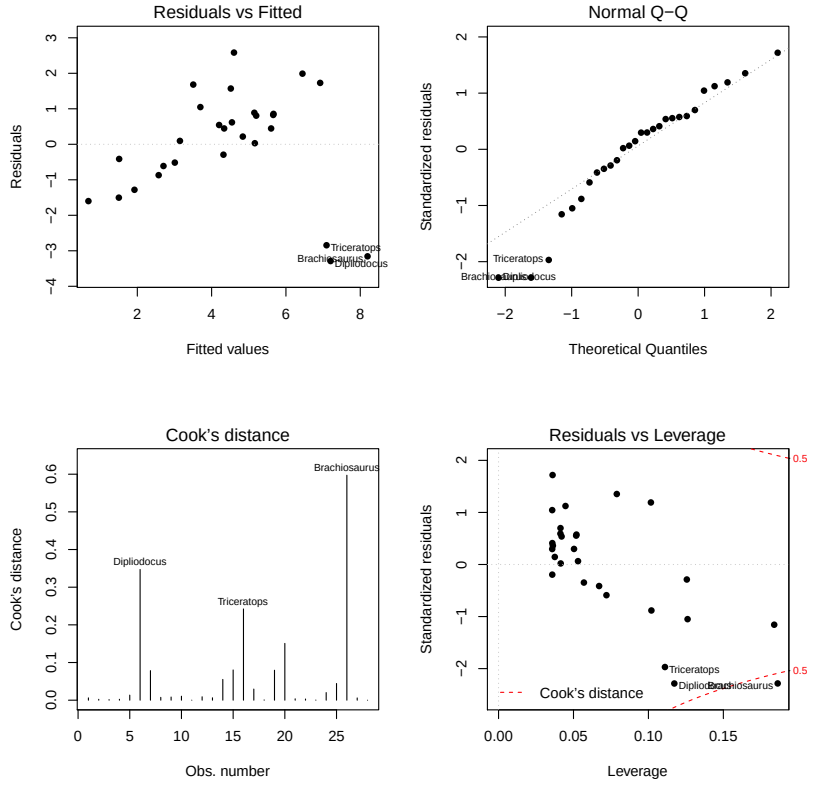
(c) Pretende-se o intervalo a 95% de confiança para β_1, ou seja:

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

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 4). 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 4.

A expressão do intervalo de confiança é a mesma que indicada na alínea 15c), 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 $\mathrm{e}^{4.755} = 116.16 < Y < 2573.443 = \mathrm{e}^{7.853}$ 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.

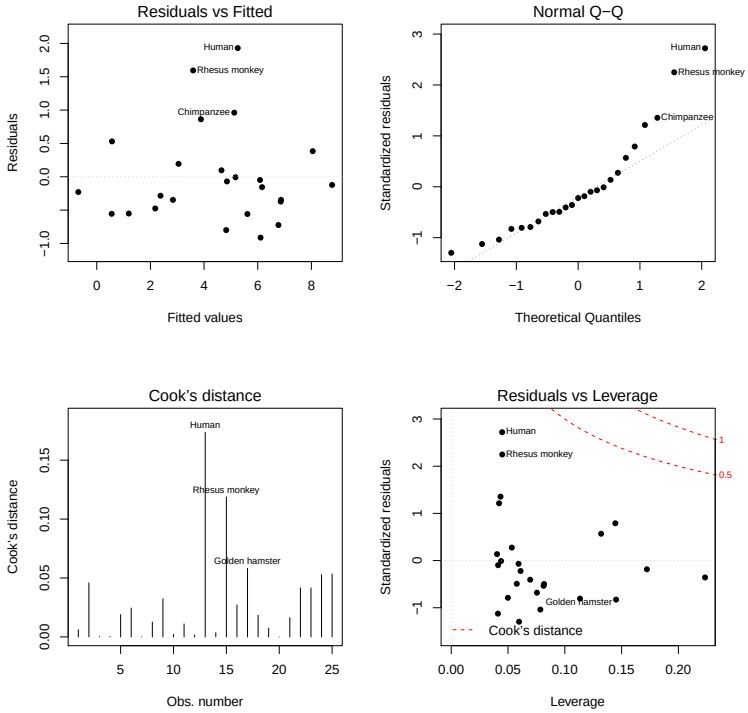
NOTA: Na alínea anterior não se pode efectuar uma transformação análoga, uma vez que valor esperado e logaritmização não são operações intercambiáveis. Ou seja, $E[\log(Y)] \neq \log(E[Y])$, pelo que não sabemos como transformar a dupla desigualdade $a < E[\log(Y)] < b$ numa dupla desigualdade equivalente 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.

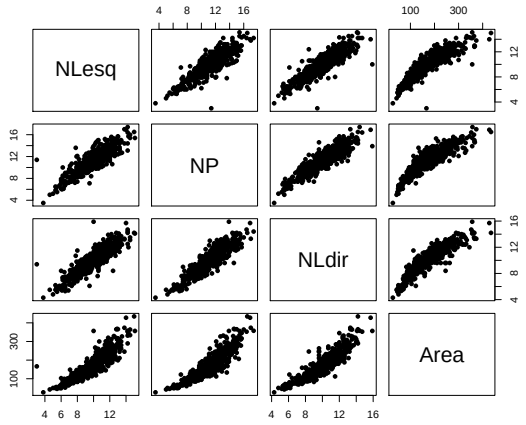


16. Na *data frame* `videiras`, a primeira coluna indica a casta, pelo que não será de utilidade neste exercício.

    (a) O comando para construir as nuvens de pontos pedidas é:

    ```
    > plot(videiras[,-1], pch=16)
    ```

    produzindo os seguintes gráficos:



Como se pode verificar, existem fortes relações lineares entre qualquer par de variáveis, o que deixa antever que uma regressão linear múltipla de área foliar sobre vários preditores venha a ter um coeficiente de determinação elevado. No entanto, nos gráficos que envolvem a variável área, existe alguma evidência de uma ligeira curvatura nas relações com cada comprimento de nervura individual.

(b) Tem-se:

```
> cor(videiras[,-1])
          NLesq        NP     NLdir      Area
NLesq 1.0000000 0.8788588 0.8870132 0.8902402
NP    0.8788588 1.0000000 0.8993985 0.8945700
NLdir 0.8870132 0.8993985 1.0000000 0.8993676
Area  0.8902402 0.8945700 0.8993676 1.0000000
```

Os valores das correlações entre pares de variáveis são todos positivos e bastante elevados, o que confirma as fortes relações lineares evidenciadas nos gráficos.

(c) Existem n observações $\{(x_{1(i)}, x_{2(i)}, x_{3(i)}, Y_i)\}_{i=1}^{n}$ nas quatro variáveis: a variável resposta área foliar (`Area`, variável aleatória Y) e as três variáveis preditoras, associadas aos comprimentos de três nervuras da folha - a principal (variável `NP`, X_1), a lateral esquerda (variável `NLesq`, X_2) e a lateral direita (variável `NLdir`, X_3). Para essas n observações admite-se que:

- A relação de fundo entre Y e os três preditores é linear, com variabilidade adicional dada por uma parcela aditiva ϵ_i chamada erro aleatório:
  $Y_i = \beta_0 + \beta_1\, x_{1(i)} + \beta_2\, x_{2(i)} + \beta_3\, x_{3(i)} + \epsilon_i$, para qualquer $i = 1, 2, ..., n$;
- os erros aleatórios têm distribuição Normal, de média zero e variância constante:
  $\epsilon_i \cap \mathcal{N}(0, \sigma^2)$, $\forall\, i$;
- Os erros aleatórios $\{\epsilon_i\}_{i=1}^{n}$ são variáveis aleatórias independentes.

(d) O comando do `R` que efectua o ajustamento pedido é o seguinte:

```
> videiras.lm <- lm(Area ~ NP + NLesq + NLdir, data=videiras)
> summary(videiras.lm)
(...)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -168.111      5.619 -29.919  < 2e-16 ***
NP             9.987      1.192   8.380  3.8e-16 ***
NLesq         11.078      1.256   8.817  < 2e-16 ***
```

```
NLdir         11.895      1.370   8.683  < 2e-16 ***
---
Residual standard error: 24.76 on 596 degrees of freedom
Multiple R-squared: 0.8649,Adjusted R-squared: 0.8642
F-statistic:  1272 on 3 and 596 DF,  p-value: < 2.2e-16
```

A equação do hiperplano ajustado é assim

$$Area \;=\; -168.111 + 9.987\,NP + 11.078\,NLesq + 11.895\,NLdir$$

O valor do coeficiente de determinação é bastante elevado: cerca de $86,49\%$ da variabilidade total nas áreas foliares é explicada por esta regressão linear sobre os comprimentos das três nervuras. Nenhum dos preditores é dispensável sem perda significativa da qualidade do modelo, uma vez que o valor de prova (*p-value*) associado aos três testes de hipóteses $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ $(j = 1, 2, 3)$ são todos muito pequenos.

O teste de ajustamento global do modelo pode ser formulado assim:

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

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

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral direita): Rej. H_0 se $F_{calc} > f_{\alpha\,(p,n-(p+1))} = f_{0.05(3,596)} \approx 2.62$.

Conclusões: O valor calculado da estatística é dado na listagem produzida pelo R ($F_{calc} = 1272$). Logo, rejeita-se (de forma muito clara) a hipótese nula, que corresponde à hipótese dum modelo inútil. Esta conclusão também resulta directamente da análise do valor de prova (*p-value*) associado à estatística de teste calculada: $p < 2.2 \times 10^{-16}$ corresponde a uma rejeição para qualquer nível de significância usual. Esta conclusão é coerente com o valor bastante elevado de R^2.

(e) São pedidos testes envolvendo a hipótese $\beta_1 = 7$ (não sendo especificada a outra hipótese, deduz-se que seja o complementar $\beta_1 \neq 7$). A hipótese $\beta_1 = 7$ é uma hipótese simples (um único valor do parâmetro β_1), que terá de ser colocada na hipótese nula e à qual corresponderá um teste bilateral.

Hipóteses: $H_0 : \beta_1 = 7 \quad$ vs. $\quad H_1 : \beta_1 \neq 7$

Estatística do Teste: $T = \frac{\hat{\beta}_1 - 7}{\hat{\sigma}_{\hat{\beta}_1}} \;\cap\; t_{(n-(p+1))}$, sob H_0.

Nível de significância: $\alpha = 0.01$.

Região Crítica: (Bilateral) Rejeitar H_0 se $|T_{\text{calc}}| > t_{0.005(596)} \approx 2.584$.

Conclusões: Tem-se $T_{\text{calc}} = \frac{b_1 - 0}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{9.987-7}{1.192} = 2.506 < 2.584$. Assim, não se rejeita a hipótese nula (que tem o benefício da dúvida), ao nível de significância de 0.01.

