INSTITUTO SUPERIOR DE AGRONOMIA

MODELOS MATEMÁTICOS e APLICAÇÕES– 2017-18

Resoluções de exercícios de Modelo Linear

1 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 $cov(x, cy) = c\, cov(x, y)$):

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

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

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

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

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

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

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

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

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

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

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

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

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

2. (a) Seguindo as instruções do enunciado, cria-se o ficheiro de texto `Azeite.txt` na directoria da sessão de trabalho do R. 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).

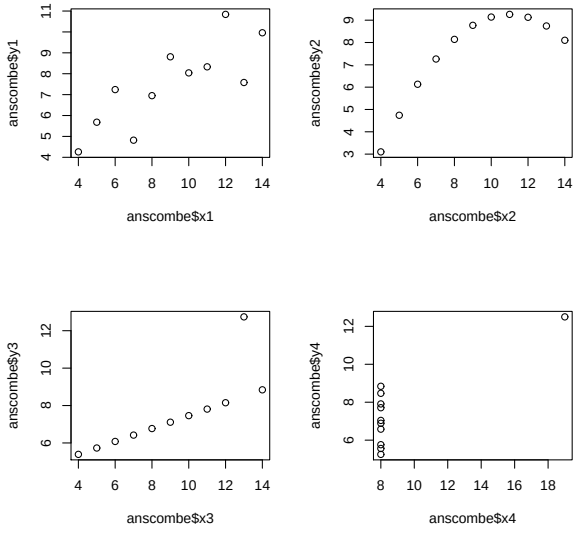
```
> apply(anscombe,2,var)
       x1        x2        x3        x4        y1        y2        y3        y4
11.000000 11.000000 11.000000 11.000000  4.127269  4.127629  4.122620  4.123249
```

(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. Tem-se:

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

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

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

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

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

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

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

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

   (c) Pela definição de coeficiente de correlação entre x e y, tem-se:

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

   (d) Por definição, $R^2 = \frac{SQR}{SQT}$. Sabemos que $SQT = (n-1)s_y^2$. Vamos verificar que $SQR = b_1^2(n-1)s_x^2$. De facto, recordando a definição dos valores ajustados de y e a expressão da ordenada na origem da recta de regressão, b_0, temos que $\hat{y}_i = b_0 + b_1\,x_i = \overline{y} + b_1(x_i-\overline{x})$. Logo,

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

   Tendo também em conta o resultado da alínea anterior, tem-se $R^2 = \frac{b_1^2\,s_x^2}{s_y^2} = (r_{xy})^2$.

   (e) Os valores ajustados $\hat{y}_i$ são dados por uma mesma transformação linear (afim) dos valores do preditor: $\hat{y}_i = b_0 + b_1\,x_i$. São conhecidas as propriedades destas transformações sobre a covariância e a variância. Assim,

$$r_{y\hat{y}}^2 = \left(\frac{cov_{y\hat{y}}}{s_y s_{\hat{y}}}\right)^2 = \frac{cov_{y,b_0+b_1x}^2}{s_y^2\,s_{b_0+b_1x}^2} = \frac{(b_1 cov_{y,x})^2}{s_y^2\,b_1^2 s_x^2} = \frac{\cancel{b_1^2}cov_{xy}^2}{\cancel{b_1^2}s_x^2\,s_y^2} = r_{xy}^2 = R^2.$$

Assim, o coeficiente de determinação duma regressão linear simples é também o quadrado do coeficiente de correlação linear entre os valores observados e os valores ajustados de y. Esta propriedade estende-se às regressões lineares múltiplas, embora seja necessário adaptar a justificação.

6. 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) É pedida a nuvem de pontos das transformações logarítimicas das duas variáveis da *data frame* `Animals`, que pode ser obtido duma das seguintes formas:

```
> 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.
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.

(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 gráfico indica que é aceitável admitir uma relação potência entre o peso do cérebro e o peso do corpo, nas espécies animais consideradas.

(d) 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)
```

(e) O declive $b_1^* = 0.49599$ 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 0.49599 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 $d = b_1^* = 0.49599$ 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 $y = 12.871\,x^{0.49599}$. 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.

(f) 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).

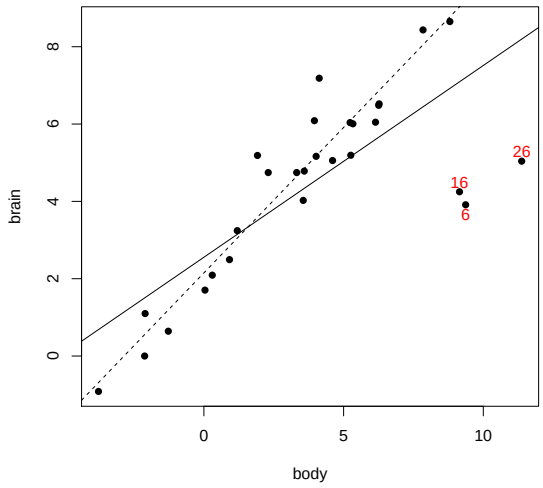
NOTA: É necessário explicitar as coordenadas dos pontos no gráfico que se vai aceder com o comando. No nosso caso, isso significa explicitar as coordenadas dos dados logaritmizados: `log(Animals)`.
O enunciado pede para identificar os pontos que se destacam da relação linear, e que são os pontos 6, 16 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
```

(g) Utilizando a indexação negativa para eliminar as três espécies de dinossáurios pode proceder-se ao reajustamento da regressão, modificando o argumento `data` do comando `lm`. Pode juntar-se a nova recta ao gráfico obtido antes, através do comando `abline`. Este comando será invocado com um argumento pedindo que a recta seja desenhada a tracejado, a fim de melhor a distinguir da recta originalmente obtida:

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

O gráfico resultante 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.



(h) 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)
(...)
```

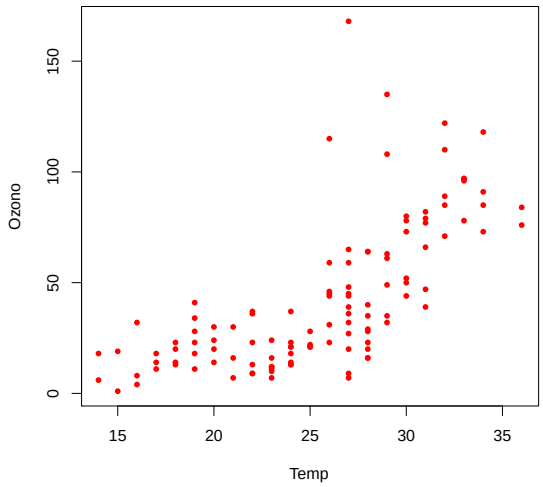
```
Multiple R-squared: 0.9217
(...)
```

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.

(i) O significado biológico dos parâmetros da recta é semelhante ao que foi visto na alínea 6e), 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.

7. (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
```

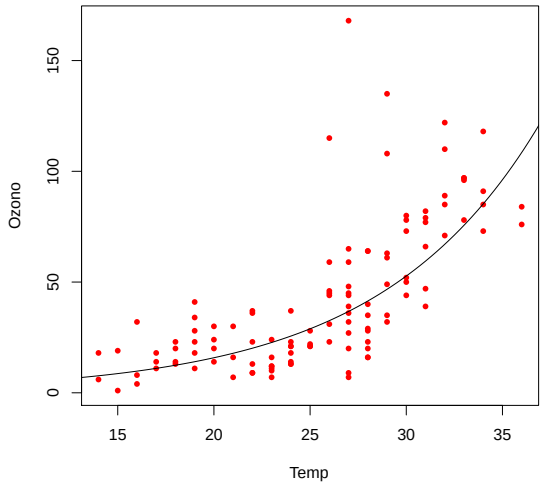
O coeficiente de determinação é de cerca de $R^2 = 0.73^2 = 0.53$ (aplicando o comando `summary` ao modelo agora ajustado verifica-se ser $R^2 = 0.5372$), o que significa que a regressão explica pouco mais de 53% da variabilidade dos log-teores de ozono.

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)
```



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

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

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

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

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

sendo os resultados obtidos os seguintes:

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

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

9. (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) 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 9a. 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. Para calcular a decomposição $SQT = SQR + SQRE$, comecemos por calcular a Soma de Quadrados Total, como $SQT = (n-1)\,s_y^2$:

```
> var(vinho.RLM$V8)
[1] 0.9977187
> dim(vinho.RLM)
[1] 178  13
> 177*var(vinho.RLM$V8)
[1] 176.5962
```

O valor de SQR é dado por: $SQR = R^2 \times SQT = 132.0004$, sendo $SQRE = SQT - SQR = 176.5962 - 132.0004 = 44.5958$.

(c) O modelo pedido no enunciado é:

```
> vinho.lm5 <- lm(V8 ~ V4 + V5 + V11 + V12 + V13 , data=vinho.RLM)
> vinho.lm5
Coefficients:
(Intercept)           V4           V5          V11          V12          V13
   -2.25196      0.53661     -0.04932      0.09053      0.95720      0.99496

> summary(vinho.lm5)$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`.

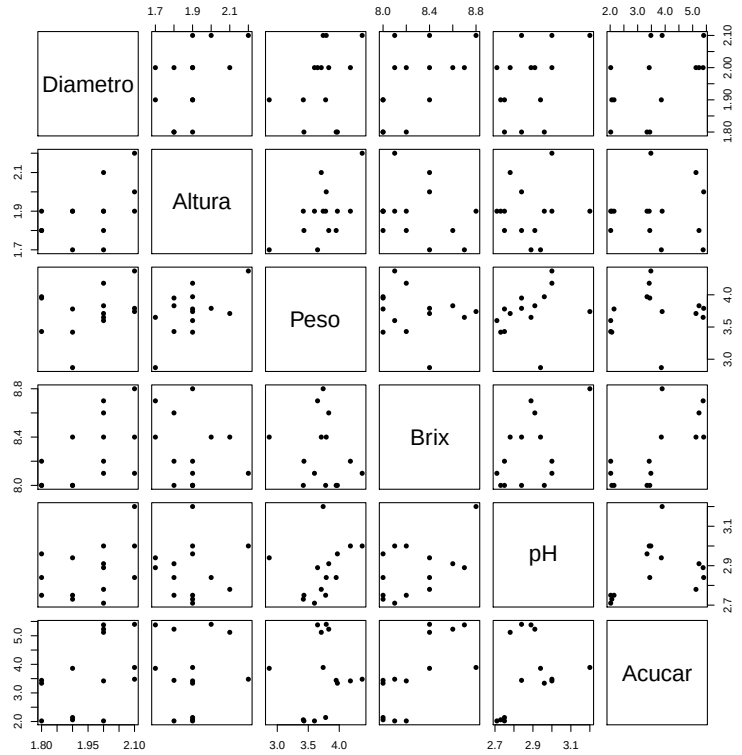
(d) Ajustando a mesma variável resposta `V8` sobre a totalidade das restantes variáveis obtem-se:

```
> vinho.lm13 <- 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(vinho.lm13))
[1] 151.4735
> 177*var(residuals(vinho.lm13))
[1] 25.12269
```

i. De novo, o valor da Soma de Quadrados Total já é conhecido da alínea acima: não depende do modelo ajustado, mas apenas da variância dos valores observados de Y (`V8`, neste exercício), que não se alterou. 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.*

10. (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 β_j, $j=1,2,3,4,5$, 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: $SQR = (n-1)\, s_{\hat{y}}^2 = 13 \times 0.06417822 = 0.8343169$. 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
```

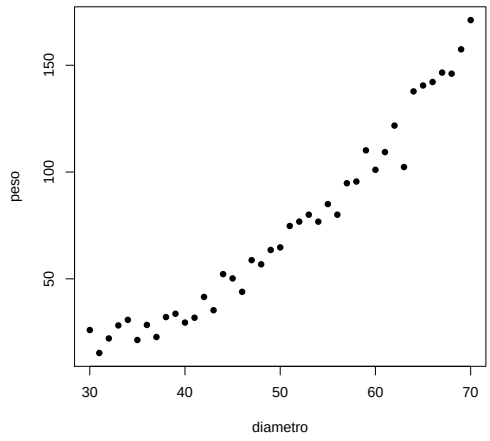
A matriz do modelo é a matriz de dimensões $n \times (p+1)$, cuja primeira coluna é uma coluna de n uns e cujas p colunas seguintes são as colunas dadas pelas n observações de cada uma das variáveis preditoras.
O vector $\mathbf{b}$ dos $p+1$ parâmetros ajustados é dado pelo produto matricial do enunciado: $\mathbf{b} = (\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\vec{\mathbf{y}})$. 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`.