Se repetirmos o teste, mas agora utilizando um nível de significância $\alpha = 0.05$, apenas a fronteira da região crítica virá diferente. Agora, a regra de rejeição será: rejeitar H_0 se $|T_{\text{calc}}| > t_{0.025(596)} \approx 1.9640$. O valor da estatística de teste não se altera ($T_{calc} = 2.506$), mas este valor pertence agora à região crítica, pelo que ao nível de significância $\alpha = 0.05$ rejeitamos a hipótese formulada, optando antes por $H_1 : \beta_1 \neq 7$. Este exercício ilustra a importância de especificar sempre o nível de significância associado às conclusões do teste.

(f) É pedido um teste à igualdade de dois coeficientes do modelo, concretamente $\beta_2 = \beta_3 \;\Leftrightarrow\; \beta_2 - \beta_3 = 0$. Trata-se dum teste à diferença de dois parâmetros, que como foi visto nas aulas, é um caso particular dum teste a uma combinação linear dos parâmetros do modelo. Mais em pormenor, tem-se:

Hipóteses: $H_0 : \beta_2 - \beta_3 = 0 \quad \text{vs.} \quad H_1 : \beta_2 - \beta_3 \neq 0$

Estatística do Teste: $T = \frac{(\hat{\beta}_2 - \hat{\beta}_3) - 0}{\hat{\sigma}_{\hat{\beta}_2 - \hat{\beta}_3}} \ \cap \ t_{(n-(p+1))}$, sob H_0

Nível de significância: $\alpha = 0.05$

Região Crítica: (Bilateral) Rejeitar H_0 se $|T_{\text{calc}}| > t_{\alpha/2 \ (n-(p+1))}$

Conclusões: Conhecem-se as estimativas $b_2 = 11.078$ e $b_3 = 11.895$, mas precisamos ainda de conhecer o valor do erro padrão associado à estimação de $\beta_2 - \beta_3$ que, como foi visto nas aulas, é dado por $\hat{\sigma}_{\hat{\beta}_2 - \hat{\beta}_3} = \sqrt{\hat{V}[\hat{\beta}_2 - \hat{\beta}_3]} = \sqrt{\hat{V}[\hat{\beta}_2] + \hat{V}[\hat{\beta}_3] - 2\widehat{Cov}[\hat{\beta}_2, \hat{\beta}_3]}$. Assim, precisamos de conhecer as variâncias estimadas de $\hat{\beta}_2$ e $\hat{\beta}_3$, bem como a covariância estimada $\hat{cov}[\hat{\beta}_2, \hat{\beta}_3]$, valores estes que surgem na matriz de (co)variâncias do estimador $\hat{\boldsymbol{\beta}}$, que é estimada por $\hat{V}[\hat{\boldsymbol{\beta}}] = QMRE\,(\mathbf{X}^t\mathbf{X})^{-1}$. Esta matriz pode ser calculada no `R` da seguinte forma:

```
> vcov(videiras.lm)
            (Intercept)         NP      NLesq      NLdir
(Intercept)  31.5707574 -1.0141321 -1.0164689 -0.9051648
NP           -1.0141321  1.4200928 -0.6014279 -0.8880395
NLesq        -1.0164689 -0.6014279  1.5784886 -0.7969373
NLdir        -0.9051648 -0.8880395 -0.7969373  1.8764582
```

Assim,

$$\begin{aligned}\hat{\sigma}_{\hat{\beta}_2 - \hat{\beta}_3} &= \sqrt{\hat{V}[\hat{\beta}_2] + \hat{V}[\hat{\beta}_3] - 2\widehat{Cov}[\hat{\beta}_2, \hat{\beta}_3]} \\ &= \sqrt{1.5784886 + 1.8764582 - 2 \times (-0.7969373)} = \sqrt{5.048821} = 2.246958,\end{aligned}$$

pelo que $T_{\text{calc}} = \frac{11.078 - 11.895}{2.246958} = -0.3636027$. Como $|T_{\text{calc}}| < t_{0.025(596)} \approx 1.9640$, não se rejeita H_0 ao nível de significância de 0.05, isto é, admite-se que $\beta_2 = \beta_3$. No contexto do problema, não se rejeitou a hipótese que a variação média provocada na área foliar seja igual, quer se aumente a nervura lateral esquerda ou a nervura lateral direita em 1cm (mantendo as restantes nervuras de igual comprimento).

(g) i. Substituindo na equação do hiperplano ajustado, obtido na alínea 16d, obtém-se os seguintes valores estimados:

- *Folha 1*: $\widehat{Area} = -168.111 + 9.987 \times 12.1 + 11.078 \times 11.6 + 11.895 \times 11.9 = 222.787\, cm^2$;
- *Folha 2*: $\widehat{Area} = -168.111 + 9.987 \times 10.6 + 11.078 \times 10.1 + 11.895 \times 9.9 = 167.3995\, cm^2$;
- *Folha 3*: $\widehat{Area} = -168.111 + 9.987 \times 15.1 + 11.078 \times 14.9 + 11.895 \times 14.0 = 314.2849\, cm^2$;

Com recurso ao comando `predict` do `R`, estas três áreas ajustadas obtêm-se da seguinte forma:

```
> predict(videiras.lm, new=data.frame(NP=c(12.1,10.6,15.1), NLesq=c(11.6,10.1,14.9),
+                                                           NLdir=c(11.9, 9.9, 14.0)))
       1        2        3
222.7762 167.3903 314.2715
```

Novamente, algumas pequenas discrepâncias nas casas decimais finais resultam de erros de arredondamento.

ii. Estes intervalos de confiança para $\mu_{Y|X} = E[Y|X_1 = x_1, X_2 = x_2, X_3 = x_3]$ (com os valores de x_1, x_2 e x_3 indicados no enunciado, para cada uma das três folhas) obtêm-se subtraindo e somando aos valores ajustados obtidos na subalínea anterior a semi-amplitude do IC, dada por $t_{\alpha/2(n-(p+1))} \cdot \hat{\sigma}_{\hat{\mu}_{Y|X}}$, sendo $\hat{\sigma}_{\hat{\mu}_{Y|X}} = \sqrt{QMRE \cdot \mathbf{a}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{a}}$ onde os vectores $\mathbf{a}$ são os vectores da forma $\mathbf{a} = (1, x_1, x_2, x_3)$. Estas contas, algo

trabalhosas, resultam fáceis recorrendo de novo ao comando `predict` do `R`, mas desta vez com o argumento `int="conf"`, como indicado de seguida:

```
> predict(videiras.lm, new=data.frame(NP=c(12.1,10.6,15.1),NLesq=c(11.6,10.1,14.9),
+                                                    NLdir=c(11.9, 9.9, 14.0)), int="conf")
       fit      lwr      upr
1 222.7762 219.1776 226.3747
2 167.3903 164.9215 169.8590
3 314.2715 308.4607 320.0823
```

Assim, tem-se para cada folha, os seguintes intervalos a 95% de confiança para $\mu_{Y|X}$:

- *Folha 1*:] 219.1776 , 226.3747 [;
- *Folha 2*:] 164.9215 , 169.8590 [;
- *Folha 3*:] 308.4607 , 320.0823 [.

Repare-se como a amplitude de cada intervalo é diferente, uma vez que depende de informação específica para cada folha (dada pelo vector $\mathbf{a}$ dos valores dos preditores).

iii. Sabemos que os intervalos de predição têm uma forma análoga aos intervalos de confiança para $E[Y|X]$, mas com uma maior amplitude, associada à variabilidade adicional de observações individuais, a que corresponde $\hat{\sigma}_{indiv} = \sqrt{QMRE \cdot [1 + \mathbf{a}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{a}]}$. De novo, recorremos ao comando `predict`, desta vez com o argumento `int="pred"`:

```
> predict(videiras.lm, new=data.frame(NP=c(12.1,10.6,15.1),NLesq=c(11.6,10.1,14.9),
+                                                    NLdir=c(11.9, 9.9, 14.0)), int="pred")
       fit      lwr      upr
1 222.7762 174.0206 271.5318
2 167.3903 118.7050 216.0755
3 314.2715 265.3029 363.2401
```

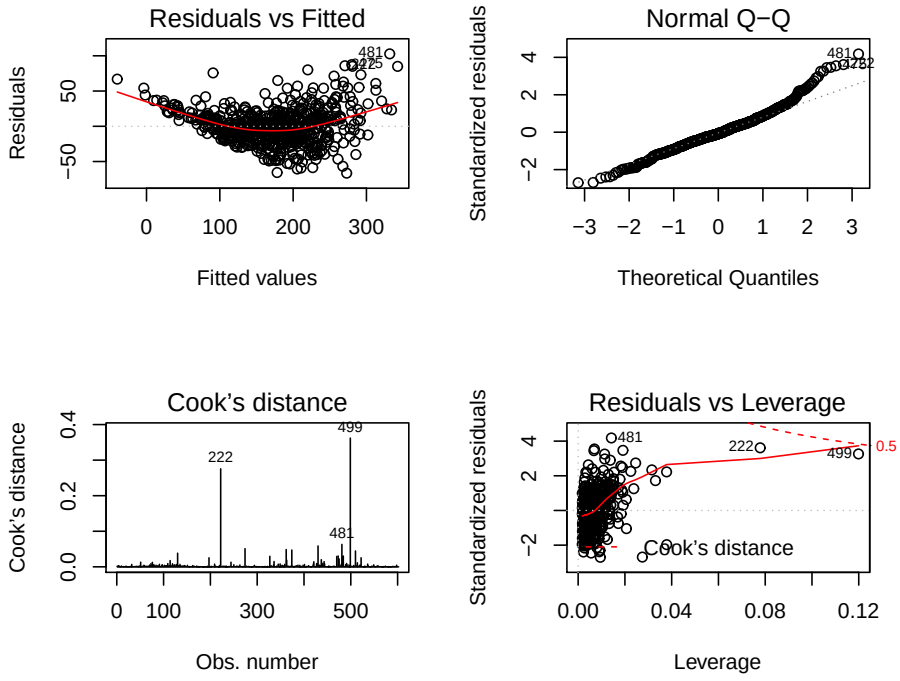
Assim, têm-se os seguintes intervalos de predição a 95% para os três valores de Y:

- *Folha 1*:] 174.0206 , 271.5318 [;
- *Folha 2*:] 118.7050 , 216.0755 [;
- *Folha 3*:] 265.3029 , 363.2401 [.

A amplitude bastante maior destes intervalos reflecte um valor elevado do Quadrado Médio Residual, que estima a variabilidade das observações individuais de Y em torno da recta.