11. (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
```

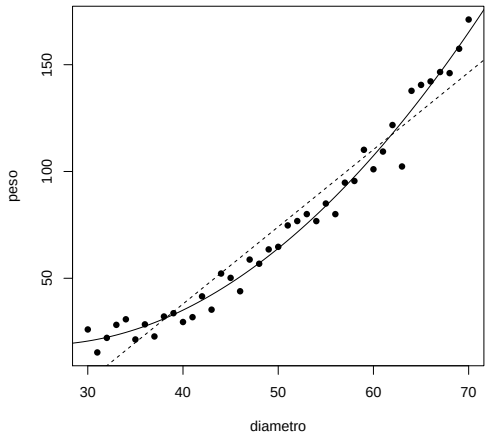
O ajustamento global deste modelo é muito bom. É possível interpretar o valor $R^2 = 0.9826$ da mesma forma que para qualquer outro modelo de regressão linear múltipla: este modelo explica cerca de $98,26\%$ da variabilidade dos pesos das ameixas.

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:

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

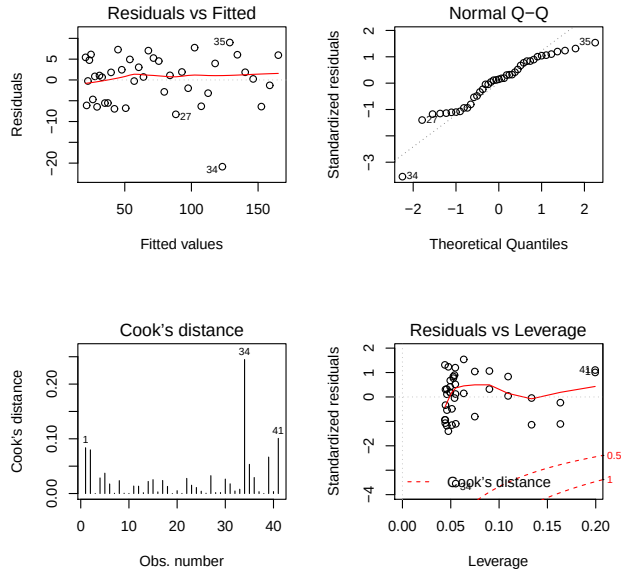
Para desenhar esta parábola em cima da nuvem de pontos criada acima, já não é possível usar o comando `abline` (que apenas serve para traçar rectas). Podemos, no entanto, usar o comando `curve`, como se ilustra seguidamente. O argumento `add=TRUE` usado nesse comando serve para que o gráfico da função 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 11b 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.

12. Comecemos por recordar alguns resultados já previamente discutidos:

- Sabemos que, para qualquer conjunto de n pares de observações, se tem:

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

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

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

- 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 $\vec{\mathbf{x}}$ e $\vec{\mathbf{y}}$ é dado por $\vec{\mathbf{x}}^t\vec{\mathbf{y}} = \sum_{i=1}^{n} x_i\, y_i$. 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: $\mathbf{1}_n^t\vec{\mathbf{x}} = \sum_{i=1}^{n} x_i = n\,\overline{x}$.
- 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 $\vec{\mathbf{x}}$. Logo, a sua transposta $\mathbf{X}^t$ é de dimensão $2 \times n$. Como o vector $\vec{\mathbf{y}}$ é de dimensão $n \times 1$, o produto $\mathbf{X}\vec{\mathbf{y}}$ 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 $\vec{\mathbf{y}}$, ou seja, por $\mathbf{1}_n^t \vec{\mathbf{y}} = \sum_{i=1}^{n} y_i = n\,\overline{y}$. 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 $\vec{\mathbf{y}}$, ou seja, por $\vec{\mathbf{x}}^t \vec{\mathbf{y}} = \sum_{i=1}^{n} x_i\, y_i = (n{-}1)\,\mathrm{cov}_{xy} + n\,\overline{x}\,\overline{y}$, 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 $\mathbf{X}$ ($\vec{\mathbf{x}}$), logo é $\mathbf{1}_n^t \vec{\mathbf{x}} = \sum_{i=1}^{n} x_i = n\,\overline{x}$. O elemento na posição (2,1) é o produto interno da segunda linha de $\mathbf{X}^t$ ($\vec{\mathbf{x}}$) 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 $\mathbf{X}^t$ ($\vec{\mathbf{x}}$) com a segunda coluna de $\mathbf{X}$ ($\vec{\mathbf{x}}$), ou seja, $\vec{\mathbf{x}}^t \vec{\mathbf{x}} = \sum_{i=1}^{n} x_i^2$. 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

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

13. 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 $\hat{\vec{\mathbf{y}}}=\mathbf{H}\vec{\mathbf{y}}$. 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 $\hat{\vec{\mathbf{y}}}$, uma vez que esse produto interno devolve a soma dos elementos de $\hat{\vec{\mathbf{y}}}$. Assim, a média dos valores ajustados é $\overline{\hat{y}}=\frac{1}{n}\mathbf{1}_n^t\hat{\vec{\mathbf{y}}}=\frac{1}{n}\mathbf{1}_n^t\mathbf{H}\vec{\mathbf{y}}=\frac{1}{n}(\mathbf{H}\mathbf{1}_n)^t\vec{\mathbf{y}}=\frac{1}{n}\mathbf{1}_n^t\vec{\mathbf{y}}$, uma vez que $\mathbf{H}\mathbf{1}_n=\mathbf{1}_n$, 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, $\frac{1}{n}\mathbf{1}_n^t\vec{\mathbf{y}}$ é a média $\overline{y}$ dos valores observados de Y (já que $\mathbf{1}_n^t\vec{\mathbf{y}}$ devolve a soma dos elementos do vector dessas observações, $\vec{\mathbf{y}}$). 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 $\mathbf{e}=\vec{\mathbf{y}}-\hat{\vec{\mathbf{y}}}=\vec{\mathbf{y}}-\mathbf{H}\vec{\mathbf{y}}$. 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) $\mathbf{1}_n^t\mathbf{e}=\mathbf{1}_n^t(\vec{\mathbf{y}}-\mathbf{H}\vec{\mathbf{y}})=\mathbf{1}_n^t\vec{\mathbf{y}}-\mathbf{1}_n^t\mathbf{H}\vec{\mathbf{y}}=\mathbf{1}_n^t\vec{\mathbf{y}}-\mathbf{1}_n^t\vec{\mathbf{y}}=0$.

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

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

(a) As estimativas dos desvios padrão associados à estimação de cada um dos parâmetros são indicadas na tabela, na coluna de nome `Std.Error` (ou seja, erro padrão). Assim, o desvio padrão associado à estimação da ordenada na origem é $\hat{\sigma}_{\hat{\beta}_0} = 0.039762$. A variância correspondente é o quadrado deste valor, $\hat{\sigma}^2_{\hat{\beta}_0} = 0.001581$. Seria igualmente possível calcular esta variância estimada a partir da sua fórmula: $\hat{\sigma}^2_{\hat{\beta}_0} = QMRE \cdot (\mathbf{X}^t\mathbf{X})^{-1}_{(1,1)}$. 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: $H_0 : \beta_1 = 0.5 \quad$ vs. $\quad H_1 : \beta_1 \neq 0.5$.

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

Nível de significância: $\alpha = 0.05$.

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

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

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

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

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

Nível de significância: $\alpha = 0.05$.

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

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

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

da pétala). Esta conclusão também se pode justificar a partir do valor de prova ($p-value$) do valor calculado da estatística, que é muito pequeno, sendo mesmo inferior à precisão de máquina, $p < 2 \times 10^{-16}$. Mesmo para níveis de significância como $\alpha = 0.01$ ou $\alpha = 0.005$, a conclusão seria a de rejeição de H_0.

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

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

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

Nível de significância: $\alpha = 0.05$.

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

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

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

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

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

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

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

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:

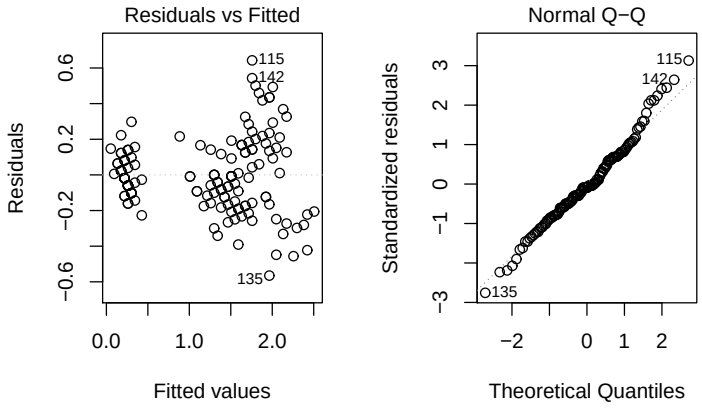
```
> predict(iris.lm, new=data.frame(Petal.Length=4.5), int="pred")
       fit      lwr      upr
1 1.507824 1.098187 1.917461
```

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))
```

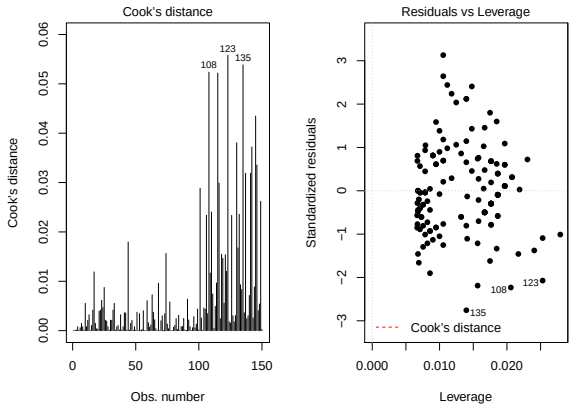
verifica-se que pode existir um problema em relação à hipótese de homogeneidade de variâncias. O gráfico da esquerda sugere que os lírios com comprimento de pétala mais pequeno (do lado esquerdo do gráfico) parecem ter menor variabilidade dos resíduos do que os restantes. Já a linearidade aproximada 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 $var(c\,x) = c^2\,var(x)$:

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

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:

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

e:

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

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

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

15. 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)

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

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

(e)

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

16. (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.

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

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

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

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

```
> Animals.lm <- lm(log(brain) ~ log(body) , data=Animals)
> summary(Animals.lm)
(...)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  2.55490    0.41314   6.184 1.53e-06
log(body)    0.49599    0.07817   6.345 1.02e-06
---
Residual standard error: 1.532 on 26 degrees of freedom
Multiple R-squared: 0.6076,Adjusted R-squared: 0.5925
F-statistic: 40.26 on 1 and 26 DF,  p-value: 1.017e-06
```

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

Hipóteses: $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

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

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral direita): Rej. H_0 se $F_{calc} > f_{\alpha\,(1,n-2)} = f_{0.05(1,26)} = 4.225201$.

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

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

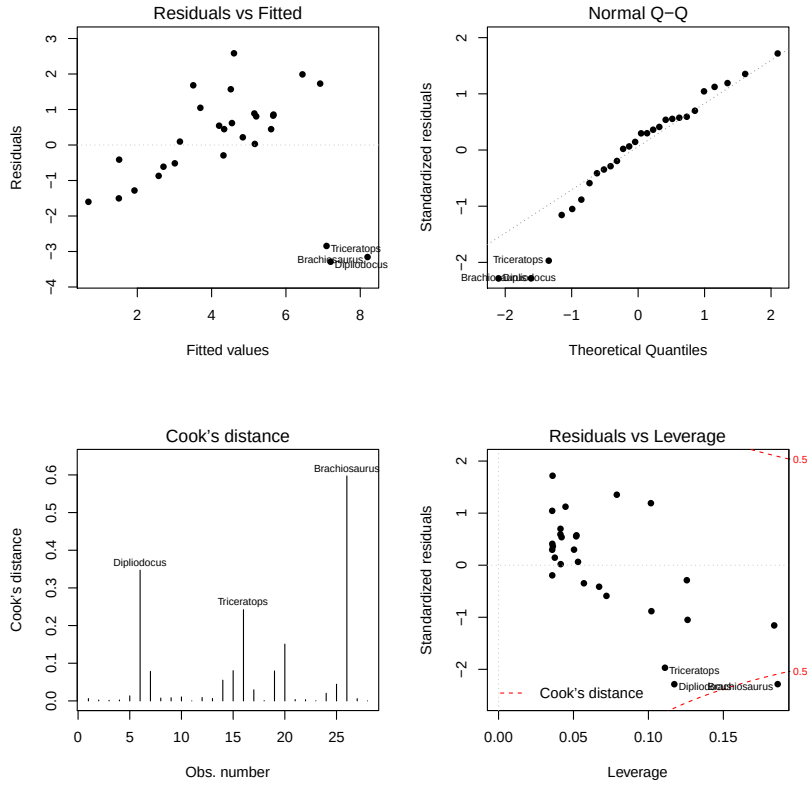
(c) Pretende-se o intervalo a 95% de confiança para β_1, ou seja:

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

com $b_1 = 0.49599$, $t_{0.025(26)} = 2.055529$ e $\hat{\sigma}_{\hat{\beta}_1} = 0.07817$. Ou seja, o intervalo é $]\,0.335\,,\,0.657\,[$. Uma relação isométrica corresponde a admitir que o declive da recta populacional é $\beta_1 = 1$, ou seja que as taxas de variação relativas de peso do corpo e peso do cérebro são iguais (ver a resolução do exercício 6). 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 6.

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

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

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

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

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

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

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

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

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

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

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

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

O valor da raíz quadrada é agora:

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

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

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

Para obter o intervalo de predição para os valores do *peso do cérebro* (sem logaritmização), basta tomar as exponenciais dos extremos do intervalo acima referido. De facto, se (ao nível 95% e para $x = 250kg$) o intervalo de predição para $Y^* = \log(Y)$ é: $4.755 < log(Y) < 7.853$, então a dupla desigualdade equivalente $\mathrm{e}^{4.755} = 116.16 < Y < 2573.443 = \mathrm{e}^{7.853}$ 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.

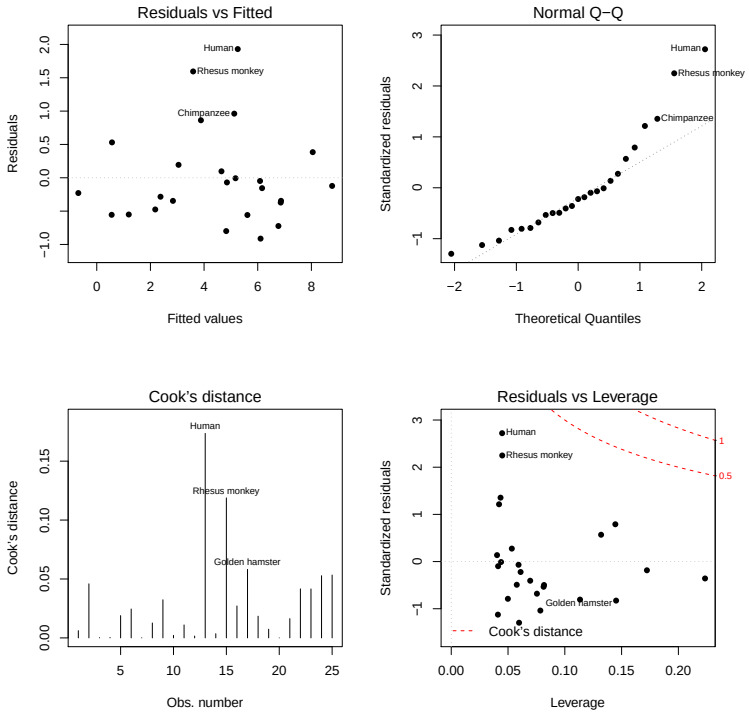
NOTA: Na alínea anterior não se pode efectuar uma transformação análoga, uma vez que valor esperado e logaritmização não são operações intercambiáveis. Ou seja, $E[\log(Y)] \neq \log(E[Y])$, pelo que não sabemos como transformar a dupla desigualdade $a < E[\log(Y)] < b$ numa dupla desigualdade equivalente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

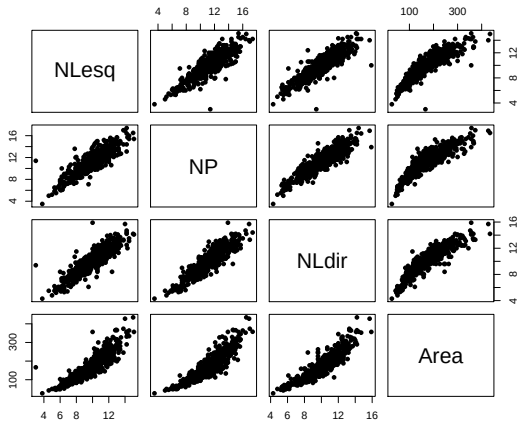


18. Na *data frame* `videiras`, a primeira coluna indica a casta, pelo que não será de utilidade neste exercício.

    (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: $H_0 : \beta_1 = 7 \quad \text{vs.} \quad H_1 : \beta_1 \neq 7$

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 $\vec{\hat{\boldsymbol{\beta}}}$, que é estimada por $\hat{V}[\vec{\hat{\boldsymbol{\beta}}}] = QMRE\,(\mathbf{X}^t\mathbf{X})^{-1}$. 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 18d, obtém-se os seguintes valores estimados:

- *Folha 1*: $\widehat{Area} = -168.111+9.987\times12.1+11.078\times11.6+11.895\times11.9 = 222.787\,cm^2$;
- *Folha 2*: $\widehat{Area} = -168.111+9.987\times10.6+11.078\times10.1+11.895\times9.9 = 167.3995\,cm^2$;
- *Folha 3*: $\widehat{Area} = -168.111+9.987\times15.1+11.078\times14.9+11.895\times14.0 = 314.2849\,cm^2$;

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
```

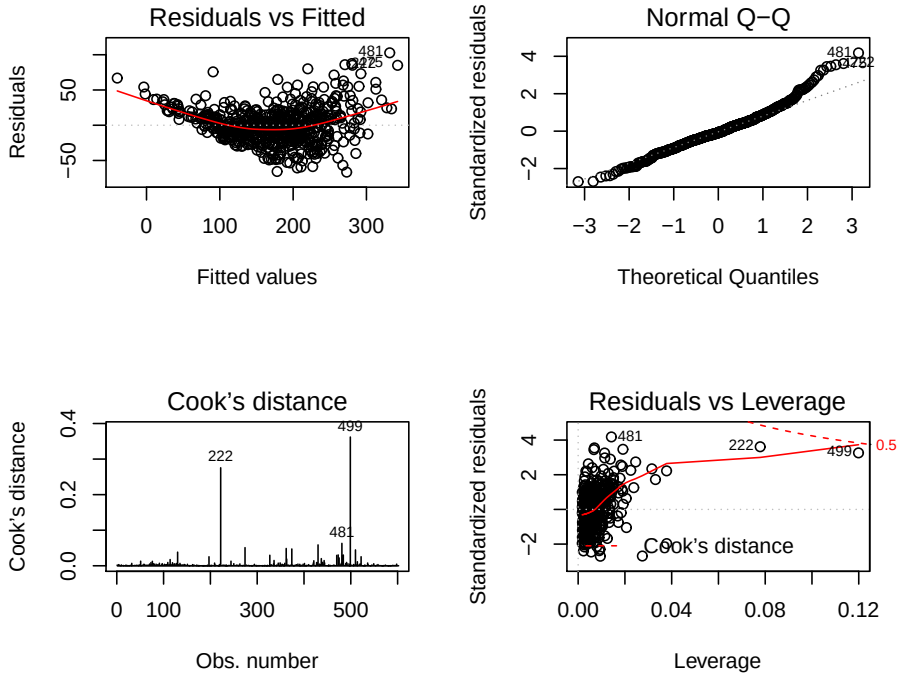
Assim, têm-se os seguintes intervalos de predição a 95% para os três valores de Y:

- *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))
```

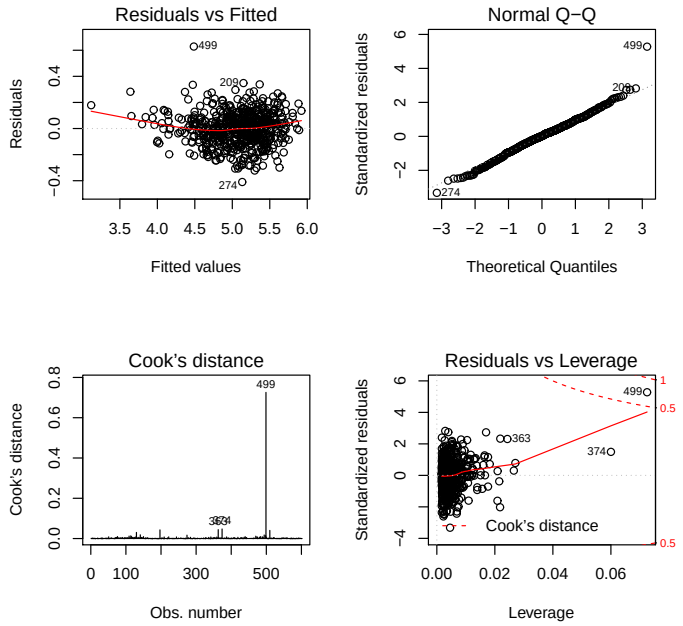
```
(...)
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              -0.53096    0.08138  -6.524 1.46e-10 ***
log(NP)                   0.67738    0.06479  10.455  < 2e-16 ***
I(log((NLesq + NLdir)))   1.33818    0.06680  20.032  < 2e-16 ***
---
(...)
Residual 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.

19. (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
```

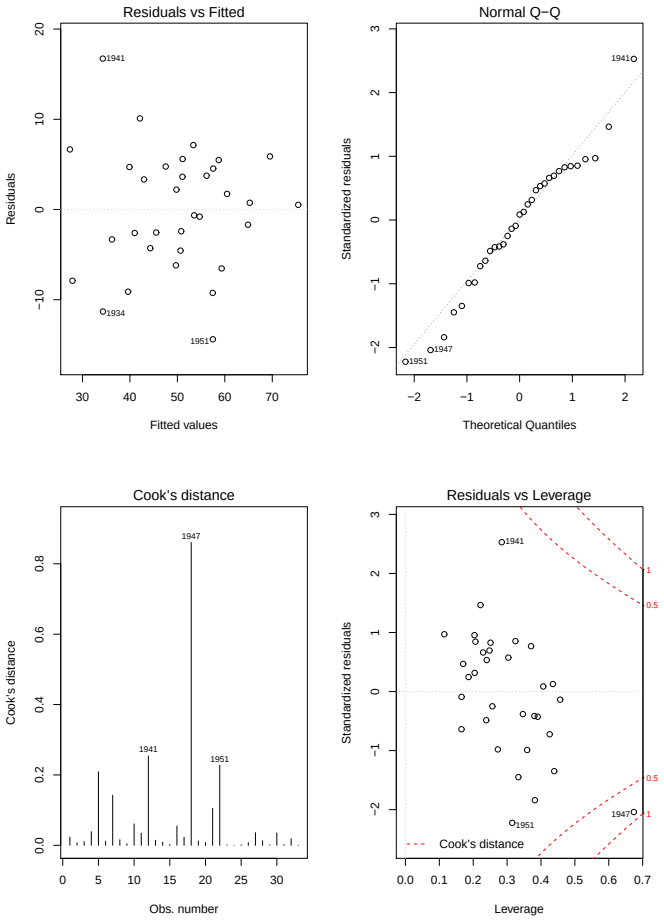
O facto mais saliente resultante da exclusão do preditor x_1 é a queda acentuada no valor do coeficiente de determinação, que é agora apenas $R^2 = 0.5074$ (repare-se como o $R^2_{mod} = 0.3432$ ainda se distancia mais do R^2 usual, reflectindo também esse ajustamento mais pobre). Assim, este modelo 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 estandardi-

zado 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,

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

Estatística do Teste: $F = \frac{n-(p+1)}{p-k} \cdot \frac{R^2_c - R^2_s}{1-R^2_c} \ \cap\ F_{(p-k,n-(p+1))}$, 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 $n = 33$, $p = 9$, $k = 5$, $R^2_c = 0.7476$ e $R^2_s = 0.6435$.

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 19a), 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 19e), 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.

20. (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:

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

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.

21. Vamos contruir o intervalo de confiança a $(1-\alpha)\times 100\%$ para $\mathbf{a}^t\vec{\boldsymbol{\beta}}$, 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:

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

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

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

22. 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 $X_1 = \text{Calibre}$ e $X_2 = \text{Calibre}^2$. A equação de base neste modelo é $\text{Peso} = \beta_0 + \beta_1 \,\text{Calibre} + \beta_2 \,\text{Calibre}^2$.

i. A equação da parábola ajustada é: $\text{Peso} = 72.33140 - 3.38747 \,\text{Calibre} + 0.06469 \,\text{Calibre}^2$. 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.

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

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

onde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ é a média dos valores do vector $\vec{\mathbf{x}}$. 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 $\vec{\mathbf{x}}$.

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

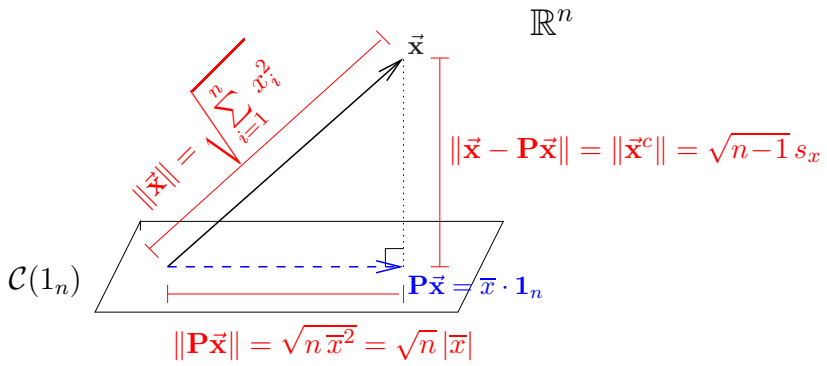
$$(\mathbf{I}-\mathbf{P})\vec{\mathbf{x}} \;=\; \mathbf{I}\vec{\mathbf{x}} - \mathbf{P}\vec{\mathbf{x}} \;=\; \vec{\mathbf{x}} - \mathbf{P}\vec{\mathbf{x}} \;=\; \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} - \begin{bmatrix} \overline{x} \\ \overline{x} \\ \overline{x} \\ \vdots \\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

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

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

ou seja, a norma é proporcional ao desvio padrão s_x dos valores do vector $\vec{\mathbf{x}}$ (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.