(h) Recorremos de novo ao `R` para construir os gráficos de resíduos. O primeiro dos dois comandos seguintes destina-se a dividir a janela gráfica numa espécie de matriz 2×2:

```
> par(mfrow=c(2,2))
> plot(videiras.lm, which=c(1,2,4,5))
```



O gráfico do canto superior esquerdo é o gráfico dos resíduos usuais (e_i) vs. valores ajustados ($\hat{y}_i$). Neste gráfico são visíveis dois problemas: uma tendência para a curvatura (já detectado nos gráficos da variável resposta contra cada preditor individual), que indica que o modelo linear pode não ser a melhor forma de relacionar área foliar com os comprimentos das nervuras; e uma forma em funil que sugere que a hipótese de homogeneidade das variâncias dos erros aleatórios pode não ser a mais adequada. Este gráfico foi usado no acetato 163 das aulas teóricas. O gráfico no canto superior direito é um *qq-plot*, de quantis empíricos vs. quantis teóricos duma Normal reduzida. A ser verdade a hipótese de Normalidade dos erros aleatórios, seria de esperar uma disposição linear dos pontos neste gráfico. É visível, sobretudo na parte direita do gráfico, um afastamento relativamente forte de muitas observações a esta linearidade, sugerindo problemas com a hipótese de Normalidade. O gráfico do canto inferior esquerdo é um diagrama de barras com as distâncias de Cook de cada observação. Embora nenhuma observação ultrapasse o limiar de guarda $D_i > 0.5$, duas observações têm um valor considerável da distância de Cook: a observação 499, com D_{499} próximo de 0.4 e a observação 222, com distância de Cook próxima de 0.3. Estas duas observações merecem especial atenção para procurar identificar as causas de tão forte influência no ajustamento. Finalmente, o gráfico do canto inferior direito relaciona resíduos (internamente) estandardizados (eixo vertical) com valor do efeito alavanca (eixo horizontal) e também com as distâncias de Cook (sendo traçadas automaticamente pelo `R` linhas de igual distância de Cook, para alguns valores particularmente elevados, como 0.5 ou 1). Este gráfico ilustra que as duas observações com maior distância de Cook (499 e 222) têm valores relativamente elevados, quer dos resíduos estandardizados, quer do efeito alavanca. Saliente-se que o efeito alavanca médio, neste ajustamento de $n = 600$ observações a um modelo com $p + 1 = 4$ parâmetros é $\overline{h} = \frac{4}{600} = 0.006667$ e as duas observações referidas têm os maiores efeitos alavanca das $n = 600$ observações com valores, respectivamente,

próximos de 0.12 e 0.08. Já a observação 481, igualmente identificada no gráfico, tem o maior resíduo estandardizado de qualquer observação, mas ao ter um valor relativamente discreto do efeito alavanca, acaba por não ser uma observação influente (como se pode confirmar no gráfico anterior). Este exemplo confirma que os conceitos de observação de resíduo elevado, observação influente e observação de elevado valor do efeito alavanca (*leverage*), são conceitos diferentes. Uma observação mais atenta dos valores observados nas folhas 222 e 499 revela que o seu traço mais saliente é o desequilíbrio nos comprimentos das nervuras laterais, sendo em ambos os casos a nervura lateral direita muito mais comprida do que a esquerda. Além disso, ambas as folhas têm uma das nervuras laterais de comprimento extremo: no caso da folha 222 tem-se a maior nervura lateral direita de qualquer das 600 folhas, enquanto que a folha 499 tem a mais pequena de todas as nervuras laterais esquerdas. Assim, trata-se de folhas com formas irregulares, diferentes da generalidade das folhas analisadas.

Este exercício visa chamar a atenção que *um modelo de regressão com um ajustamento bastante forte pode revelar, no estudo dos resíduos, problemas* que levantam dúvidas sobre a validade das conclusões inferenciais (testes de hipóteses, intervalos de confiança e predição) obtidas nas alíneas anteriores.

(i) O modelo proposto corresponde à equação $Area = NP * \left(\frac{NLesq+NLdir}{2}\right)$.

i. Esta equação *não* é linearizável nas três variáveis preditoras. Apenas pode ser linearizada se se considerar que há *duas* variáveis preditoras: o comprimento da nervura principal NP e a média (ou a soma) das nervuras laterais.

ii. Considerando a soma das nervuras laterais como uma única variável, e logaritmizando a relação referida no ponto anterior, obtêm-se $\ln(Area) = \ln(NP) + \ln(NLesq + NLdir) - \ln(2)$, que é uma relação de tipo linear entre a variável resposta $y^* = \ln(Area)$ e os preditores $x_1^* = \ln(NP)$ e $x_2^* = \ln(NLesq + NLdir)$, com $\beta_1 = \beta_2 = 1$ e $\beta_0 = -\ln(2)$. Ora, ajustando um modelo linear com essas três variáveis, obtém-se:

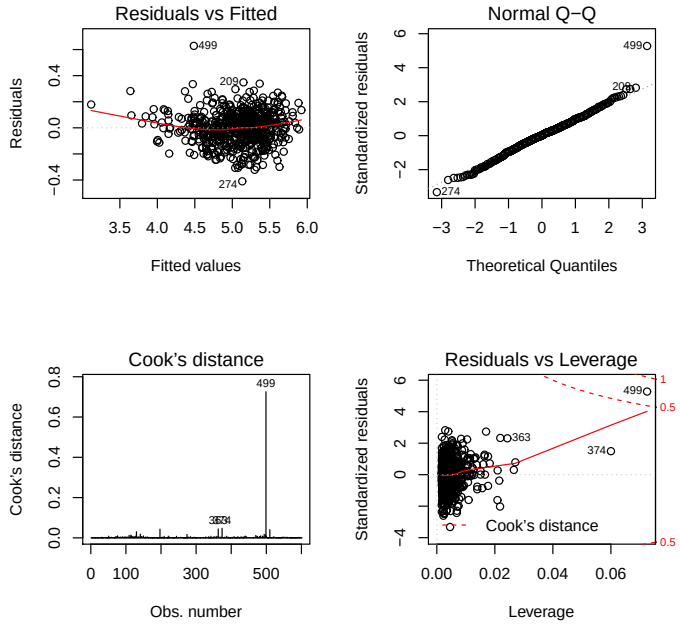
```
> summary(lm(log(Area) ~ log(NP) + I(log((NLesq+NLdir))), data=videiras))
(...)
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.53096    0.08138  -6.524 1.46e-10 ***
log(NP)                  0.67738    0.06479  10.455  < 2e-16 ***
I(log((NLesq + NLdir)))  1.33818    0.06680  20.032  < 2e-16 ***
---
(...)
Residual standard error: 0.1236 on 597 degrees of freedom
Multiple R-squared: 0.9112,Adjusted R-squared: 0.911
F-statistic:  3065 on 2 and 597 DF,  p-value: < 2.2e-16
```

Os intervalos a 95% de confiança para os três parâmetros β_j ($j = 0, 1, 2$) são:

```
> confint(lm(log(Area) ~ log(NP) + I(log((NLesq+NLdir))), data=videiras))
                             2.5 %     97.5 %
(Intercept)             -0.6907827 -0.3711297
log(NP)                  0.5501401  0.8046179
I(log((NLesq + NLdir)))  1.2069883  1.4693791
```

Assim, os valores $\beta_1 = 1$ e $\beta_2 = 1$ *não* são admissíveis (tal como não o é, embora por pouco, o valor $\beta_0 = -\ln(2) = -0.6931472$). Assim, o modelo proposto deve ser rejeitado.

iii. Independentemente do resultado insatisfatório obtido no ponto anterior, considerem-se os gráficos de residuos usuais:



Quando comparados com os modelo linear ajustado precedentemente, verifica-se uma maior correspondência destes gráficos com o que seria de exigir para validar os pressupostos do modelo: curvatura menor e redução clara do "efeito funil" no gráfico de resíduos vs. valores ajustados de y e linearidade mais clara no qq-plot, indiciando a validade das hipóteses de homogeneidade de variâncias e de Normalidade dos erros aleatórios. Assim, a logaritmização teve um efeito benéfico do ponto de vista dos pressupostos do modelo linear. Surge uma observação discordante (a observação no. 499), que tem uma distância de Cook elevada (acima de 0.7) e também um resíduo elevado e o maior de todos os valores do efeito alavanca. Trata-se claramente duma observação a merecer análise mais pormenorizada.

17. (a) Eis a regressão linear múltipla de rendimento sobre todos os preditores:

```
> summary(lm(y ~ . , data=milho))
[...]
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 51.03036   85.73770   0.595 0.557527
x1           0.87691    0.18746   4.678 0.000104 ***
x2           0.78678    0.43036   1.828 0.080522 .
x3          -0.46017    0.42906  -1.073 0.294617
x4          -0.77605    1.05512  -0.736 0.469464
x5           0.48279    0.57352   0.842 0.408563
x6           2.56395    1.38032   1.858 0.076089 .
x7           0.05967    0.71881   0.083 0.934556
x8           0.40590    1.03322   0.393 0.698045
x9          -0.65951    0.67034  -0.984 0.335426
---
Residual standard error: 7.815 on 23 degrees of freedom
Multiple R-squared:  0.7476,Adjusted R-squared:  0.6488
F-statistic: 7.569 on 9 and 23 DF,  p-value: 4.349e-05
```

Não sendo um ajustamento que se possa considerar excelente, apesar de tudo as variáveis

preditivas conseguem explicar quase 75% da variabilidade nos rendimentos. Um teste de ajustamento global rejeita a hipótese nula (inutilidade do modelo) com um valor de prova de $p = 0.00004349$.

(b) O coeficiente de determinação modificado tem valor dado no final da penúltima linha da listagem produzida pelo `R`: $R^2_{mod} = 0.6488$. Este coeficiente modificado é definido como $R^2_{mod} = 1 - \frac{QMRE}{QMT} = 1 - \frac{SQRE}{SQT} \cdot \frac{n-1}{n-(p+1)} = 1 - (1-R^2) \cdot \frac{n-1}{n-(p+1)}$. O facto de, neste exercício o valor do R^2 usual e do R^2 modificado serem bastante diferentes resulta do facto de se tratar dum modelo com um valor de R^2 (usual) não muito elevado, e que é ajustado com um número de observações ($n = 33$) não muito grande, quando comparado com o número de parâmetros do modelo ($p+1 = 10$). Efectivamente, e considerando a última das expressões acima para R^2_{mod}, vemos que o factor multiplicativo $\frac{n-1}{n-(p+1)} = \frac{32}{23} = 1.3913$. Assim, a distância do R^2 usual em relação ao seu máximo ($1 - R^2 = 0.2524$) é aumentado em cerca de 40% antes de ser subtraído de novo ao valor máximo 1, pelo que $R^2_{mod} = 1 - 0.2524 \times 1.3913 = 1 - 0.3512 = 0.6488$. Em geral, o R^2_{mod} penaliza modelos ajustados com relativamente poucas observações (em relação ao número de parâmetros do modelo), em especial quando o valor de R^2 não é muito elevado. Por outras palavras, R^2_{mod} penaliza modelos com ajustamentos modestos, baseados em relativamente pouca informação, face à complexidade do modelo.

(c) Eis o resultado do ajustamento pedido, sem o preditor x_1:

```
> summary(lm(y ~ . - x1 , data=milho))
[...]
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.387333 109.724668   1.753   0.0923 .
x2            0.305508   0.571461   0.535   0.5978
x3           -0.469256   0.586748  -0.800   0.4317
x4           -1.526474   1.426129  -1.070   0.2951
x5           -0.133203   0.763345  -0.174   0.8629
x6            3.312695   1.874882   1.767   0.0900 .
x7           -1.580293   0.858146  -1.842   0.0779 .
x8            1.239484   1.391780   0.891   0.3820
x9           -0.008387   0.896726  -0.009   0.9926
---
Residual standard error: 10.69 on 24 degrees of freedom
Multiple R-squared:  0.5074,Adjusted R-squared:  0.3432
F-statistic: 3.091 on 8 and 24 DF,  p-value: 0.01524
```

O facto mais saliente resultante da exclusão do preditor x_1 é a queda acentuada no valor do coeficiente de determinação, que é agora apenas $R^2 = 0.5074$ (repare-se como o $R^2_{mod} = 0.3432$ ainda se distancia mais do R^2 usual, reflectindo também esse ajustamento mais pobre). Assim, este modelo sem a variável preditiva x_1 apenas explica cerca de metade da variabilidade nos rendimentos. Outro facto saliente é a grande perturbação nos valores ajustados dos parâmetros (quando comparados com o modelo com todos os preditores).

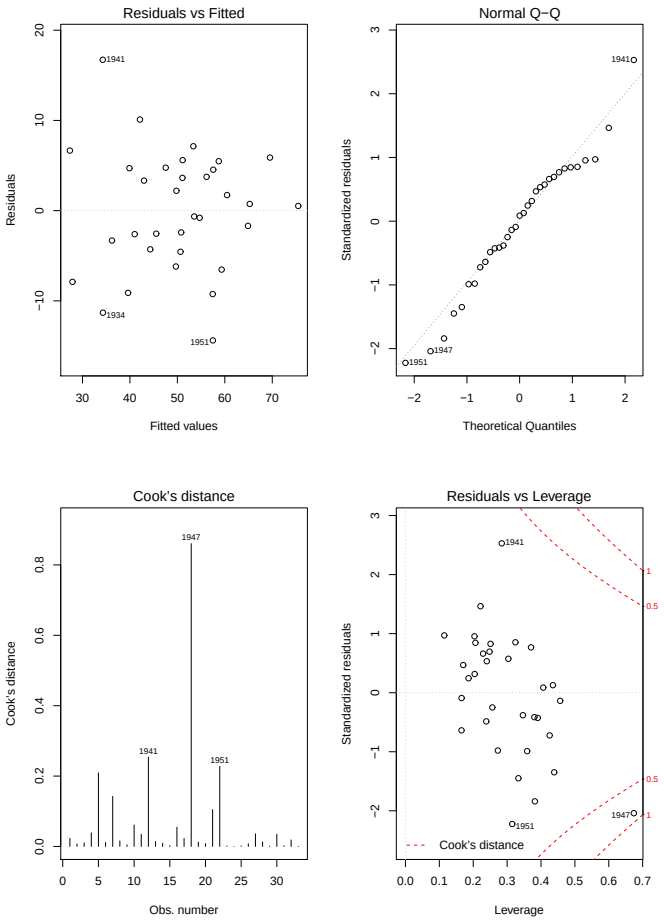
Este enorme impacto da exclusão do preditor x_1 é digno de nota, tanto mais quanto essa variável preditiva é apenas um contador dos anos que passam. Há dois aspectos a salientar:

- o preditor x_1 parece funcionar aqui como uma variável substituta (*proxy variable*, em inglês) para um grande número de outras variáveis, muitas das quais de difícil medição, tais como desenvolvimentos técnicos ou tecnológicos associados à cultura do milho nos anos em questão. A sua importância resulta de ser um indicador simples para levar em conta os aspectos não meteorológicos que, nos anos em questão, tiveram grande

impacto na produção (variável resposta do modelo), mas que não eram contemplados pelos restantes preditores.

- este exemplo ilustra bem o facto de os modelos estudarem *associações estatísticas*, o que não é sinónimo com *relações de causa e efeito*. No ajustamento do modelo com todos os preditores, a estimativa do coeficiente da variável x_1 é $b_1 = 0.87691$. Tendo em conta a natureza e unidades de medida das variáveis, podemos afirmar que, a cada ano que passa (e mantendo as restantes variáveis constantes) o valor da produção aumenta, em média, 0.87691 *bushels/acre*. Mas não faz evidentemente sentido dizer que cada ano que passa *provoca* esse aumento na produção. Não é a mera passagem do tempo que *causa* a produção. Pode exisitir uma relação de causa e efeito entre alguns preditores e a variável resposta, mas pode apenas existir uma *associação*, como neste caso. A existência, ou não, de uma relação de causa e efeito nunca poderá ser afirmada pela via estatística, mas apenas com base nos conhecimentos teóricos associados aos fenómenos sob estudo.

Quanto ao estudo dos resíduos, eis os gráficos produzidos com as opções 1, 2, 4 e 5 do comando `plot` do `R`:



O gráfico de resíduos usuais *vs.* valores ajustados $\hat{y}_i$ (no canto superior esquerdo) não apresenta qualquer padrão digno de registo, dispersando-se os resíduos numa banda horizontal. Assim, nada sugere que não se verifiquem os pressupostos de linearidade e de homgeneidade

de variâncias, admitidos no modelo RLM. Analogamente, no *qq-plot* comparando quantis teóricos duma Normal reduzida e quantis empíricos (canto superior direito), existe linearidade aproximada dos pontos, pelo que a hipótese de Normalidade dos erros aleatórios também parece admissível. Já no diagrama de barras das distâncias de Cook (canto inferior esquerdo) há um facto digno de registo: a observação correspondente ao ano 1947 tem um valor elevadíssimo da distância de Cook (superior a 0.8), pelo que se trata dum ano muito influente no ajustamento do modelo. Dado o elevado número de variáveis preditoras, não é possível visualizar a nuvem de pontos associada aos dados, mas uma análise mais atenta da tabela de valores observados (disponível no enunciado) sugere possíveis causas para este facto. O ano de 1947 teve uma precipitação pré-Junho particularmente intensa, a que se seguiu um mês de Agosto anormalmente quente e seco (nas três variáveis registam-se observações extremas, para os anos observados). O valor muito elevado da distância de Cook indica que a exclusão deste ano do conjunto de dados provocaria alterações importantes no modelo ajustado. Finalmente, o gráfico de resíduos internamente estandardizados (R_i) *vs.* valores do efeito alavanca (h_{ii}) confirmam a elevada distância de Cook da observação correspondente a 1947, e mostram que ela resulta dum resíduo internamente estandardizado relativamente grande, em valor absoluto (embora não extraordinariamente grande), mas sobretudo dum valor muito elevado (cerca de 0.7) do efeito alavanca. Este último valor sugere que esta observação está a "atrair" o hiperplano ajustado, facto que ajuda a esconder a natureza atípica desta observação. Este exemplo é ainda digno de nota por outra razão: todas as observações têm valores relativamente elevados dos efeitos alavanca. Trata-se duma consequência de se ajustar um modelo complexo (p+1 parâmetros) com relativamente poucas observações ($n = 33$). Assim, o valor médio dos efeitos alavanca, que numa RLM é dada por $\frac{p+1}{n}$, é cerca de 0.3, existindo várias observações com valores superiores do efeito alavanca.

A discussão dos resíduos para o modelo sem o preditor x_1 é muito semelhante. A distância de Cook da observação relativa a 1947 baixa um pouco, mas permanece muito elevada (cerca de 0.6), mantendo-se os restantes aspectos já referidos acima.

(d) O submodelo pedido aqui é o submodelo com os preditores de x_1 a x_5. Eis o seu ajustamento:

```
> summary(milhoJun.lm)
[...]
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.6476    50.4835   0.251   0.8041
x1            1.0381     0.1655   6.272 1.04e-06 ***
x2            0.8606     0.4198   2.050   0.0502 .
x3           -0.5710     0.4558  -1.253   0.2210
x4           -1.4878     1.0708  -1.389   0.1761
x5            0.6427     0.5747   1.118   0.2733
---
Residual standard error: 8.571 on 27 degrees of freedom
Multiple R-squared:  0.6435,Adjusted R-squared:  0.5775
F-statistic: 9.749 on 5 and 27 DF,  p-value: 2.084e-05
```

Tratando-se dum submodelo do modelo original (com todos os preditores), pode também aqui efectuar-se um teste F parcial para comparar modelo e submodelo. Temos:

Hipóteses: $H_0 : \beta_j = 0\ ,\ \forall\, j = 6,7,8,9$ [modelos equivalentes] vs. $H_1 : \exists\, j = 6,7,8,9$ tal que $\beta_j \neq 0$ [modelos diferentes]

ou alternativamente,

$$H_0 : \mathcal{R}_c^2 = \mathcal{R}_s^2 \qquad \text{vs.} \qquad H_1 : \mathcal{R}_c^2 > \mathcal{R}_s^2$$

Estatística do Teste: $F = \frac{n-(p+1)}{p-k} \cdot \frac{R_c^2 - R_s^2}{1-R_c^2} \ \cap \ F_{(p-k,n-(p+1))}$, sob H_0

Nível de significância: $\alpha = 0.05$

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{\text{calc}} > f_{\alpha(p-k,n-(p+1))}$

Conclusões: Temos $n = 33$, $p = 9$, $k = 5$, $R_c^2 = 0.7476$ e $R_s^2 = 0.6435$.

Logo, $F_{\text{calc}} = \frac{23}{4} \times \frac{0.7476-0.6435}{1-0.7476} = 2.371533 < f_{0.05(4,23)} = 2.78$. Assim, não se rejeita H_0, ou seja, o modelo e o submodelo não diferem significativamente ao nível 0.05.

Esta conclusão pode ser confirmada utilizando o comando `anova` do `R`:

```
> anova(milhoJun.lm, milho.lm)
Analysis of Variance Table
Model 1: y ~ x1 + x2 + x3 + x4 + x5
Model 2: y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
Res.Df    RSS Df Sum of Sq    F  Pr(>F)
1     27 1983.7
2     23 1404.7  4    578.98 2.37 0.08231 .
```

Apenas aceitando trabalhar com uma probabilidade de cometer o erro de Tipo I maior, por exemplo $\alpha = 0.10$, é que seria possível rejeitar H_0 e considerar os modelos como tendo ajustamentos significativamente diferentes.

Esta conclusão sugere a possibilidade de ter, já em finais de Junho, previsões de produção que expliquem quase dois terços da variabilidade observada na produção. No entanto, deve recordar-se que se trata dum modelo ajustado com relativamente poucas observações.

(e) Vamos aplicar o algoritmo de exclusão sequencial, baseado nos testes t aos coeficientes β_j e usando um nível de significância $\alpha = 0.10$.

Partindo do ajustamento do modelo com todos os preditores, efectuado na alínea 17a), conclui-se que há várias variáveis candidatas a sair (os *p-values* correspondentes aos testes a $\beta_j = 0$ são superiores ao limiar acima indicado). De entre estas, é a variável x_7 que tem de longe o maior *p-value*, pelo que é a primeira variável a excluir.