24. 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 23. Assim, o vector $\vec{\mathbf{Y}} - \mathbf{P}\vec{\mathbf{Y}}$ é o vector centrado

das observações de $\vec{\mathbf{Y}}$:

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

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 $\hat{\vec{\mathbf{Y}}}$ dos valores ajustados é dado por $\hat{\vec{\mathbf{Y}}} = \mathbf{X}\hat{\vec{\boldsymbol{\beta}}} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}} = \mathbf{H}\vec{\mathbf{Y}}$, temos que o vector $\mathbf{H}\vec{\mathbf{Y}} - \mathbf{P}\vec{\mathbf{Y}}$ tem como elementos $\hat{Y}_i - \overline{Y}$:

$$\mathbf{H}\vec{\mathbf{Y}} - \mathbf{P}\vec{\mathbf{Y}} \;=\; \begin{bmatrix} \hat{Y}_1 - \overline{Y} \\ \hat{Y}_2 - \overline{Y} \\ \hat{Y}_3 - \overline{Y} \\ \vdots \\ \hat{Y}_n - \overline{Y} \end{bmatrix}$$

pelo que o quadrado da sua norma é $SQR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$. Finalmente, o vector $\vec{\mathbf{Y}} - \mathbf{H}\vec{\mathbf{Y}} = \vec{\mathbf{Y}} - \hat{\vec{\mathbf{Y}}}$ 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:

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

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

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

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 13) $\mathbf{H}$ 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 ($\vec{\mathbf{Y}}^t\mathbf{H}\vec{\mathbf{Y}}$), 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, $\mathbf{H}\mathbf{P} = \mathbf{1}_n(\mathbf{1}_n^t\mathbf{1}_n)^{-1}\mathbf{1}_n^t = \mathbf{P}$. Assim, a última parcela da equação (4) vem igual à segunda ($\vec{\mathbf{Y}}^t\mathbf{P}\vec{\mathbf{Y}}$), mas com sinal trocado, pelo que essas duas parcelas também cancelam e o produto interno indicado nessa equação anula-se.

25. 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 $\hat{\beta}_0 = \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{\mathbf{Y}}$. Como

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

    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 $\hat{\beta}_0 = \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{\mathbf{Y}}$ 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.

26. 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 $\vec{\mathbf{Y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \epsilon$, embora agora $\vec{\boldsymbol{\beta}}$ 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, $\hat{\vec{\boldsymbol{\beta}}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}}$. Mas $\mathbf{X}^t\vec{\mathbf{Y}}$ é o produto interno dos dois vectores $\mathbf{X}$ e $\vec{\mathbf{Y}}$, 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 $\hat{\vec{\boldsymbol{\beta}}} = \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i\, Y_i}{\sum_{j=1}^{n} x_j^2}$.

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

27. (a) O modelo de regressão linear múltipla relaciona uma variável resposta Y com p variáveis preditoras X_1, X_2, ..., X_p. Designando por $\vec{\mathbf{Y}}$ o vector das n observações da variável resposta Y, $\vec{\boldsymbol{\epsilon}}$ o vector dos n erros aleatórios correspondentes, $\vec{\boldsymbol{\beta}}$ o vector dos $p+1$ parâmetros do modelo, $\beta_0, \beta_1, ..., \beta_p$, e $\mathbf{X}$ a matriz $n \times (p+1)$, cuja primeira coluna é constituída por n

uns e cada uma das restantes p colunas contém as n observações duma variável preditora, tem-se:

$$\vec{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & x_{1_{(1)}} & x_{2_{(1)}} & \cdots & x_{p_{(1)}} \\ 1 & x_{1_{(2)}} & x_{2_{(2)}} & \cdots & x_{p_{(2)}} \\ 1 & x_{1_{(3)}} & x_{2_{(3)}} & \cdots & x_{p_{(3)}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1_{(n)}} & x_{2_{(n)}} & \cdots & x_{p_{(n)}} \end{bmatrix}, \ \vec{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \ \vec{\boldsymbol{\epsilon}} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

O modelo de regressão linear múltipla é então dado por:

i. $\vec{\mathbf{Y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}$

ii. $\vec{\boldsymbol{\epsilon}} \cap \mathcal{N}_n(\vec{\mathbf{0}},\ \sigma^2\mathbf{I}_n)$,

sendo $\vec{\mathbf{0}}$ o vector de n zeros e $\mathbf{I}_n$ a matriz identidade $n \times n$. Na segunda condição, indica-se que o vector dos erros aleatórios segue uma distribuição Multinormal, com vector médio dado pelo vector de zeros (ou seja, cada erro aleatório individual tem valor esperado zero) e matriz de variâncias-covariâncias diagonal, com os elementos diagonais todos iguais a σ^2. Uma vez que, numa matriz de (co-)variâncias os elementos diagonais representam as variâncias de cada componente do vector, esta condição indica que $V[\epsilon_i] = \sigma^2$, $\forall\, i$. O facto de os elementos não diagonais da matriz $\sigma^2\,\mathbf{I}_n$ serem todos nulos equivale a dizer que a covariância entre elementos diferentes do vector aleatório dos erros é sempre nula (ou seja, $Cov[\epsilon_i, \epsilon_j] = 0$, sempre que $i \neq j$) e, como sabemos, numa distribuição Multinormal tal facto implica a independência desses elementos.

(b) O vector $\hat{\vec{\boldsymbol{\beta}}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)^t$ dos estimadores dos $p+1$ parâmetros dum modelo linear é dado (ver formulário) por $\hat{\vec{\boldsymbol{\beta}}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}}$. Mas, pelo modelo, tem-se $\vec{\mathbf{Y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}$. Substituindo, tem-se:

$$\hat{\vec{\boldsymbol{\beta}}} \ = \ (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}) \ = \ \underbrace{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}}_{=\mathbf{I}}\vec{\boldsymbol{\beta}} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}} \ = \ \vec{\boldsymbol{\beta}} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}\,,$$

como se pedia para mostrar.

(c) A expressão da alínea anterior é a soma dum vector não aleatório, $\vec{\boldsymbol{\beta}}$, com um vector aleatório, $(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}$. Ora, para qualquer vector aleatório $\vec{\mathbf{W}}$ e vector não aleatório $\mathbf{a}$ verifica-se $E[\vec{\mathbf{W}} + \mathbf{a}] = E[\vec{\mathbf{W}}] + \mathbf{a}$. Logo, no nosso caso, tem-se: $E[\hat{\vec{\boldsymbol{\beta}}}] = E[\vec{\boldsymbol{\beta}} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}] = \vec{\boldsymbol{\beta}} + E[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}]$. A segunda parcela é o vector esperado dum vector que resulta de multiplicar uma matriz não aleatória ($(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$) por um vector aleatório ($\vec{\boldsymbol{\epsilon}}$). Por outra propriedade operatória dos vectores esperados, tem-se $E[\mathbf{B}\vec{\mathbf{W}}] = \mathbf{B}\,E[\vec{\mathbf{W}}]$, onde $\mathbf{B}$ é uma matriz não aleatória. Assim, $E[\hat{\vec{\boldsymbol{\beta}}}] = \vec{\boldsymbol{\beta}} + E[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}] = \vec{\boldsymbol{\beta}} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\underbrace{E[\vec{\boldsymbol{\epsilon}}]}_{=\vec{\mathbf{0}}} = \vec{\boldsymbol{\beta}} + \vec{\mathbf{0}} = \vec{\boldsymbol{\beta}}$.

Por outro lado, tendo em conta a propriedade operatória geral de matrizes de (co-)variâncias, $V[\vec{\mathbf{W}} + \mathbf{a}] = V[\vec{\mathbf{W}}]$, tem-se $V[\hat{\vec{\boldsymbol{\beta}}}] = V[\vec{\boldsymbol{\beta}} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}] = V[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}]$. Outra propriedade operatória de matrizes de (co-)variâncias diz-nos que $V[\mathbf{B}\vec{\mathbf{W}}] = \mathbf{B}V[\vec{\mathbf{W}}]\mathbf{B}^t$, para uma matriz não aleatória $\mathbf{B}$. Logo (e sendo no nosso caso $\mathbf{B} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$), tem-se:

$$\begin{aligned} V[\hat{\vec{\boldsymbol{\beta}}}] \ &= \ V[(\mathbf{X}^t\mathbf{X})^{-1}\,\mathbf{X}^t\vec{\boldsymbol{\epsilon}}] \ = \ (\mathbf{X}^t\mathbf{X})^{-1}\,\mathbf{X}^t\,V[\vec{\boldsymbol{\epsilon}}]\,\left[(\mathbf{X}^t\mathbf{X})^{-1}\,\mathbf{X}^t\right]^t \\ &= \ (\mathbf{X}^t\mathbf{X})^{-1}\,\mathbf{X}^t\,\sigma^2\mathbf{I}_n\,\mathbf{X}\left[(\mathbf{X}^t\mathbf{X})^{-1}\right]^t \ = \ \sigma^2\cancel{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}}\left[(\mathbf{X}^t\mathbf{X})^t\right]^{-1} \ = \ \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\,. \end{aligned}$$

28. O R^2 modificado definiu-se como $R^2_{mod} = 1 - \frac{QMRE}{QMT}$ onde $QMT = \frac{SQT}{n-1} = s^2_y$.

    (a) Tem-se $R^2_{mod} = 1 - \frac{QMRE}{QMT} = 1 - \frac{\frac{SQRE}{n-(p+1)}}{\frac{SQT}{n-1}} = 1 - \frac{n-1}{n-(p+1)} \times \frac{SQRE}{SQT}$. Mas $\frac{SQRE}{SQT} = \frac{SQT-SQR}{SQT} = 1 - R^2$. Substituindo na expressão anterior, tem-se o resultado pretendido: $R^2_{mod} = 1 - (1-R^2)\frac{n-1}{n-(p+1)}$.

    (b) Por definição, a estatística do teste F de ajustamento global tem valor $F_{calc} = \frac{QMR}{QMRE} = \frac{n-(p+1)}{p} \cdot \frac{R^2}{1-R^2}$. Ora, usando a expressão da alínea anterior, tem-se $R^2 - R^2_{mod} = R^2 - 1 + (1 - R^2)\frac{n-1}{n-(p+1)} = (1-R^2)\left[-1 + \frac{n-1}{n-(p+1)}\right] = (1-R^2)\left[\frac{\cancel{-n}+(p+\cancel{1})+\cancel{n}-\cancel{1}}{n-(p+1)}\right] = (1-R^2)\frac{p}{n-(p+1)}$. Logo, $\frac{R^2}{R^2 - R^2_{mod}} = \frac{R^2}{(1-R^2)\frac{p}{n-(p+1)}} = \frac{n-(p+1)}{p} \times \frac{R^2}{1-R^2} = F_{calc}$, como se queria mostrar.

    (c) Usando os resultados da primeira alínea, tem-se $R^2_{mod} < 0 \Leftrightarrow 1 - (1-R^2)\frac{n-1}{n-(p+1)} < 0 \Leftrightarrow 1 < (1-R^2)\frac{n-1}{n-(p+1)} \Leftrightarrow \frac{n-(p+1)}{n-1} < 1 - R^2 \Leftrightarrow R^2 < 1 - \frac{n-(p+1)}{n-1} = \frac{\cancel{n}-\cancel{1}-\cancel{n}+p-\cancel{1}}{n-1} = \frac{p}{n-1}$, como se pedia para mostrar. Se esta condição se verifica, tem-se, a partir da expressão da alínea anterior, que F_{calc} terá valor inferior a 1. Uma rápida vista de olhos pelas tabelas da distribuição F mostra que valores de F_{calc} inferiores a 1 nunca conduzem (para os níveis de significância usuais) à rejeição da H_0, pelo que o modelo ajustado não passa o teste de ajustamento global. Assim, ajustamentos de modelos em que p seja pouco menor do que $n-1$ podem não passar o teste de ajustamento global, mesmo com valores relativamente elevados de R^2.

2 Análise de Variância

1. (a) Trata-se dum delineamento a um único factor (as variedades de tomate), sendo a variável resposta Y a resistência da película (em gf). Em cada um dos $k = 6$ níveis do factor há $n_c = 3$ repetições (as parcelas). O número igual de repetições nas 6 situações experimentais significa que o delineamento é equilibrado. O modelo ANOVA a um factor corresponde a:

    i. A resistência Y_{ij}, na j-ésima parcela ($j = 1, 2, 3$) associada à variedade i ($i = 1, ..., 6$), é dada por:

    $$Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij} \ , \quad \forall\, i, j \ ,$$

    sendo μ_1 a resistência esperada da primeira variedade; $\alpha_i = \mu_i - \mu_1$ o efeito (acréscimo à resistência média da primeira variedade) da variedade i (com $\alpha_1 = 0$); e ϵ_{ij} o erro aleatório da observação Y_{ij}. Iremos (tal como o programa `R`) admitir que as variedades estão ordenadas por ordem alfabética, com os nomes de nível numéricos à cabeça, pelo que a primeira variedade acima referida é a variedade *18*.

    ii. Admite-se que os erros aleatórios são todos Normais, de média zero e variâncias homogéneas, ou seja, para qualquer i, j:

    $$\epsilon_{ij} \cap \mathcal{N}(0, \sigma^2) \ .$$

    iii. Admite-se que os erros aleatórios ϵ_{ij} são independentes.