Após a exclusão do preditor x_7 é necessário re-ajustar o modelo:

```
> summary(lm(y ~ . - x7, data=milho))
[...]
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  54.8704    70.6804   0.776   0.4451
x1            0.8693     0.1602   5.425 1.42e-05 ***
x2            0.7751     0.3983   1.946   0.0634 .
x3           -0.4590     0.4199  -1.093   0.2852
x4           -0.7982     0.9995  -0.799   0.4324
x5            0.4814     0.5613   0.858   0.3996
x6            2.5245     1.2687   1.990   0.0581 .
x8            0.4137     1.0074   0.411   0.6849
x9           -0.6426     0.6252  -1.028   0.3143
---
Residual standard error: 7.652 on 24 degrees of freedom
Multiple R-squared:  0.7475,Adjusted R-squared:  0.6633
F-statistic: 8.882 on 8 and 24 DF,  p-value: 1.38e-05
```

Assinale-se que o valor do coeficiente de determinação quase não se alterou com a exclusão de x_7. Continuam a existir várias variáveis com valor de prova superiores ao limiar estabelecido, e de entre estas é a variável x_8 que tem o maior *p-value*: $p = 0.6849$. Exclui-se essa variável e ajusta-se novamente o modelo.

```
> summary(lm(y ~ . - x7 - x8, data=milho))
[...]
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  58.4750    68.9575   0.848   0.4045
x1            0.8790     0.1558   5.641 7.17e-06 ***
x2            0.8300     0.3689   2.250   0.0335 *
x3           -0.4592     0.4128  -1.112   0.2765
x4           -0.8354     0.9787  -0.854   0.4015
x5            0.5287     0.5401   0.979   0.3370
x6            2.4392     1.2306   1.982   0.0586 .
x9           -0.7254     0.5819  -1.247   0.2240
---
Residual standard error: 7.523 on 25 degrees of freedom
Multiple R-squared:  0.7457,Adjusted R-squared:  0.6745
F-statistic: 10.47 on 7 and 25 DF,  p-value: 4.333e-06
```

O valor de R^2 mantem-se próximo do original e continuam a existir variáveis candidatas a sair do modelo. De entre estas, é o preditor x_4 que tem o maior *p-value* ($p = 0.4015$), pelo que será o próximo preditor a excluir. O re-ajustamento do modelo sem os três preditores já excluídos (`x7, x8` e `x4`) produz os seguintes resultados:

```
> summary(lm(y ~ . - x7 - x8 - x4, data=milho))
[...]
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.9486    64.2899   0.590   0.5601
x1            0.8854     0.1548   5.718 5.11e-06 ***
x2            0.7685     0.3599   2.135   0.0423 *
x3           -0.3603     0.3941  -0.914   0.3690
x5            0.6338     0.5231   1.212   0.2366
x6            2.7275     1.1772   2.317   0.0286 *
x9           -0.6829     0.5767  -1.184   0.2471
---
Residual standard error: 7.484 on 26 degrees of freedom
Multiple R-squared:  0.7383,Adjusted R-squared:  0.6779
F-statistic: 12.23 on 6 and 26 DF,  p-value: 1.624e-06
```

Após a exclusão de três preditores, o coeficiente de determinação continua próximo do valor original: $R^2 = 0.7383$. Esta quebra pequena reflecte os valores elevados dos *p-values* associados aos preditores excluídos. Mas há mais preditores candidatos à exclusão, sendo x_3 a próxima variável a excluir do lote de preditores ($p = 0.3690 > 0.10$).

```
> summary(lm(y ~ . - x7 - x8 - x4 - x3, data=milho))
[...]
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.3646    64.0755   0.614   0.5441
```

```
x1            0.8870     0.1544   5.747 4.13e-06 ***
x2            0.7562     0.3586   2.109   0.0444 *
x5            0.4725     0.4910   0.962   0.3444
x6            2.4893     1.1445   2.175   0.0386 *
x9           -0.8320     0.5515  -1.509   0.1430
---
Residual standard error: 7.461 on 27 degrees of freedom
Multiple R-squared:  0.7299,Adjusted R-squared:  0.6799
F-statistic: 14.59 on 5 and 27 DF,  p-value: 5.835e-07
```

Há ainda candidatos à exclusão, sendo x_5 a exclusão seguinte.

```
> summary(lm(y ~ . - x7 - x8 - x4 - x3 - x5, data=milho))
[...]
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  87.1589    40.4371   2.155   0.0399 *
x1            0.8519     0.1498   5.688 4.25e-06 ***
x2            0.5989     0.3187   1.879   0.0707 .
x6            2.3613     1.1353   2.080   0.0468 *
x9           -0.9755     0.5302  -1.840   0.0764 .
---
Residual standard error: 7.451 on 28 degrees of freedom
Multiple R-squared:  0.7206,Adjusted R-squared:  0.6807
F-statistic: 18.06 on 4 and 28 DF,  p-value: 1.954e-07
```

Tendo em conta que fixámos o limiar de exclusão no nível de significância $\alpha = 0.10$, não há mais variáveis candidatas à exclusão, pelo que o algoritmo termina aqui. O modelo final escolhido pelo algoritmo tem quatro preditores (`x1, x2, x6` e `x9`), e um coeficiente de determinação $R^2 = 0.7206$. Ou seja, com menos de metade dos preditores iniciais, apenas se perdeu 0.027 no valor de R^2.

O valor relativamente alto ($\alpha = 0.10$) do nível de significância usado é aconselhável, na aplicação deste algoritmo, uma vez que variáveis cujo *p-value* cai abaixo deste limiar podem, se excluídas, gerar quebras mais pronunciadas no valor de R^2. Tal facto é ilustrado pela exclusão de x_9 (a exclusão seguinte, caso se tivesse optado por um limiar $\alpha = 0.05$):

```
> summary(lm(y ~ . - x7 - x8 - x4 - x3 - x5 - x9, data=milho))
[...]
Residual standard error: 7.752 on 29 degrees of freedom
Multiple R-squared:  0.6869,Adjusted R-squared:  0.6545
F-statistic:  21.2 on 3 and 29 DF,  p-value: 1.806e-07
```

Dado o número de exclusões efectuadas, pode desejar-se fazer um teste F parcial, comparando o submodelo final produzido pelo algoritmo e o modelo original com todos os preditores:

```
> anova(milhoAlgExc.lm, milho.lm)
Analysis of Variance Table

Model 1: y ~ x1 + x2 + x6 + x9
Model 2: y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     28 1554.6
2     23 1404.7  5     149.9 0.4909 0.7796
```

O *p-value* muito elevado ($p = 0.7796$) indica que não se rejeita a hipótese de modelo e submodelo serem equivalentes.

Como foi indicado nas aulas, existe uma função do `R`, a função `step`, que automatiza um algoritmo de exclusão sequencial, mas utilizando o valor do Critério de Informação de Akaike (AIC) como critério de exclusão dum preditor em cada passo do algoritmo. Esta função produz neste exemplo o mesmo submodelo final, como se pode constatar na parte final desta listagem:

```
> step(milho.lm)
Start:  AIC=143.79
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
[...]
Step:  AIC=137.13
y ~ x1 + x2 + x6 + x9
Df Sum of Sq    RSS    AIC
<none>              1554.6 137.13
- x9    1    187.95 1742.6 138.90
- x2    1    196.01 1750.6 139.05
- x6    1    240.20 1794.8 139.87
- x1    1   1796.22 3350.8 160.47
Call: lm(formula = y ~ x1 + x2 + x6 + x9, data = milho)
Coefficients:
(Intercept)           x1           x2           x6           x9
87.1589       0.8519       0.5989       2.3613      -0.9755
```

Refira-se que as variáveis meteorológicas mais associadas à previsão da produção são a precipitação pré-Junho (x_2), a precipitação em Julho (x_6) e a temperatura em Agosto (x_9).

Finalmente, refira-se que, caso esteja disponível *software* adequado, pode recorrer-se a uma pesquisa completa de todos os subconjuntos, a fim de escolher os melhores, para cada número k de preditores. Como referido nas aulas, o módulo `leaps` do `R` disponibiliza um comando de igual nome para fazer essas escolhas. Eis os comandos e a listagem produzida, para o conjunto de dados deste Exercício.

```
> library(leaps)
> leaps(y=milho$y , x=milho[,-10], method="r2", nbest=1)
$which
1     2     3     4     5     6     7     8     9
1 TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
2 TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
3 TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
4 TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE
5 TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE
6 TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE
7 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE
8 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
9 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[...]
$r2
[1] 0.5633857 0.6566246 0.6868757 0.7206491 0.7299145 0.7383258 0.7457353
[8] 0.7475100 0.7475856
```

Na matriz de valores lógicos, cada linha corresponde a uma cardinalidade (número de variáveis) do subconjunto preditor, e cada coluna corresponde a uma das variáveis preditoras.

As colunas que tenham o valor lógico `TRUE`, na linha correspondente a k preditores, correspondem a variáveis que pertencem ao melhor subconjunto de k preditores. Repare-se como o melhor subconjunto de quatro preditores é o subconjunto `x1`, `x2`, `x6` e `x9`, escolhido pelo algoritmo de exclusão sequencial (nas suas duas versões). Aliás, em todos os passos intermédios do algoritmo, o subconjunto de k preditores escolhido acaba por revelar-se o subconjunto óptimo, ou seja, o subconjunto de preditores que está associado aos maiores valores do Coeficiente de Determinação.

(f) O ajustamento pedido nesta alínea produziu os seguintes resultados:

```
> summary(lm(I(y*0.06277) ~  x1 + I(x2*25.4) + I(x6*25.4) + I(5/9*(x9-32)), data=milho))
[...]
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.5114712  1.5019053   2.338   0.0268 *
x1                  0.0534744  0.0094015   5.688 4.25e-06 ***
I(x2 * 25.4)        0.0014800  0.0007877   1.879   0.0707 .
I(x6 * 25.4)        0.0058354  0.0028055   2.080   0.0468 *
I(5/9 * (x9 - 32)) -0.1102213  0.0599066  -1.840   0.0764 .
---
Residual standard error: 0.4677 on 28 degrees of freedom
Multiple R-squared:  0.7206,Adjusted R-squared:  0.6807
F-statistic: 18.06 on 4 and 28 DF,  p-value: 1.954e-07
```

Comparando esta listagem com os resultados do modelo final produzido pelo algoritmo de exclusão sequencial, nas unidades de medida originais (ver alínea 17e), constata-se que as quantidades associadas à qualidade do ajustamento global (R^2, valor da estatística F no teste de ajustamento global) mantêm-se inalteradas. Trata-se duma consequência do facto de que as transformações de variáveis foram todas transformações lineares (afins). No entanto, e tal como sucedia na RLS, os valores das estimativas b_j são diferentes. O facto de que a informação relativa aos testes a $\beta_j = 0$ se manter igual, para os coeficientes β_j que multiplicam as variáveis preditoras (isto é, quando $j > 0$), sugere que se trata de alterações que apenas visam adaptar as estimativas às novas unidades de medida, não alterando globalmente o ajustamento.

18. (a) i. **Hipóteses:** $H_0 : \beta_1 = \beta_2 = 0$, vs. $H_1 : \beta_1 \neq 0$ ou $\beta_2 \neq 0$.
Estatística do teste: $F = \frac{n-(p+1)}{p} \frac{R^2}{1-R^2} \cap F_{(p,n-(p+1))}$, sob H_0.
Nível de significância: $\alpha = 0.05$.
Região Crítica (Unilateral direita): Rejeitar H_0 se $F_{calc} > f_{\alpha\,(p,n-(p+1))} = f_{0.05(2,28)} \approx 3.33$ (entre 3.32 e 3.39, nas tabelas).
Conclusões: O enunciado indica que o valor calculado da estatística é $F_{calc} = 255$. Assim, *rejeita-se* H_0, indicando que o modelo RLM difere significativamente do modelo nulo.

ii. Nos testes a que o coeficiente β_j de cada preditor ($j = 1, 2$) seja nulo, os valores de prova dados no enunciado indicam que ambos são inferiores a $\alpha = 0.05$, pelo que haverá rejeição de $H_0 : \beta_j = 0$ em ambos os casos e, ao nível $\alpha = 0.05$, qualquer das regressões lineares simples possíveis terá uma qualidade de ajustamento significativamente pior. Já ao nível $\alpha = 0.01$ a situação é diferente. Enquanto o *p-value* para o teste a $H_0 : \beta_1 = 0$ é $p < 2 \times 10^{-16}$, ou seja, indistinguível de zero e portanto indicando com grande convicção que $\beta_1 \neq 0$, já o valor de prova no teste a $H_0 : \beta_2 = 0$ é $p = 0.0145$ e portanto superior a $\alpha = 0.01$. Assim, e embora por pouco, não se rejeita a hipótese $H_0 : \beta_2 = 0$ ao nível

de significância $\alpha = 0.01$. Como tal, uma regressão linear simples de `Volume` sobre `Diametro` não difere significativamente (para $\alpha = 0.01$) da regressão com dois preditores ajustada no enunciado.

iii. Sabemos que numa regressão linear simples, o coeficiente de determinação é o quadrado do coeficiente de correlação entre o preditor e a variável resposta. Com base na matriz de correlações disponível no enunciado geral, temos que, na RLS de `Volume` sobre `Diametro` o coeficiente de determinação é $R^2 = 0.9671194^2 = 0.9353199$, enquanto que na RLS de `Volume` sobre `Altura` o coeficiente de determinação é $R^2 = 0.5982497^2 = 0.3579027$. Estes valores são coerentes com os resultados da alínea anterior. Quanto aos valores das estatísticas F nos testes de ajustamento global, podem ser obtidos pela fórmula da RLS, $F = (n-2)\frac{R^2}{1-R^2}$. Os valores nas duas regressões lineares simples são (e indicando o preditor pela sua inicial) $F_D = 29 \times \frac{0.9353199}{1-0.9353199} = 419.3605$ e $F_A = 29 \times \frac{0.3579027}{1-0.3579027} = 16.16449$.

(b) Consideremos agora o modelo com base nas transformações logarítmicas das três variáveis originais. Designaremos por y o log-volume, por x_1 o log-diâmetro e por x_2 a log-altura.

i. Partindo da relação linear entre as variáveis logaritmizadas, tem-se:

$$\begin{aligned} \ln(y) = b_0 + b_1 \ln x_1 + b_2 \ln x_2 \quad &\Leftrightarrow \quad y = \mathrm{e}^{b_0 + b_1 \ln x_1 + b_2 \ln x_2} \\ &\Leftrightarrow \quad y = \mathrm{e}^{b_0}\, \mathrm{e}^{b_1 \ln x_1}\, \mathrm{e}^{b_2 \ln x_2} \\ &\Leftrightarrow \quad y = \underbrace{\mathrm{e}^{b_0}}_{=b_0^*}\, \mathrm{e}^{\ln x_1^{b_1}}\, \mathrm{e}^{\ln x_2^{b_2}} \\ &\Leftrightarrow \quad y = b_0^*\, x_1^{b_1}\, x_2^{b_2} \ . \end{aligned}$$

Assim, y é proporcional ao produto de potências de cada um dos preditores. A superfície em R^3 ajustada à nuvem de pontos das observações originais terá, tendo em conta os valores disponíveis no enunciado, equação $y = \mathrm{e}^{-6.63162}\, x_1^{1.98265}\, x_2^{1.11712}$, ou seja, $Volume = 0.001318\, Diametro^{1.98265}\, Altura^{1.11712}$.

ii. Esta frase baseia-se numa comparação errada, uma vez que as escalas da variável resposta y (usadas para medir, resíduos e todas as Somas de Quadrados numa regressão, logo também usadas para obter os coeficientes de determinação e portanto também o valor da estatística F) são diferentes nos dois modelos ajustados. Enquanto que na alínea anterior o volume era medido na escala original, nesta alínea a regressão linear usa a escala logarítmica para os volumes. Assim, o R^2 da alínea anterior media a proporção da variabilidade *dos volumes* observados que era explicada pela regressão então usada, nesta alínea o R^2 mede a variabilidade *dos log-volumes* observados que é explicada pela nova regressão. Os SQTs de cada alínea não são iguais. Não são correctas as comparações referidas na frase do enunciado.

iii. Com base na relação entre as variáveis originais estabelecida duas alíneas acima, podemos verificar que na regra simples $v = \pi r^2 h$ corresponde a ter-se uma relação do tipo $Volume = \beta_0^* \left(\frac{Diametro}{2}\right)^{\beta_1} Altura^{\beta_2}$, com $\beta_2 = 1$, $\beta_1 = 2$ e (tendo também em conta as unidades de medida do diâmetro - polegadas - que eram diferentes das restantes) $\beta_0^* = \exp(\beta_0) = \pi \times \left(\frac{1}{12} \times \frac{1}{2}\right)^2$, logo $\beta_0 = \ln\left(\frac{\pi}{24^2}\right) = -5.211378$. A matéria estudada sugere que se façam testes de hipóteses para cada dos parâmetros, com as hipóteses da forma $H_0 : \beta_j = c_j$, a fim de saber se os valores c_j (acima referidos) são admissíveis. Nos três casos, a estatística do teste terá valor $T_{calc} = \frac{b_j - c_j}{\hat{\sigma}_{\hat{\beta}_j}}$. Uma vez que

$t_{0.025(28)} = 2.048407$, as regras de rejeição, nos três testes, serão: rejeitar $H_0 : \beta_j = c_j$ se $|T_{calc}| > 2.048407$. Com base nos valores de b_j e $\hat{\sigma}_{\hat{\beta}_j}$ dados na listagem dos resultados, tem-se, para o teste a $H_0 : \beta_2 = c_2 = 1$, $T_{calc} = \frac{1.11712-1}{0.20444} = 0.572882$, pelo que não se rejeita H_0. De forma análoga, no teste a $H_0 : \beta_1 = c_1 = 2$, tem-se $T_{calc} = \frac{1.98265-2}{0.07501} = -0.2313025$, pelo que também não se rejeita H_0. Finalmente, no teste a $H_0 : \beta_0 = c_0 = -5.211378$, tem-se $T_{calc} = \frac{-6.63162-(-5.211378)}{0.79979} = -1.775769$, pelo que mais uma vez não se rejeita H_0. A admissibilidade de cada um destes valores sugere que a regra simples que foi proposta é uma alternativa simples viável. **NOTA:** Seria possível fazer um teste multivariado para testar a admissibilidade simultânea do conjunto dos três valores, mas essa matéria mais avançada não faz parte do programa da disciplina.

(c) A troca de variável resposta piorou claramente o valor de R^2 do ajustamento. Este resultado pode parecer surpreendente à primeira vista, uma vez que do ponto de vista algébrico, uma relação da forma $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ é equivalente a $x_2 = \frac{y-\beta_0-\beta_1 x_1}{\beta_2} = \beta_0^* + \beta_1^* x_1 + \beta_2^* y$ (com $\beta_0^* = \frac{-\beta_0}{\beta_2}$, $\beta_1^* = \frac{-\beta_1}{\beta_2}$ e $\beta_2^* = \frac{1}{\beta_2}$). Além disso, numa regressão linear simples, a troca do preditor e da variável resposta, se bem que muda a equação da recta ajustada, não muda a qualidade do ajustamento (uma vez que $R^2 = r_{xy}^2$, e o coeficiente de correlação é simétrico nos seus argumentos). Mas numa regressão linear múltipla, permutar a variável resposta com um dos preditores pode, como este exemplo ilustra, gerar um modelo de qualidade bastante diferente. O exemplo sugere a razão de ser deste facto: as variáveis `Volume` e `Diametro` estão fortemente correlacionadas entre si. Qualquer modelo de regressão linear que tenha uma dessas variáveis como variável resposta, e a outra como preditor, terá de ter $R^2 \geq (0.9671194)^2 = 0.9353199$. Mas a variável `Altura`, que foi agora colocada como variável resposta, não está fortemente correlacionada com nenhuma das duas outras. Ao desempenhar o papel de variável resposta, com as outras duas variáveis como preditores, o valor do R^2 resultante poderá ser elevado, mas como este exemplo ilustra, poderá não o ser.

19. Vamos contruir o intervalo de confiança a $(1-\alpha) \times 100\%$ para $\mathbf{a}^t\boldsymbol{\beta}$, a partir da distribuição indicada no enunciado. Sendo $t_{\frac{\alpha}{2}}$ o valor que, numa distribuição $t_{n-(p+1)}$, deixa à direita uma região de probabilidade $\alpha/2$, temos a seguinte afirmação probabilística, na qual trabalhamos a dupla desigualdade até deixar a combinação linear (para a qual se quer o intervalo de confiança) isolada no meio:

$$\begin{aligned}
P\left[-t_{\frac{\alpha}{2}} \;<\; \frac{\mathbf{a}^t\hat{\boldsymbol{\beta}} - \mathbf{a}^t\boldsymbol{\beta}}{\hat{\sigma}_{a^t\hat{\boldsymbol{\beta}}}} \;<\; t_{\frac{\alpha}{2}}\right] &= 1-\alpha \\
P\left[-t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{a^t\hat{\boldsymbol{\beta}}} \;<\; \mathbf{a}^t\hat{\boldsymbol{\beta}} - \mathbf{a}^t\boldsymbol{\beta} \;<\; t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{a^t\hat{\boldsymbol{\beta}}}\right] &= 1-\alpha \\
P\left[t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{a^t\hat{\boldsymbol{\beta}}} \;>\; \mathbf{a}^t\boldsymbol{\beta} - \mathbf{a}^t\hat{\boldsymbol{\beta}} \;>\; -t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{a^t\hat{\boldsymbol{\beta}}}\right] &= 1-\alpha \qquad (\textit{multiplicando por } -1) \\
P\left[\mathbf{a}^t\hat{\boldsymbol{\beta}} - t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{a^t\hat{\boldsymbol{\beta}}} \;<\; \mathbf{a}^t\boldsymbol{\beta} \;<\; \mathbf{a}^t\hat{\boldsymbol{\beta}} + t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{a^t\hat{\boldsymbol{\beta}}}\right] &= 1-\alpha
\end{aligned}$$