    (b) A tabela-resumo terá apenas duas linhas (além da linha correspondente aos Totais), associadas respectivamente aos efeitos do Factor e à variabilidade Residual.

    i. Sabemos que os graus de liberdade dos efeitos do factor são $k - 1 = 5$ e que os graus de liberdade residuais são $n - k = 18 - 6 = 12$. As fórmulas para as Somas de Quadrados

são dadas no formulário. A Soma de Quadrados Residual é $SQRE = \sum_{i=1}^{k}(n_i-1)s_i^2$ e, tratando-se dum delineamento equilibrado com $n_c = 3$ repetições em cada nível, tem-se $SQRE = (n_c-1)\sum_{i=1}^{k} s_i^2$. Usando as variâncias amostrais de nível dadas no enunciado, vem $SQRE = 2 \times (14713.08 + 367.9434 + 5881.921 + 33132.64 + 5.414433 + 47.11163) = 108\,296.2$. É possível calcular SQF através da sua fórmula, uma vez que são disponibilizadas as médias amostrais de nível e globais. Mas é mais simples obter esse valor constatando que, numa ANOVA a um factor, se tem $SQF = SQT - SQRE$. No nosso caso $SQT = (n-1)s_y^2 = 17 \times 34\,517.82 = 586\,802.9$. Logo, $SQF = 478\,506.7$. Dividindo estas Somas de Quadrados pelos graus de liberdade antes referidos obtêm-se os Quadrados Médios, e dividindo QMF por $QMRE$ obtem-se o valor calculado da estatística do teste F aos efeitos do factor. Eis a tabela-resumo:

	g.l.	SQs	Quadrados Médios	F_{calc}
Factor	5	478 506.7	$\frac{478\,506.7}{5} = 95\,701.35$	$F_{calc} = \frac{QMF}{QMRE} = \frac{95\,701.35}{9\,024.685} = 10.6044$
Residual	12	108 296.2	$\frac{108\,296.2}{12} = 9\,024.685$	–

ii. Usando o `R`, confirmamos a tabela-resumo agora obtida:

```
> tomate.aov <- aov(res.pel ~ variedade , data=tomate)
> summary(tomate.aov)
            Df Sum Sq Mean Sq F value   Pr(>F)
variedade    5 478507   95701    10.6 0.000448
Residuals   12 108296    9025
```

(c) Eis o teste aos efeitos do factor (variedade):

Hipóteses: $H_0 : \alpha_i = 0\ ,\ \forall\ i$ vs. $H_1 : \exists\, i$ tal que $\alpha_i \neq 0$.

Estatística do Teste: $F = \frac{QMF}{QMRE} \cap F_{[k-1,n-k]}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(5,12)} = 3.11$.

Conclusões: Como $F_{calc} = 10.6044 > 3.11$, rejeita-se H_0, concluindo-se que existem de efeitos de variedade (ao nível $\alpha = 0.05$), o que corresponde a afirmar que existem variedades de tomate cujas películas têm resistência média diferentes de outras.

(d) O valor de prova (*p-value*) associado ao valor calculado da estatística de teste é $p = 0.000448$. Pela própria definição de *p-value*, esta é a área à direita de $F_{calc} = 10.6044$, numa distribuição $F_{[5,12]}$. Logo, seria preciso fazer um teste de hipóteses com nível de significância $\alpha = 0.000448$ (ou inferior) para que F_{calc} não pertencesse à Região Crítica e a conclusão do teste pudesse ser a de não rejeitar H_0.

(e) Tal como nas regressões lineares, a primeira coluna da matriz $\mathbf{X}$ é uma coluna de uns. No contexto duma ANOVA a um factor, as restantes colunas são variáveis indicatrizes de pertença de cada observação a um dos níveis do factor, ou seja, colunas com apenas dois valores: "1" associado a observações que pertencem ao nível do factor em causa, e "0" associado a observações associadas a outros níveis do factor. A restrição imposta no modelo ($\alpha_1 = 0$) implica que não há indicatriz do primeiro nível do factor, neste caso, o nível "`18`". Assim, neste caso teremos uma primeira coluna de $n = 18$ uns e cinco colunas indicatrizes dos segundo, terceiro, quarto, quinto e sexto níveis do factor ($\mathcal{I}_2$, $\mathcal{I}_3$, $\mathcal{I}_4$, $\mathcal{I}_5$ e $\mathcal{I}_6$), como se pode confirmar através do comando referido no enunciado:

```
> model.matrix(tomate.aov)
   (Intercept) variedade28 variedade29 variedade40C variedadeAce variedadeRoma
1            1           0           0            0            0             0
```

2	1	0	0	0	0	0
3	1	0	0	0	0	0
4	1	1	0	0	0	0
5	1	1	0	0	0	0
6	1	1	0	0	0	0
7	1	0	1	0	0	0
8	1	0	1	0	0	0
9	1	0	1	0	0	0
10	1	0	0	1	0	0
11	1	0	0	1	0	0
12	1	0	0	1	0	0
13	1	0	0	0	0	1
14	1	0	0	0	0	1
15	1	0	0	0	0	1
16	1	0	0	0	1	0
17	1	0	0	0	1	0
18	1	0	0	0	1	0

A ordem dos níveis do factor no `R` é, por omissão, a ordem alfabética dos nomes dos níveis. Mas essa pode não ser a ordem pela qual as observações surgem nas linhas da *data frame* com os dados. Neste exemplo, a variedade `Roma` surge como último nível (última coluna de $\mathbf{X}$), mas as observações dessa variedade não estão nas linhas finais da *data frame*, razão pela qual as duas colunas finais de $\mathbf{X}$ parecem 'trocadas'.

(f) Os valores ajustados $\hat{Y}_{ij}$, numa ANOVA a um factor, são as médias amostrais do nível a que cada observação pertence. Assim, tem-se:

```
> fitted(tomate.aov)
       1        2        3        4        5        6        7        8
560.6433 560.6433 560.6433 241.4833 241.4833 241.4833 290.9500 290.9500
       9       10       11       12       13       14       15       16
290.9500 705.7800 705.7800 705.7800 332.1067 332.1067 332.1067 377.2533
      17       18
377.2533 377.2533
```

Estas são as médias de variedade dadas no enunciado (arredondadas a uma casa decimal).

(g) O facto dos resíduos se encontrarem 'empilhados' em seis colunas é o reflexo natural do facto, referido na alínea anterior, de apenas haver seis diferentes valores ajustados nesta ANOVA: as seis médias amostrais de cada variedade, $\hat{y}_{ij} = \overline{y}_{i\cdot}$ $(j = 1, 2, 3)$. Este facto ajuda a identificar as observações associadas aos resíduos de maior magnitude. Assim, por exemplo, o maior resíduo (em módulo) corresponde ao ponto no canto inferior direito. Por estar associado a uma média $\overline{y}_{i\cdot}$ de aproximadamente 700, tem de corresponder à variedade `40C`. Por ser um resíduo negativo, tem de corresponder a uma observação com valor inferior à média dessa variedade, o que apenas acontece com a primeira das três observações desse nível. Assim, a observação a que corresponde o referido resíduo é a observação $y_{4,1} = 503.51$.

Embora o número de repetições em cada nível ($n_c = 3$) seja muito baixo, e portanto susceptível de gerar impressões enganadoras, o gráfico sugere alguma heterogeneidade nas variâncias de Y_{ij} em cada nível. Os valores das variâncias amostrais de nível indicam que há variedades com muito pouca variabilidade nas resistências observadas (como a *Ace*, com $s_5^2 = 5.414433$) e outras com uma variabilidade muito maior (como a *29*, com $s_3^2 = 5881.921$, mais de cem vezes maior).

2. Neste exercício sobre os estomas das folhas de café, não estão disponíveis os dados originais. Apenas se conhece a tabela dos valores médios e variâncias amostrais de cada variedade.

   (a) A variável resposta Y é o comprimento médio dos estomas das folhas duma planta. Para explicar a variabilidade dos valores desta variável, apenas se dispõe de um factor: o factor variedade, com $k=3$ níveis (as três variedades indicadas no enunciado). O modelo ANOVA é assim o modelo a um factor, semelhante ao do primeiro exercício. É um delineamento equilibrado, pois existem $n_i=12$ observações para qualquer variedade ($i=1,2,3$), perfazendo um total de $n=3\times 12=36$ observações Y_{ij}. Eis o modelo:

      i. $Y_{ij}=\mu_1+\alpha_i+\epsilon_{ij},\ \forall\, i=1,2,3,\ ,\ j=1,2,...,12\ ,\quad$ com $\alpha_1=0$, onde
         - Y_{ij} indica o comprimento médio dos estomas das folhas da planta j da variedade i;
         - μ_1 indica o comprimento médio populacional dos estomas das folhas de plantas da primeira variedade ($i=1$) que é, por ordem alfabética, a variedade `CA`;
         - α_i indica o efeito (acréscimo em relação à média da variedade `CA`) da variedade i; e
         - ϵ_{ij} indica o erro aleatório associado à observação Y_{ij}.

      ii. $\epsilon_{ij}\cap\mathcal{N}(0,\sigma^2),\ \forall\, i,j.$

      iii. $\{\epsilon_{ij}\}_{i,j}$ constitui um conjunto de variáveis aleatórias independentes.

   (b) Comecemos pelo cálculo das Somas de Quadrados. Uma vez que o delineamento é equlibrado (igual número de observações em cada nível), a média global da totalidade das 36 observações ($\overline{y}_{..}$) é a média simples das três médias de nível dadas na tabela: $\overline{y}_{..}=(22.85833+19.49333+25.31583)/3=22.55583$. Tendo em conta as fórmulas vistas nas aulas teóricas e os valores dados no enunciado, temos:

$$\begin{aligned} SQRE &= (n_c-1)\sum_{i=1}^{3}s_i^2 = 11\times(13.69303+2.725424+9.388936) = 284.1983\ ; \\ SQF &= n_c\sum_{i=1}^{3}(\overline{y}_{i.}-\overline{y}_{..})^2 \\ &= 12\times((22.85833-22.55583)^2+(19.49333-22.55583)^2+(25.31583-22.55583)^2) \\ &= 205.0561, \end{aligned}$$

   Logo, tem-se a seguinte tabela-resumo:

Fonte	g.l.	SQ	QM	F_{calc}
Factor	$k-1=2$	$SQF=205.0561$	$QMF=\frac{SQF}{k-1}=102.5281$	$\frac{QMF}{QMRE}=11.90516$
Resíduos	$n-k=33$	$SQRE=284.1983$	$QMRE=\frac{SQRE}{n-k}=8.61207$	

   (c) Neste caso, e uma vez que não são conhecidas as observações individuais, apenas é possível calcular a variância da totalidade das $n=36$ observações recorrendo à decomposição da Soma de Quadrados Total correspondente a esta ANOVA:

$$s_y^2 = \frac{SQT}{n-1} = \frac{SQF+SQRE}{n-1} = \frac{205.0561+284.1983}{35} = \frac{489.2544}{35} = 13.9787\ .$$

   Repare-se que este valor *não é* a média das variâncias amostrais de nível.

(d) Embora se possa escrever as hipóteses do teste com base nos efeitos α_i do factor (como se fez no exercício anterior), nas ANOVAs a um único factor é equivalente formular as hipóteses em termos das médias populacionais (valores esperados das observações $E[Y_{ij}] = \mu_i = \mu_1 + \alpha_i$) em cada nível do factor. Eis o teste com $\alpha = 0.05$:

Hipóteses: $H_0 : \mu_1 = \mu_2 = \mu_3$ vs. $H_1 : \exists\, i, i'$ tal que $\mu_i \neq \mu_{i'}$.

Estatística do teste: $F = \frac{QMF}{QMRE} \cap F_{(k-1\,,\,n-k)}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(2,33)} \approx 3.30$ (entre os valores tabelados 3.23 e 3.32).

Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 11.90516$. É um valor significativo ao nível $\alpha = 0.05$ e rejeita-se H_0 a favor da hipótese de que existem efeitos do factor, ou seja, de que o comprimento médio dos estomas das folhas não é igual em todas as variedades.

O valor de prova associado à estatística calculada é (tendo em conta a natureza unilateral direita do teste) $P[F_{(2,33)} > F_{calc}] = P[F_{(2,33)} > 11.90516]$. Não é possível obter este valor nas tabelas, mas pode calcular-se essa probabilidade com o auxílio do **R**:

```
> 1-pf(11.90516, 2,33)
[1] 0.000128065
```

Assim, tem-se $p = 0.000128065$.

(e) **[Material Complementar]** Sabemos que duas médias de nível μ_i e $\mu_{i'}$ devem ser consideradas diferentes caso as respectivas médias amostrais difiram (em módulo) mais do que o termo de comparação $q_{\alpha(k,n-k)}\sqrt{\frac{QMRE}{n_c}}$, onde $q_{\alpha(k,n-k)}$ corresponde ao valor que deixa à sua direita uma região de probabilidade α numa distribuição de Tukey de parâmetros k e $n-k$, e n_c indica o número comum de observações em cada nível do factor (o resultado que sustenta o teste de Tukey parte do pressuposto que o delineamento é equilibrado). No nosso caso tem-se $k = 3$ e $n = 36$. Trabalhando (como pedido no enunciado) com $\alpha = 0.05$, e recorrendo às tabelas da distribuição de Tukey (tabelas específicas, disponíveis na página *web* da disciplina), tem-se $q_{0.05(3,33)} = 3.47$. Um valor mais preciso pode ser obtido através do comando `qtukey` do **R**:

```
> qtukey(0.95, 3, 33)
[1] 3.470189
```

Sabemos pela alínea (b) que $QMRE = 8.61207$ e também que $n_c = 12$. Logo, o termo de comparação é dado por $q_{\alpha(k,n-k)}\sqrt{\frac{QMRE}{n_c}} = 3.470189 \times \sqrt{\frac{8.61207}{12}} = 2.490459$. Calculando as diferenças entre as médias amostrais de cada variedade, obtém-se a seguinte tabela:

$\lvert\overline{y}_{i\cdot} - \overline{y}_{i'\cdot}\rvert$	CA ($i'=1$)	CL ($i'=2$)	PR ($i'=3$)
CA ($i=1$)	–	3.3650	2.4575
CL ($i=2$)	3.3650	–	5.8225
PR ($i=3$)	2.4575	5.8225	–

Assim, ao nível de significância $\alpha = 0.05$, o comprimento médio dos estomas de folhas da variedade `CL` é diferente, quer do comprimento médio da variedade `CA`, quer do comprimento médio da variedade `PR`. No entanto, não se pode considerar (por pouco) significativamente diferentes os comprimentos médios dos estomas de folhas das variedades `CA` e `PR`.