Assim, calculando o valor $\mathbf{a}^t\mathbf{b} = a_0b_0 + a_1b_1 + ... + a_pb_p$ do estimador $\mathbf{a}^t\hat{\boldsymbol{\beta}}$ e o erro padrão $\hat{\sigma}_{a^t\hat{\boldsymbol{\beta}}}$, para a nossa amostra, temos o intervalo a $(1-\alpha)\times 100\%$ de confiança para $\mathbf{a}^t\boldsymbol{\beta} = a_0\beta_0 + a_1\beta_1 + ... + a_p\beta_p$:

$$\Big]\quad \mathbf{a}^t\mathbf{b} - t_{\frac{\alpha}{2}[n-(p+1)]} \cdot \hat{\sigma}_{a^t\hat{\boldsymbol{\beta}}} \quad , \quad \mathbf{a}^t\mathbf{b} + t_{\frac{\alpha}{2}[n-(p+1)]} \cdot \hat{\sigma}_{a^t\hat{\boldsymbol{\beta}}} \quad \Big[$$

20. Parte-se duma regressão linear simples relacionando a variável resposta `Peso` e o preditor `Calibre`.

(a) A ordenada na origem natural é $\beta_0 = 0$: a calibre nulo corresponde inexistência de fruto, ou seja, peso nulo. O intervalo a 95% de confiança para a ordenada na origem é dado por:

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

No enunciado verifica-se que $b_0 = -210.3137$, com erro padrão associado $\hat{\sigma}_{\hat{\beta}_0} = 3.8078$. Tem-se ainda $t_{0.025(1271)} \approx 1.96$. Logo, o IC pedido é $]\,-217.777\,,\,-202.8504\,[$. Este intervalo está muito longe de incluir o valor natural $\beta_0 = 0$, pelo que essa eventualidade pode ser excluída. Não sendo um resultado encorajador, a verdade é que não faz sentido utilizar um modelo deste tipo para frutos de calibre próximo de zero. Os calibres utilizados no ajustamento do modelo variaram entre 53 e 79, pelo que deve evitar-se utilizar este modelo para calibres muito distantes da gama de calibres observados.

(b) Nesta alínea ajustou-se um polinómio de segundo grau, através dum modelo de regressão múltipla em que $X_1 =$ Calibre e $X_2 =$ Calibre^2. A equação de base neste modelo é Peso $= \beta_0 + \beta_1$ Calibre $+ \beta_2$ Calibre^2.

i. A equação da parábola ajustada é: Peso $= 72.33140 - 3.38747$ Calibre $+ 0.06469$ Calibre^2. Observe como a ordenada na origem e o coeficiente da variável `Calibre` são radicalmente diferentes do que eram na regressão linear simples.

ii. O modelo linear e o modelo quadrático são equivalentes caso $\beta_2 = 0$. Essa hipótese pode ser testada como qualquer outro teste t a um parâmetro β_j individual do modelo:

Hipóteses: $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$.

Estatística do teste: $T = \frac{\hat{\beta}_2 - 0}{\hat{\sigma}_{\hat{\beta}_2}} \;\cap\; t_{n-(p+1)}$

Nível de significância: $\alpha = 0.05$.

Região Crítica (Bilateral): Rejeitar H_0 se $|T_{calc}| > t_{\alpha/2\,(n-3)} = t_{0.025(1270)} \approx 1.962$.

Conclusões: O valor calculado da estatística do teste é dado no enunciado, na penúltima coluna da tabela *Coefficients*: $T_{calc} = \frac{0.06469}{0.01067} = 6.064$. Logo, rejeita-se claramente a hipótese nula $\beta_2 = 0$, pelo que o modelo polinomial (quadrático) tem um ajustamento significativamente melhor que o modelo linear. Repare-se como este resultado está associado a um aumento bastante pequeno do coeficiente de determinação R^2 (de 0.8638 para 0.8677). Este facto está, mais uma vez, associado à grande dimensão da amostra ($n = 1273$), que permite considerar significativas diferenças tão pequenas quanto estas.

21. (a) A matriz de projecção ortogonal $\mathbf{P} = \mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}_n^t$ é de dimensão $n \times n$ (confirme!), uma vez que o vector $\mathbf{1}_n$ é $n \times 1$. Mas o seu cálculo é facilitado pelo facto de que $\mathbf{1}_n^t\mathbf{1}_n$ é, neste caso, um escalar. Concretamente, $\mathbf{1}_n^t\mathbf{1}_n = n$, pelo que $(\mathbf{1}_n^t\mathbf{1}_n)^{-1} = \frac{1}{n}$. Logo $\mathbf{P} = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t$. O produto $\mathbf{1}_n\mathbf{1}_n^t$ resulta numa matriz $n \times n$ com todos os elementos iguais a 1 (não confundir com o produto pela ordem inversa, $\mathbf{1}_n^t\mathbf{1}_n$: recorde-se que o produto de matrizes **não** é comutativo). Assim,

$$\mathbf{P} \;=\; \frac{1}{n}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix}$$

(b) A projecção ortogonal do vector $\mathbf{x} = (x_1, x_2, ..., x_n)^t$ (cujos elementos serão por nós encarados como n observações duma variável X) sobre o subespaço gerado pelo vector dos uns $\mathbf{1}_n$ é:

$$\mathbf{P}\mathbf{x} \;=\; \frac{1}{n}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \;=\; \begin{bmatrix} \overline{x} \\ \overline{x} \\ \overline{x} \\ \vdots \\ \overline{x} \end{bmatrix} \;=\; \overline{x}\cdot\mathbf{1}_n$$

onde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ é a média dos valores do vector $\mathbf{x}$. Ou seja, o vector projectado é um múltiplo escalar do vector dos uns (como são todos os vectores que pertencem a $\mathcal{C}(\mathbf{1}_n)$, uma vez que as combinações lineares dum qualquer vector são sempre múltiplos escalares desse vector). Mas a constante de multiplicação desse vector projectado tem significado estatístico: é a média dos valores do vector $\mathbf{x}$.

(c) É característico da matriz identidade $\mathbf{I}$ que, para qualquer vector $\mathbf{x}$ se tem $\mathbf{I}\mathbf{x} = \mathbf{x}$. Logo, tendo em conta o resultado da alínea anterior, tem-se:

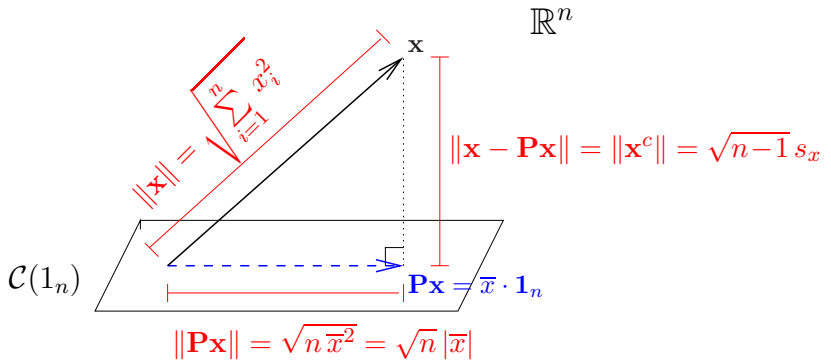
$$(\mathbf{I}-\mathbf{P})\mathbf{x} \;=\; \mathbf{I}\mathbf{x}-\mathbf{P}\mathbf{x} \;=\; \mathbf{x}-\mathbf{P}\mathbf{x} \;=\; \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} - \begin{bmatrix} \overline{x} \\ \overline{x} \\ \overline{x} \\ \vdots \\ \overline{x} \end{bmatrix} \;=\; \begin{bmatrix} x_1-\overline{x} \\ x_2-\overline{x} \\ x_3-\overline{x} \\ \vdots \\ x_n-\overline{x} \end{bmatrix} \;=\; \mathbf{x}^c$$

(d) A norma do vector $\mathbf{x}^c$ é, por definição, a raíz quadrada da soma dos quadrados dos seus elementos. Logo, tendo em conta a natureza dos elementos do vector $\mathbf{x}^c$ (ver a alínea anterior), tem-se:

$$\|\mathbf{x}^c\| \;=\; \sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2} \;=\; \sqrt{(n-1)\,s_x^2} \;=\; \sqrt{n-1}\;s_x\;,$$

ou seja, a norma é proporcional ao desvio padrão s_x dos valores do vector $\mathbf{x}$ (sendo a constante de proporcionalidade $\sqrt{n-1}$).

(e) A situação considerada nas alíneas anteriores tem a seguinte representação gráfica:



Nota: O subespaço $\mathcal{C}(\mathbf{1}_n)$ é gerado por um único vector, $\mathbf{1}_n$, pelo que em termos geométricos é uma linha recta que atravessa a origem (um subespaço de dimensão 1). Esse subespaço

foi representado aqui por um plano para manter coerência com as representações gráficas usadas nas aulas, salientando que se trata do mesmo conceito de projecções ortogonais.
Aplicando o Teorema de Pitágoras ao triângulo rectângulo indicado, tem-se:

$$\sum_{i=1}^{n} x_i^2 \;=\; (n-1)\, s_x^2 + n\,\overline{x}^2 \quad \Leftrightarrow \quad s_x^2 \;=\; \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 \;-\; n\,\overline{x}^2\right) \;,$$

que é a fórmula computacional da variância dada na disciplina de Estatística dos primeiros ciclos do ISA.

22. Note-se que a matriz $\mathbf{P}_{\mathbf{1}_n}$ referida neste exercício (e que será representada apenas por $\mathbf{P}$ no que se segue) é a mesma que foi discutida no Exercício 21. Assim, o vector $\mathbf{Y} - \mathbf{PY}$ é o vector centrado das observações de $\mathbf{Y}$:

$$\mathbf{Y} - \mathbf{PY} \;=\; \begin{bmatrix} Y_1 - \overline{Y} \\ Y_2 - \overline{Y} \\ Y_3 - \overline{Y} \\ \vdots \\ Y_n - \overline{Y} \end{bmatrix} \;=\; \mathbf{Y}^c$$

A norma deste vector, ao quadrado, é a soma dos quadrados dos seus elementos, ou seja, $SQT = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$. De forma análoga, e como o vector $\hat{\mathbf{Y}}$ dos valores ajustados é dado por $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} = \mathbf{HY}$, temos que o vector $\mathbf{HY} - \mathbf{PY}$ tem como elementos $\hat{Y}_i - \overline{Y}$:

$$\mathbf{HY} - \mathbf{PY} \;=\; \begin{bmatrix} \hat{Y}_1 - \overline{Y} \\ \hat{Y}_2 - \overline{Y} \\ \hat{Y}_3 - \overline{Y} \\ \vdots \\ \hat{Y}_n - \overline{Y} \end{bmatrix}$$

pelo que o quadrado da sua norma é $SQR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$. Finalmente, o vector $\mathbf{Y}-\mathbf{HY} = \mathbf{Y}-\hat{\mathbf{Y}}$ é o vector dos resíduos, e a sua norma ao quadrado é $SQRE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$.