Existem duas formas frequentes de representar esta conclusão, sendo usual em ambas ordenar os níveis do factor por ordem crescente das respectivas médias, e:

i. sublinhando-se com traços os grupos de níveis cujas médias não diferem significativamente o que, nesta alínea (ao nível $\alpha = 0.05$) produz o seguinte resultado:

CL	CA	PR
19.49333	22.85833	25.31583

ii. ou colocando uma mesma letra ao lado das variedades cujas médias não se consideram significativamente diferentes, por exemplo:

CL	CA	PR
19.49333^a	22.85833^b	25.31583^b

Assim, a média de `CL` é significativamente diferente das médias, quer de `CA`, quer de `PR` (com quem não partilha letras em comum), mas já a média da variedade `CA` não difere significativamente da média de `PR` (uma vez que partilham a mesma letra).

3. A variável resposta Y é, neste caso, a variação de massa (coluna `variacao.massa` na *data frame*). Existem ao todo $n = 50$ observações.

(a) Para estudar este problema através duma ANOVA, ignora-se os valores numéricos das concentrações de dióxido de carbono, tratando cada diferente concentração apenas como um diferente tratamento. Assim, o factor CO_2 terá $k = 5$ níveis, havendo $(n_i = 10 = n_c)$ observações para cada concentração de CO_2 (nível do factor). O modelo ANOVA associado a este delineamento é o seguinte:

i. $Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij}, \ \forall\, i = 1, 2, 3, 4, 5 \ , \ j = 1, 2, ..., 10 \ , \quad$ com $\alpha_1 = 0$, onde
   - Y_{ij} indica a variação de massa para a j-ésima repetição associada à i-ésima concentração de CO_2;
   - μ_1 indica o variação de massa média (populacional) na ausência de CO_2 $(i = 1)$;
   - α_i indica o efeito (acréscimo em relação à média populacional do primeiro nível) da i-ésima concentração de dióxido de carbono, isto é, $\alpha_i = \mu_i - \mu_1$; e
   - ϵ_{ij} indica o erro aleatório associado à observação Y_{ij}.

ii. $\epsilon_{ij} \cap \mathcal{N}(0, \sigma^2), \ \forall\, i, j.$

iii. $\{\epsilon_{ij}\}_{i,j}$ constitui um conjunto de variáveis aleatórias independentes.

(b) Vamos construir a tabela-resumo da ANOVA com o auxílio do R, uma vez que os dados estão disponíveis na *data frame* `CO2`, com os valores da variável resposta na coluna `variacao.massa` e os diferentes níveis de CO_2 no factor `CO2.factor` (alternativamente, podem sempre usar-se as fórmulas disponíveis no formulário para SQF e $SQRE$ em delineamentos a um factor, sabendo-se também que os graus de liberdade associados ao Factor são $k - 1 = 4$ e os residuais $n - k = 45$):

```
> summary(aov(variacao.massa ~ CO2.factor, data=CO2))
Df Sum Sq Mean Sq F value Pr(>F)
CO2.factor   4  11274  2818.6   101.6 <2e-16 ***
Residuals   45   1248    27.7
```

O teste F desta ANOVA diz respeito à possível existência de efeitos do Factor, ou seja,

Hipóteses: $H_0 : \alpha_i = 0 \ , \ \forall \ i = 2, 3, 4, 5$ vs. $H_1 : \exists\, i = 2, 3, 4, 5$ tal que $\alpha_i \neq 0$.

Estatística do teste: $F = \frac{QMF}{QMRE} \ \cap \ F_{(k-1\,,\,n-k)}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(4,45)} \ \approx \ 2.58$.

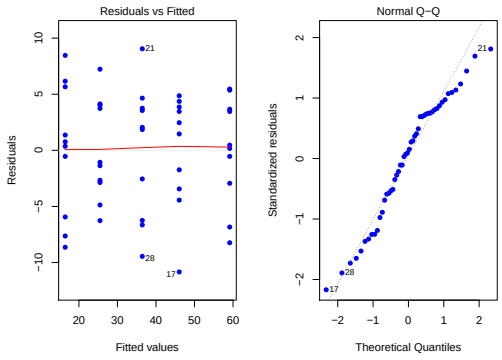
Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 101.6$. É um valor claramente significativo ao nível $\alpha = 0.05$ e rejeita-se H_0 a favor da hipótese de que existem efeitos do Factor, ou seja, que as concentrações de CO_2 estão associadas a diferentes variações médias na massa das culturas do *Pseudomonas fragi*.

(c) Como em qualquer modelo linear, o resíduo é a diferença entre cada valor observado da variável resposta e o correspondente valor ajustado pelo modelo, ou seja, e usando a notação da ANOVA a 1 Factor, $e_{ij} = y_{ij} - \hat{y}_{ij}$. Sabe-se que, num modelo ANOVA a um factor, o valor ajustado duma dada observação corresponde à média amostral das observações no mesmo nível do factor: $\hat{y}_{ij} = \overline{y}_{i.}$. Assim, todas as observações do primeiro grupo têm valor ajustado igual a $\hat{y}_{1j} = \overline{y}_{1.} = 59.14$. O resíduo da primeira observação do primeiro grupo será $e_{11} = 62.6 - 59.14 = 3.46$ e o da segunda observação desse grupo é $e_{12} = 59.6 - 59.14 = 0.46$. De forma análoga, os valores ajustados de qualquer observação no segundo grupo são dados por $\hat{y}_{2j} = \overline{y}_{2.} = 46.04$. O resíduo da terceira observação do segundo grupo é assim $e_{23} = y_{23} - \overline{y}_{2.} = 47.5 - 46.04 = 1.46$. Para calcular a totalidade dos resíduos podemos recorrer ao R (arredondando a três casas decimais):

```
> round(residuals(CO2.aov), d=3)
    1     2     3     4     5     6     7     8     9    10    11    12    13
 3.46  0.46  5.36  0.16 -0.54  5.46 -8.24 -2.94 -6.84  3.66  4.86 -1.74  1.46
   14    15    16    17    18    19    20    21    22    23    24    25    26
 3.46  2.46  4.36 -10.84  3.86 -3.44 -4.44  9.05  4.65 -6.65  1.85  3.75  2.05
   27    28    29    30    31    32    33    34    35    36    37    38    39
-6.25 -9.45  3.55 -2.55  4.03 -2.67 -6.27 -4.87  3.73 -1.37 -2.87  7.23 -1.07
   40    41    42    43    44    45    46    47    48    49    50
 4.13  8.46  0.76 -8.64 -5.94  1.36  5.66  6.16  0.36 -0.54 -7.64
```

Com o auxílio do R, podemos obter os dois gráficos de resíduos já considerados no estudo dos modelos de Regressão Linear, através do comando:

```
> plot(CO2.aov, which=c(1,2), pch=16, col="blue")
```



O gráfico da esquerda é o gráfico de resíduos usuais (no eixo vertical) vs. valores ajustados da variável resposta (eixo horizontal). O facto de os resíduos surgirem "empilhados" em colunas é característico numa ANOVA a um factor e resulta do já referido facto de todas as observações dum dado nível terem o mesmo valor ajustado $\hat{y}_{ij} = \overline{y}_{i.}$, logo, a mesma coordenada no eixo horizontal. Neste caso, observam-se $k = 5$ colunas. Não parece existir problema com a hipótese de homogeneidade das variâncias, uma vez que a variabilidade dos resíduos não parece diferir muito nos cinco níveis do factor. O *qq-plot* (gráfico à direita) não

indicia problemas graves com a Normalidade, dada a disposição aproximadamente linear dos pontos.

Os restantes diagnósticos que foram considerados aquando do estudo da regressão (distâncias de Cook, efeito alavanca) são geralmente de menor utilidade no contexto duma ANOVA. Em relação às distâncias de Cook, por exemplo, sabe-se de antemão qual o efeito de retirar uma observação: além de desequilibrar um delineamento equilibrado, afectará a média das observações no mesmo nível do factor (ou seja, os valores ajustados $\hat{y}$ nesse nível). Assim valores elevados da distância de Cook correspondem a observações atípicas (*outliers*) no seio dum dado nível. Mas para identificar tais observações, basta o gráfico usual de resíduos contra $\hat{y}$, não sendo necessário um diagnóstico específico. Em relação aos efeitos alavanca, é possível mostrar que o efeito alavanca de qualquer observação y_{ij} numa ANOVA a um factor é dada por $\frac{1}{n_i}$, onde n_i indica o número de observações no nível i da observação. Em delineamentos equilibrados, esse valor é igual para todas as observações (no nosso caso, todas teriam efeito alavanca igual a $\frac{1}{10}$). O gráfico obtido no `R` com a opção `which=5` tinha, na regressão linear, os valores do efeito alavanca (h_{ii}, ou *leverages*) de cada observação no eixo horizontal. No entanto, para ANOVAs com delineamentos equilibrados a um factor, o R substitui esse eixo por uma simples indicação dos diferentes níveis do factor (ordenados por ordem crescente das médias $\overline{y}_{i.}$), uma vez que um gráfico análogo ao construído na regressão linear apenas empilharia todos os resíduos numa única coluna. O gráfico alternativo produzido pelo R quando os delineamentos são equilibrados fica assim semelhante ao primeiro gráfico de resíduos, embora sem qualquer efeito de escala no eixo horizontal e com os resíduos (internamente) estandardizados no eixo vertical, em vez dos resíduos usuais.

(d) Nesta alínea pede-se para aproveitar os valores das concentrações de CO_2 utilizadas, e tratar essa variável preditora como uma variável numérica, estudando a regressão linear simples de `variacao.massa` sobre `CO2.numerico`.

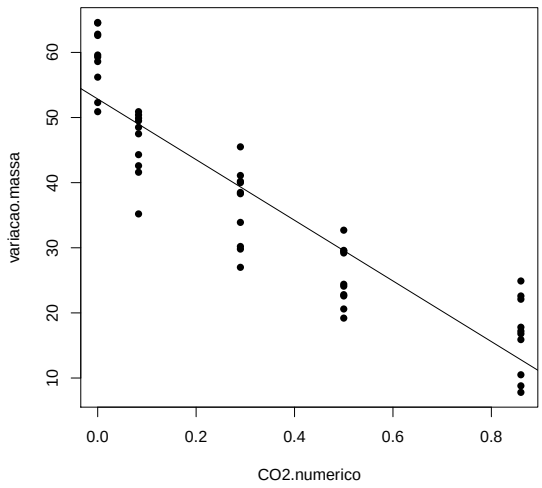
i. O gráfico pedido pode ser construído com o seguinte comando do `R`. O resultado é mostrado na alínea seguinte.

```
> plot(variacao.massa ~ CO2.numerico, data=CO2, pch=16)
```

ii. A regressão linear pedida é dada por:

```
> CO2.lm <- lm(variacao.massa ~ CO2.numerico, data=CO2)
> summary(CO2.lm)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    52.849      1.408   37.52   <2e-16 ***
CO2.numerico  -46.569      3.030  -15.37   <2e-16 ***
---
Residual standard error: 6.637 on 48 degrees of freedom
Multiple R-squared: 0.8312,Adjusted R-squared: 0.8276
F-statistic: 236.3 on 1 and 48 DF,  p-value: < 2.2e-16
```

A nuvem de pontos pedida na alínea anterior, já com a recta de regressão (traçada com o comando `abline(CO2.lm)`) é:



Apesar de alguma tendência para uma relação curvilinear, uma regressão linear simples pode constituir uma modelação aproximada da relação entre concentrações de dióxido de carbono e variação na massa das culturas de *Pseudomonas fragi* (repare-se como seria impossível tirar esta ilação se o número de níveis fosse mais pequeno, *e.g.*, $k = 3$). O valor do coeficiente de determinação é claramente significativo ($p < 2.2 \times 10^{-16}$) e bastante elevado ($R^2 = 0.8312$), explicando mais de 83% da variabilidade total observada na variável resposta.

iii. Os testes F de ajustamento global do contexto regressão linear simples e do contexto ANOVA a um factor, não são os mesmos. Como se viu nas aulas teóricas, a ANOVA a um factor pode ser vista como uma espécie de regressão linear múltipla em que as variáveis preditoras são as indicatrizes dos níveis (excepto o primeiro) do factor. Assim, a informação disponível para prever os valores da variável resposta é, no caso da regressão considerada nesta alínea, a variável `CO2.numerico`, com valores numéricos diferentes em cada nível (mas repetidos para as observações dum mesmo nível). No caso da ANOVA a um factor, é o conjunto das indicatrizes de nível e o vector dos n uns. Sendo diferente a informação preditora, serão diferentes os valores ajustados e os valores dos respectivos F_{calc} e coeficientes de determinação. Em relação a este último, e embora não seja hábito utilizá-lo no contexto duma ANOVA a um factor, o seu valor é aqui $R^2 = 0.9003$, superior ao que se obteve na regressão ($R^2 = 0.8312$), como se pode constatar através do ajustamento obtido utilizando simultaneamente o comando `lm` e o factor preditor `CO2.factor`:

```
> summary(lm(variacao.massa ~ CO2.factor, data=CO2))
(...)
Residual standard error: 5.266 on 45 degrees of freedom
Multiple R-squared: 0.9003,Adjusted R-squared: 0.8915
F-statistic: 101.6 on 4 and 45 DF,  p-value: < 2.2e-16
```

Repare-se como o valor da estatística calculada, $F_{calc} = 101.6$, é o que foi obtido usando o comando `aov`.

Um comentário final: o modelo ANOVA não permite, ao contrário da regressão, fazer previsões sobre as variações de massa com concentrações de CO_2 não observadas na experiência, uma vez que os níveis do factor CO_2 não têm escala (são apenas categorias diferentes).

4. Trata-se dum delineamento factorial a dois factores (`terreno` e `variedade`), mas com uma única observação em cada célula (em cada terreno, apenas há uma parcela com cada variedade). Logo, só é possível ajustar um modelo a dois factores sem interacção.

   (a) A tabela-resumo correspondente é:

```
> terrenos.aov <- aov(rend ~ variedade + terreno, data=terrenos)
> summary(terrenos.aov)
            Df Sum Sq Mean Sq F value  Pr(>F)
variedade    3  1.799  0.5997   6.145 0.00175 **
terreno     12  2.407  0.2006   2.056 0.04737 *
Residuals   36  3.513  0.0976
```

   Desta tabela depreende-se que, aos níveis de significância usuais, deve considerar-se a existência de efeitos do factor variedade:

   Hipóteses: $H_0 : \alpha_i = 0 \ , \ \forall \ i = 2, 3, 4$ vs. $H_1 : \exists\, i = 2, 3, 4$ tal que $\alpha_i \neq 0$.
   Estatística do teste: $F = \frac{QMA}{QMRE} \ \cap \ F_{(a-1\,,\, n-(a+b-1))}$, sob H_0.
   Nível de significância: $\alpha = 0.05$.
   Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(3,36)} \ \approx \ 2.87$.
   Conclusões: $F_{calc} = 6.145$, um valor significativo mesmo ao nível $\alpha = 0.005$. Logo, rejeita-se H_0 a favor da hipótese de que existem efeitos do factor. Assim, é de concluir que diferentes variedades estejam associadas a diferentes rendimentos médios.

   (b) Um teste aos efeitos do factor `terreno` permite tirar a conclusão que os efeitos deste factor são menos importantes que os efeitos do factor `variedade`, embora ao nível de significância $\alpha = 0.05$ sejam (por pouco) significativos. Assim,

   Hipóteses: $H_0 : \beta_j = 0 \ , \ \forall \ j = 2, .., 13$ vs. $H_1 : \exists\, j = 2, ..., 13$ tal que $\beta_j \neq 0$.
   Estatística do teste: $F = \frac{QMB}{QMRE} \ \cap \ F_{(b-1\,,\, n-(a+b-1))}$, sob H_0.
   Nível de significância: $\alpha = 0.05$.
   Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(12,36)} \ \approx \ 2.04$.
   Conclusões: $F_{calc} = 2.056$, um valor significativo (por muito pouco) ao nível $\alpha = 0.05$. Logo, rejeita-se H_0 a favor da hipótese de que existem efeitos do factor `terreno`.

   NOTA: Num caso como este, em que a conclusão dependende do nível de significância usado, é especialmente importante que eventuais fontes de variabilidade, exteriores ao factor sob estudo, mas que afectem a variável resposta, sejam tidas em conta, de forma a reduzir a variabilidade não explicada pelo modelo, isto é, o valor de $QMRE$.

5. Trata-se dum delineamento factorial a dois factores, o factor A (Fósforo), com $a = 3$ níveis (Baixa, Média e Elevada dosagem de adubação) e o Factor B (Potássio), igualmente com $b = 3$ níveis (Baixa, Média e Elevada dosagem de adubação). O delineamento é equilibrado, uma vez que em cada uma das $ab = 9$ situações experimentais (células) há igual número de observações $n_{ij} = n_c = 3$. Havendo repetições nas células, é possível estudar o modelo ANOVA a 2 factores, com interacção. A equação de base deste modelo é $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \ \forall\, i = 1, 2, 3 \ , \ j = 1, 2, 3, \ k = 1, 2, 3$, onde Y_{ijk} indica o rendimento obtido na k-ésima repetição da adubação correspondente à célula que cruza o nível i do fósforo e o nível j do potássio. Impõem-se as restrições $\alpha_1 = 0$, $\beta_1 = 0$, $(\alpha\beta)_{1j} = 0$ para qualquer j, e $(\alpha\beta)_{i1} = 0$ para qualquer i.

   (a) A tabela-resumo é dada no enunciado, mas com seis valores omissos. Os graus de liberdade do factor A (fósforo) são $a{-}1 = 2$. Os graus de liberdade associados aos efeitos de interacção

são $(a-1)(b-1) = 4$. O Quadrado Médio associado ao factor B (potássio) é $QMB = \frac{SQB}{b-1} = \frac{18.7563}{2} = 9.37815$. O Quadrado Médio Residual é $QMRE = \frac{SQRE}{n-ab} = \frac{2.59333}{18} = 0.1440739$. O valor da estatística F para o teste aos efeitos principais do factor A é $F_A = \frac{QMA}{QMRE} = \frac{1.121481}{0.1440739} = 7.784068$. Finalmente, o valor da estatística F no teste aos efeitos principais do factor B é $F_B = \frac{QMB}{QMRE} = \frac{9.37815}{0.1440739} = 65.09264$.

(b) Há três tipos de efeitos: principais do factor fósforo, associados às parcelas α_i; principais do factor potássio, associados às parcelas β_j; e de interacção entre os dois tipos de adubação, associados às parcelas $(\alpha\beta)_{ij}$. Existe um teste F para testar hipóteses associadas a cada um destes tipos de efeitos. Em concreto:

Teste à interacção. As hipóteses são:

$$H_0 : \quad (\alpha\beta)_{ij} = 0 \ , \ \forall\, i,j \qquad vs. \qquad H_1 : \quad \exists\, i,j \text{ tal que } \quad (\alpha\beta)_{ij} \neq 0 \ .$$

Teste aos efeitos principais do factor A. As hipóteses são:

$$H_0 : \quad \alpha_i = 0 \ , \ \forall\, i \qquad vs. \qquad H_1 : \quad \exists\, i \text{ tal que } \quad \alpha_i \neq 0 \ .$$

Teste aos efeitos principais do factor B. As hipóteses são:

$$H_0 : \quad \beta_j = 0 \ , \ \forall\, j \qquad vs. \qquad H_1 : \quad \exists\, j \text{ tal que } \quad \beta_j \neq 0 \ .$$

Para cada um destes testes, as estatísticas F são definidas como $F = \frac{QMxx}{QMRE}$, onde $QMxx$ indica o quadrado médio associado ao respectivo tipo de efeitos. As distribuições destas estatísticas de teste, caso seja verdadeira cada uma das hipóteses nulas, são F com graus de liberdade dados pelos g.l. dos quadrados médios no numerador e denominador, respectivamente, da estatística correspondente. Todas as regiões críticas são unilaterais direitas. Assim, e tendo em conta os valores da tabela-resumo e utilizando o nível de significância $\alpha = 0.05$, tem-se que se rejeitam as hipóteses nulas dos três testes. De facto, rejeita-se a inexistência de efeitos de interacção, uma vez que $F_{AB_{calc}} = 3.36504 > f_{0.05(4,18)} = 2.927744$. Rejeita-se a inexistência de efeitos principais do factor fósforo uma vez que $F_{A_{calc}} = 7.784068 > f_{0.05(2,18)} = 3.554557$. Finalmente, rejeita-se clarissimamente a inexistência de efeitos principais do factor potássio já que $F_{B_{calc}} = 65.09264 > f_{0.05(2,18)} = 3.554557$. Assim, conclui-se pela existência dos três tipos de efeitos. Estas conclusões poderiam também ser obtidas directamente a partir dos valores de prova (*p-values*) correspondentes às três estatísticas de teste, disponíveis no enunciado. O valor de prova mais elevado, no caso do teste aos efeitos de interacção ($p = 0.03187154$) indica que, ao nível de significância $\alpha = 0.01$, a conclusão já seria a não rejeição da hipótese nula, isto é, não seria possível concluir pela existência de efeitos de interacção. Já a existência de efeitos principais do factor potássio está associado a um *p-value* da ordem de 10^{-8}.

(c) Nesta alínea pede-se para considerar-se o modelo sem efeitos de interacção, ou seja, cuja equação de base é $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \epsilon_{ijk}$, $\forall\, i,j,k$, e com as restrições $\alpha_1 = \beta_1 = 0$. O facto de o modelo não prever efeitos de interacção significa que a respectiva Soma de Quadrados (indicada no enunciado) passa a englobar a Soma de Quadrados Residual (uma vez que já não corresponde a efeitos previstos pelo modelo). Tem-se agora $SQRE = 2.59333 + 1.93926 = 4.53259$. Os graus de liberdade sofrem uma transformação análoga (este modelo tem agora menos $(a-1)(b-1)$ parâmetros do que anterior, pelo que os graus de liberdade residuais aumentam nesse montante). Assim, $g.l.(SQRE) = 18 + 4 = 22$. Logo o novo Quadrado

Médio Residual vem: $QMRE = \frac{4.53259}{22} = 0.2060268$. As somas de quadrados, graus de liberdade e quadrados médios associados aos efeitos principais de cada factor permanecem iguais (são calculados de forma análoga) pelo que a tabela-resumo é agora a seguinte:

variação	g.l.	SQs	QMs	F_{calc}
fosforo	2	2.24296	1.121481	5.443374
potassio	2	18.75630	9.37815	45.51908
residual	22	4.53259	0.2060268	–

Para identificar os valores de prova (*p-values*) dos novos valores das estatísticas F sobrantes, é necessário ter em conta os novos valores dos graus de liberdade residuais. Tem-se:

```
> 1-pf(5.443374, 2, 22)
[1] 0.01200658
> 1-pf(45.51908, 2, 22)
[1] 1.517658e-08
```

Assim, os dois valores calculados das estatísticas continuam a ser significativos ao nível $\alpha = 0.05$. No entanto, os efeitos do factor fósforo já não seriam considerados significativos ao nível $\alpha = 0.01$. Este exemplo ilustra o perigo de ignorar a existência de efeitos que realmente existam (neste caso, ignorar os efeitos de interacção): pode ajudar a camuflar a existência de outros tipos de efeitos, mesmo dos que são previstos no modelo, através do inflacionamento da variabilidade residual ($QMRE$).

6. (a) Trata-se dum delineamento factorial a dois factores, sendo a variável resposta Y a altura aos dois anos (em cm) dos pinheiros; o primeiro factor (A) a proveniência, com $a = 5$ níveis e o segundo factor (B) o local do ensaio (com $b = 2$ níveis). O delineamento é equilibrado, uma vez que em cada uma das $ab = 10$ células (situações experimentais) existem $n_c = 6$ observações, num total de $n = n_c ab = 60$ observações. Existem repetições nas células, logo é possível (e desejável) estudar a existência de eventuais efeitos de interacção.

   O modelo ajustado é o modelo ANOVA a dois factores, com efeitos de interacção. Admite-se que os níveis de cada factor estão ordenados por ordem alfabética (que corresponde à ordem em que aparecem no enunciado). Eis o modelo:

   - $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, para qualquer $i = 1, 2, 3, 4, 5$, $j = 1, 2$ e $k = 1, 2, 3, 4, 5, 6$, sendo μ_{11} a altura esperada (aos dois anos) dos pinheiros gregos em Sines; α_i o efeito principal (acréscimo à altura) associado à proveniência i (com a restrição $\alpha_1 = 0$); β_j o efeito principal (acréscimo à altura) associado a $j = 2$ (dada a restrição $\beta_1 = 0$); $(\alpha\beta)_{ij}$ o efeito de interacção, isto é, o acréscimo na altura específico da combinação da proveniência i com o local j. Dadas as restrições $(\alpha\beta)_{ij} = 0$ se $i = 1$ e/ou $j = 1$, o modelo apenas prevê efeitos de interacção nas situações experimentais correspondentes a Tavira ($j = 2$) e para proveniências diferentes da Grécia ($i > 1$). Finalmente ϵ_{ijk} é o erro aleatório da observação Y_{ijk}.
   - Admite-se que os erros aleatórios são todos Normais, de média zero e variâncias homogéneas: $\epsilon_{ijk} \cap \mathcal{N}(0, \sigma^2)$, para qualquer i, j, k.
   - Admite-se que os erros aleatórios ϵ_{ijk} são independentes.

   (b) Tratando-se dum modelo ANOVA factorial, a dois factores com interacção, a tabela-resumo terá de ter quatro linhas, correspondentes aos três tipos de efeitos previstos (principal de cada factor e de interacção), bem como à variabilidade residual e, opcionalmente, uma quinta linha associada à variabilidade total. A tabela terá as habituais colunas de graus

de liberdade, Somas de Quadrados, Quadrados Médios e valor das estatísticas F. Vejamos como se pode preencher esta tabela.