Nas aulas viu-se geometricamente que o Teorema de Pitágoras garante que $SQT = SQR + SQRE$. Neste exercício pede-se para confirmar tal facto do ponto de vista algébrico. Tendo em conta que as Somas de Quadrados são os quadrados das normas acima indicados, e recordando as propriedades de normas, temos:

$$\begin{aligned} SQT &= \|\mathbf{Y} - \mathbf{PY}\|^2 \;=\; \|(\mathbf{Y} - \mathbf{HY}) + (\mathbf{HY} - \mathbf{PY})\|^2 \\ &= \|\mathbf{Y} - \mathbf{HY}\|^2 + \|\mathbf{HY} - \mathbf{PY}\|^2 + 2(\mathbf{Y} - \mathbf{HY})|(\mathbf{HY} - \mathbf{PY}) \\ &= SQR + SQRE + 2(\mathbf{Y} - \mathbf{HY})|(\mathbf{HY} - \mathbf{PY}) \end{aligned}$$

onde na última parcela surge o produto interno entre os vectores $\mathbf{Y} - \mathbf{HY}$ e $\mathbf{HY} - \mathbf{PY}$. Este produto interno tem de ser nulo, para ser verdade a relação entre as Somas de Quadrados. Ora,

$$\begin{aligned} (\mathbf{Y} - \mathbf{HY})|(\mathbf{HY} - \mathbf{PY}) &= (\mathbf{Y} - \mathbf{HY})^t(\mathbf{HY} - \mathbf{PY}) \\ &= \mathbf{Y}^t\mathbf{HY} - \mathbf{Y}^t\mathbf{PY} - (\mathbf{HY})^t\mathbf{HY} + (\mathbf{HY})^t\mathbf{PY} \\ &= \mathbf{Y}^t\mathbf{HY} - \mathbf{Y}^t\mathbf{PY} - \mathbf{Y}^t\mathbf{H}^t\mathbf{HY} + \mathbf{Y}^t\mathbf{H}^t\mathbf{PY} \;, \end{aligned} \tag{4}$$

tendo em conta que, em qualquer produto matricial, a transposta do produto é o produto das transpostas pela ordem inversa ($(\mathbf{AB})^t = \mathbf{B}^t\mathbf{A}^t$). Mas (tal como se viu no Exercício 11) $\mathbf{H}$ é uma matriz simétrica: $\mathbf{H}^t = [\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t]^t = \mathbf{X}[(\mathbf{X}^t\mathbf{X})^{-1}]^t\mathbf{X}^t = \mathbf{X}[(\mathbf{X}^t\mathbf{X})^t]^{-1}\mathbf{X}^t = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{H}$, tendo em conta que, para qualquer matriz invertível, a inversa da transposta é a transposta da inversa ($(\mathbf{A}^t)^{-1} = (\mathbf{A}^{-1})^t$), e que a transposta duma transposta é a matriz original ($(\mathbf{A}^t)^t = \mathbf{A}$). Por outro lado, $\mathbf{HH} = \mathbf{H}$, porque $\mathbf{HH} = [\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t][\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t] = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\cancel{(\mathbf{X}^t\mathbf{X})}\cancel{(\mathbf{X}^t\mathbf{X})^{-1}}\mathbf{X}^t = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{H}$. Logo, a terceira parcela na equação (4) vem igual à primeira ($\mathbf{Y}^t\mathbf{HY}$), mas de sinal contrário, cancelando. Por seu lado, e de novo usando a simetria de $\mathbf{H}$, a matriz da última parcela em (4) vem $\mathbf{H}^t\mathbf{P} = \mathbf{HP} = \mathbf{H}\mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}_n^t$. Mas (como se viu nas aulas teóricas) $\mathbf{H}\mathbf{1}_n = \mathbf{1}_n$, uma vez que o vector $\mathbf{1}_n$ pertence ao subespaço $\mathcal{C}(\mathbf{X})$ sobre o qual a matriz $\mathbf{H}$ projecta, e qualquer vector fica invariante quando projectado sobre um subespaço ao qual pertence. Logo, $\mathbf{HP} = \mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}_n^t = \mathbf{P}$. Assim, a última parcela da equação (4) vem igual à segunda ($\mathbf{Y}^t\mathbf{PY}$), mas com sinal trocado, pelo que essas duas parcelas também cancelam e o produto interno indicado nessa equação anula-se.

23. Em notação matricial/vectorial, a equação base deste modelo é $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ com $\mathbf{X} = \mathbf{1}_n$ e $\boldsymbol{\beta}$ o vector com um único elemento, β_0 (o único parâmetro do modelo).

    (a) A fórmula para o vector dos estimadores de mínimos quadrados do modelo linear contínua válida, pelo que $\hat{\beta}_0 = \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Como

$$\mathbf{X}^T\mathbf{Y} = \begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}\begin{bmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{bmatrix} = \sum_{i=1}^{n} Y_i, \qquad \text{e} \qquad \mathbf{X}^T\mathbf{X} = \begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix} = n,$$

    temos que $(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{n}$ e $\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}$. Ou seja, o estimador de mínimos quadrados de β_0 é a média amostral da variável Y.

    (b) $V[\hat{\beta}_0] = V[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \frac{\sigma^2}{n}$.

    (c) Admitindo os habituais pressupostos do modelo de regressão linear, continua válido que $\hat{\beta}_0 = \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ tem distribuição normal (multinormal com uma só componente), de média $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} = \beta_0$ e variância $V[\hat{\boldsymbol{\beta}}] = \frac{\sigma^2}{n}$ (como determinado na alínea b). Ou seja, $\hat{\beta}_0 \cap \mathcal{N}(\beta_0, \sigma^2/n)$.

    (d) Por definição $SQR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$. Considerando o modelo em estudo e o resultado obtido na alínea a), $\hat{Y}_i = \hat{\beta}_0 = \overline{Y}, \forall i = 1, \ldots, n$, pelo que $SQR = \sum_{i=1}^{n}(\overline{Y} - \overline{Y})^2 = 0$. Assim,

$$SQR = 0 \quad \text{e} \quad SQRE = SQT - SQR = SQT.$$

    Isto é, este modelo explica 0% da variabilidade total de Y. Toda a variabilidade de Y é residual.

    (e) Seja $\{Y_1, Y_2, \cdots, Y_n\}$ uma amostra aleatória duma população normal com média μ e variância σ^2, isto é, $Y_i \cap \mathcal{N}(\mu, \sigma^2), \forall i$ e $\{Y_i\}_{i=1}^n$ v.a. independentes. De acordo com os conhecimentos adquiridos na disciplina introdutória de Estatística (primeiros ciclos do ISA), $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ é um estimador de μ e $\overline{Y} \cap \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

    Considerando o modelo linear sem preditores e admitindo os usuais pressupostos, sabemos que $Y_i \cap \mathcal{N}(\beta_0, \sigma^2), \forall i$ e $\{Y_i\}_{i=1}^n$ são v.a. independentes, ou seja, estamos na situação considerada na outra disciplina de Estatística (com $\beta_0 = \mu$). Não admira assim que $\hat{\beta}_0 = \overline{Y}$

e que, como se viu na alínea c), $\hat{\beta}_0 \cap \mathcal{N}(\beta_0, \sigma^2/n)$. Isto é, numa situação em que só temos informação sobre a variável Y, a melhor maneira de a modelar, de estimar um novo valor dessa variável, é através da sua média amostral. A regressão linear, com um ou mais preditores, estende esta situação, admitindo que para prever novos valores de Y utilizamos informação extra, informação fornecida pelas variáveis preditoras.

(f) Vamos utilizar o teste F parcial para comparar um modelo com p preditores e o seu submodelo *sem preditores* ($k = 0$):

Modelo completo (C) $\quad Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

Submodelo (S) $\quad Y = \beta_0$

Hipóteses: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad$ vs. $\quad H_1 : \beta_1 \neq 0 \vee \beta_2 \neq 0 \vee \cdots \vee \beta_p \neq 0$

Estatística do Teste:

$$F = \frac{(SQRE_s - SQRE_c)/(p-k)}{SQRE_c/(n-(p-1))} \ \cap \ F_{(p-k,n-(p+1))}, \text{ sob } H_0$$

Como $k = 0$ e $SQRE_s = SQT$, temos que

$$F = \frac{(SQT - SQRE_c)/p}{SQRE_c/(n-(p-1))} = \frac{SQR_c/p}{SQRE_c/(n-(p-1))} = \frac{QMR_c}{QMRE_c} \ \cap \ F_{(p,n-(p+1))},$$

o que corresponde à estatística (e às hipóteses) do teste de ajustamento global do modelo completo (com p preditores). Ou seja, o teste de ajustamento global de um modelo não é mais do que um teste F parcial que compara esse modelo com o modelo nulo (sem preditores). A Hipótese Nula no teste de ajustamento global corresponde a dizer que o modelo completo não se distingue do modelo nulo.

24. Trata-se dum modelo linear, mas sem constante aditiva β_0. Neste caso, a matriz $\mathbf{X}$ do modelo (cujas colunas geram o subespaço onde se pretende ajustar o modelo) será constituída por uma única coluna, com os valores da variável preditora X (não existindo a usual coluna de uns, que estava associada à constante aditiva do modelo). O modelo, em forma matricial/vectorial, continua a escrever-se como $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$, embora agora $\boldsymbol{\beta}$ seja um escalar: β_1.

(a) Existe um único parâmetro no modelo (β_1) e a fórmula usual para o vector dos estimadores dos parâmetros no modelo linear (que continua válida, mas com a nova matriz $\mathbf{X}$ acima referida) vai produzir um único estimador. De facto, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$. Mas $\mathbf{X}^t\mathbf{Y}$ é o produto interno dos dois vectores $\mathbf{X}$ e $\mathbf{Y}$, com valor $\sum_{i=1}^{n} x_i\, Y_i$. Analogamente, $\mathbf{X}^t\mathbf{X} = \sum_{j=1}^{n} x_j^2$, pelo que $(\mathbf{X}^t\mathbf{X})^{-1} = \frac{1}{\sum_{i=j}^{n} x_j^2}$, ficando então $\hat{\boldsymbol{\beta}} = \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i\, Y_i}{\sum_{j=1}^{n} x_j^2}$.

(b) Com os pressupostos usuais no modelo de regressão linear, o vector das observações $\mathbf{Y}$ tem distribuição Multinormal, com vector médio $\mathbf{X}\boldsymbol{\beta} = \beta_1\mathbf{X}$ e matriz de variâncias-covariâncias $\sigma^2\mathbf{I}_n$, como no caso usual. Também se mantém válido que $\hat{\boldsymbol{\beta}} = \hat{\beta}_1 = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$ é o produto duma matriz constante, $(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, e do vector Multinormal $\mathbf{Y}$, logo terá distribuição Normal (Multinormal, mas com uma única componente), de média $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} = \beta_1$ e variância $V[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1} = \frac{\sigma^2}{\sum_{j=1}^{n} x_j^2}$.