Sabemos que, neste tipo de modelo, os graus de liberdade associados a $QMRE$ são dados por $n-ab$, onde $n=60$ é o número total de observações e $ab=10$ é o número de parâmetros existentes no modelo. Assim, $g.l.(SQRE)=50$. Sabemos ainda que, para os vários tipos de efeitos, os graus de liberdade são dados pelo número de parcelas de cada tipo de efeito, após a introdução das restrições, ou seja, associado a SQA há $a-1=4$ g.l., associado a SQB há $b-1=1$ g.l., e associado a $SQAB$ há $(a-1)(b-1)=4$ graus de liberdade.

No enunciado é dada a Soma de Quadrados associada ao que foi designado factor A, tendo-se $SQA = 280.61$, donde se conclui que $QMA = \frac{SQA}{a-1} = \frac{280.61}{4} = 70.1525$. No enunciado é também dado o Quadrado Médio Residual, tendo-se $QMRE = 16.59$, donde $SQRE = QMRE \times (n - ab) = 16.59 \times 50 = 829.50$. Ora, sabemos pelo formulário que:

$$\begin{aligned} SQB &= a\, n_c \sum_{j=1}^{2} (\overline{y}_{.j.} - \overline{y}_{...})^2 \\ &= 5 \times 6 \times [(28.14 - 31.76298)^2 + (35.38 - 31.76298)^2] = 786.2645 \ . \end{aligned}$$

Donde $QMB = \frac{SQB}{b-1} = 786.2645$. O enunciado refere ainda a variância da totalidade das 60 observações, $s_y^2 = 34.49584$, donde se pode concluir que a Soma de Quadrados Total é $SQT = (n-1)\, s_y^2 = 59 \times 34.49584 = 2035.255$. Uma vez que sabemos que esta Soma de Quadrados Total se pode decompor como $SQT = SQA + SQB + SQAB + SQRE$, torna-se possível calcular $SQAB = SQT - (SQA + SQB + SQRE) = 2035.255 - (280.61 + 786.2645 + 829.50) = 138.8801$. Assim, o Quadrado Médio associado à interacção é dado por $QMAB = \frac{SQAB}{(a-1)(b-1)} = \frac{138.8801}{4} = 34.7200$.

Finalmente, os valores das estatísticas F são dados, para os três tipos de efeitos, pela razão entre o Quadrado Médio do referido tipo de efeito e $QMRE$. A tabela completa fica assim:

	g.l.	Soma de Quadrados	Quadrado Médio	F
Proveniência	4	280.61	70.1525	4.229
Local	1	786.2645	786.2645	47.394
Interacção	4	138.8801	34.7200	2.093
Residual	50	829.50	16.59	–

(c) Vai-se efectuar em pormenor o teste aos efeitos principais do Factor A (proveniência dos pinheiros), e descrever sinteticamente os testes aos efeitos principais do Factor B (local) e aos efeitos de interacção.

Hipóteses: $H_0 : \alpha_i = 0 \ , \ \forall\, i$ vs. $H_1 : \exists\, i$ tal que $\alpha_i \neq 0$.

Estatística do Teste: $F_A = \frac{QMA}{QMRE} \cap F_{[a-1,n-ab]}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(4,50)} \approx 2.57$ (entre os valores tabelados 2.53 e 2.61).

Conclusões: Como $F_{calc} = \frac{QMA}{QMRE} = 4.229 > 2.57$, rejeita-se H_0, sendo possível concluir pela existência de efeitos principais de proveniência (ao nível $\alpha = 0.05$).

No teste aos efeitos principais do factor local do estudo, as hipóteses do teste podem ser escritas apenas como $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$, uma vez que após a imposição da restrição $\beta_1 = 0$, apenas sobra um efeito deste tipo, o efeito β_2 associado a Tavira. O valor calculado

da estatística de teste é muito grande ($F_{calc} = 47.394$) deixando antever a rejeição de H_0, facto que é confirmado determinando nas tabelas o limiar da região crítica unilateral direita: $f_{0.05(1,50)} \approx 4.04$ (entre os valores tabelados 4.00 e 4.08). Assim, conclui-se claramente pela existência de efeitos principais de localidade, o que neste caso significa que existe um efeito associado à passagem do local de plantação de Sines para Tavira. Uma rápida inspecção das médias de local sugere que se trata dum maior crescimento dos pinheiros em Tavira, pelo que se deduz que β_2 terá um valor positivo.

No teste aos efeitos de interacção, com hipóteses $H_0 : (\alpha\beta)_{ij} = 0$, para todo o i e j, contra a hipótese alternativa de que existe pelo menos uma célula (i, j) onde $(\alpha\beta)_{ij} \neq 0$, o valor calculado da estatística de teste é $F_{calc} = 2.093$, inferior ao limiar da região crítica, que é (por coincidência) igual ao do teste aos efeitos do factor A, $f_{0.05(4,50)} \approx 2.57$. Logo, não se rejeita H_0 (para $\alpha = 0.05$), e conclui-se pela inexistência de efeitos significativos de interacção.

(d) **[Material Complementar]** Nesta alínea é pedido para verificar se o facto da maior altura média amostral de Sines (31.16, para pinheiros provenientes de Marrocos) ser menor que a mais baixa altura média amostral em Tavira (33.56, para pinheiros da segunda proveniência italiana) é uma relação que se possa estender à população. Vamos responder efectuando, como solicitado no enunciado, um teste de Tukey, e usando $\alpha = 0.05$. Ora, o termo de comparação é (como indicado no formulário e usando as tabelas da distribuição de Tukey):

$$q_{\alpha(ab,n-ab)} \sqrt{\frac{QMRE}{n_c}} = q_{0.05(10,50)} \sqrt{\frac{16.59}{6}} = 4.68 \times 1.662829 = 7.782039 \ .$$

Ora, a diferença entre as médias amostrais das duas células referidas acima é apenas $|31.16 - 33.56| = 2.40$, logo inferior ao termo de comparação, pelo que não é uma diferença significativa (ao nível $\alpha = 0.05$). Assim, não é possível afirmar que as médias populacionais em Tavira sejam sempre maiores às de Sines, independentemente das proveniências. Alguns pares de médias populacionais podem ser consideradas diferentes (por exemplo, o crescimento médio dos pinheiros gregos em Sines e em Tavira), mas será preciso levar em conta as proveniências, e não apenas o local da realização do estudo.

7. (a) Trata-se dum delineamento factorial a dois factores: *localidade* (Factor A, com $a = 4$ níveis) e *cultivar* (Factor B, com $b = 9$ níveis). Existem $n_{ij} = 4 = n_c$ repetições em todas as $ab = 36$ situações experimentais (células), pelo que se trata dum delineamento equilibrado. Existem ao todo $n = abn_c = 144$ observações da variável resposta Y (rendimento, em kg/ha). O modelo ANOVA adequado é o modelo ANOVA a dois factores, com interacção, dado por:

   i. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \ \forall\, i = 1, 2, 3, 4 \ , \ j = 1, 2, ..., 9 \ , \ k = 1, 2, 3, 4,$
   com $\alpha_1 = 0$, $\beta_1 = 0$, $(\alpha\beta)_{1j} = 0$ para qualquer j, e $(\alpha\beta)_{i1} = 0$ para qualquer i, onde
      - Y_{ijk} indica o rendimento na k-ésima parcela da localidade i, associada à cultivar j;
      - μ_{11} indica o rendimento médio (populacional) da cultivar *Celta*, em Elvas;
      - α_i indica o efeito principal da localidade i;
      - β_j indica o efeito principal da cultivar j;
      - $(\alpha\beta)_{ij}$ indica o efeito de interacção entre a localidade i e a cultivar j; e
      - ϵ_{ijk} indica o erro aleatório associado à observação Y_{ijk}.

   ii. $\epsilon_{ijk} \cap \mathcal{N}(0, \sigma^2), \ \forall\, i, j, k.$

   iii. $\{\epsilon_{ijk}\}_{i,j,k}$ constitui um conjunto de variáveis aleatórias independentes.

(b) i. Os nove valores em falta na tabela são dados por:

- $g.l.(SQA) = a - 1 = 3$;
- $g.l.(SQB) = b - 1 = 8$;
- $g.l.(SQAB) = (a-1)(b-1) = 3 \times 8 = 24$;
- $g.l.(SQRE) = n - ab = 144 - 36 = 108$;
- $SQB = QMB\,(b-1) = 964\,060 \times 8 = 7\,712\,480$;
- $SQAB = SQT - (SQA + SQB + SQRE) = (n-1)\,s_y^2 - 219\,628\,472 =$
  $= 143 \times 1\,714\,242 - 219\,628\,472 = 25\,508\,134$;
- $QMA = \frac{SQA}{a-1} = \frac{183\,759\,916}{3} = 61\,253\,305$;
- $QMAB = \frac{SQAB}{(a-1)(b-1)} = \frac{25\,508\,134}{24} = 1\,062\,839$;
- $F_B = \frac{QMB}{QMRE} = \frac{964\,060}{260\,704} = 3.69791$.

ii. Em qualquer modelo linear (regressão ou ANOVA), a variância dos erros aleatórios do modelo ($V[\epsilon_i] = \sigma^2$) é estimado pelo Quadrado Médio Residual. No nosso caso, a estimativa de σ^2 é dada no enunciado: $QMRE = 260\,704$. O valor muito elevado nada indica de especial, uma vez que a sua interpretação tem de levar em conta as unidades de medida dos dados, que são $(kg\,ha^{-1})^2$. De facto sabemos pelo enunciado que as unidades de medida da variável resposta são kg/ha. Sabemos que os resíduos ($e_i = y_i - \hat{y}_i$) têm as mesmas unidades de medida que a variável resposta. Sabemos que o QMRE é a Soma de Quadrados dos Resíduos a dividir pelos graus de liberdade associados, pelo que as unidades de medida do QMRE são o quadrado das unidades de medida da variável resposta. Bastava que os valores da variável resposta tivessem sido medidos em toneladas por hectare, para que o Quadrado Médio Residual viesse em $(t\ ha^{-1})^2$, ou seja, que fosse um milhão de vezes inferior ao valor acima indicado: $QMRE = 0.260704$. Mas isso não altera os dados, nem a significância de cada tipo de efeitos previsto no modelo. Assim, não é possível avaliar a estimativa de σ^2 apenas olhando para o valor absoluto de $QMRE$: é essencial ter em conta as unidades de medida associadas.

iii. Pedem-se os três testes F para cada tipo de efeitos previstos no modelo. Efectuemos em pormenor o teste à existência de efeitos de interacção entre localidade e cultivar:

Hipóteses: $H_0 : (\alpha\beta)_{ij} = 0\ ,\ \forall\ i = 2,3,4$ e $j = 2,3,...,9$ [não há interacção]
vs. $H_1 : \exists\, i = 2,3,4\ ,\ j = 2,3,...,9$ tais que $(\alpha\beta)_{ij} \neq 0$ [há interacção].

Estatística do teste: $F = \frac{QMAB}{QMRE} \ \cap\ F_{[(a-1)(b-1)\,,\,n-ab]}$, sob H_0.

Nível de significância: $\alpha = 0.01$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.01(24,108)} \approx 1.97$.

Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 4.0768$. É um valor significativo ao nível $\alpha = 0.01$, rejeitando-se H_0 a favor da hipótese alternativa de que existem efeitos de interacção entre localidade e cultivar.

No que respeita ao teste para os efeitos principais do factor *localidade*, as hipóteses em confronto são $H_0 : \alpha_i = 0\ ,\ \forall\ i = 2,3,4$ vs. $H_1 : \exists\, i = 2,3,4\ ,$ tal que $\alpha_i \neq 0$. A Região Crítica é agora dada pela rejeição de H_0 caso $F_{calc} > f_{0.01\,(3,108)} \approx 3.97$. O valor elevadíssimo da estatística calculada $F_{calc} = 234.9531$ leva à rejeição clara de H_0, concluindo-se pela existência de importantes efeitos de localidade, nos rendimentos.
Finalmente, no teste aos efeitos principais do factor *cultivar*, as hipóteses em confronto são $H_0 : \beta_j = 0\ ,\ \forall\ j = 2,3,...,9$ vs. $H_1 : \exists\, j = 2,3,...,9\ ,$ tal que $\beta_j \neq 0$. A Região Crítica é agora dada pela rejeição de H_0 caso $F_{calc} > f_{0.01\,(8,108)} \approx 2.68$. O valor da

estatística calculada $F_{calc} = 3.698$ pertence à Região Crítica, levando à rejeição de H_0, concluindo-se também pela existência de efeitos de cultivar sobre os rendimentos.
Assim, conclui-se pela existência dos três tipos de efeitos, ao nível $\alpha = 0.01$, com destaque para a existência clara de efeitos de localidade.

iv. Pede-se para discutir o efeito sobre a tabela resultante de dividir a variável resposta por mil (passando o rendimento a ser expresso em t/ha). Os graus de liberdade não são, naturalmente, afectados. O mesmo não se passa com as Somas de Quadrados. À nova variável $Y^* = Y/1000$ corresponderão novas médias de nível, de célula e global, que também resultam de dividir por mil (para ficarem em t/ha). Tendo em conta que no modelo em questão, as médias de célula definem os valores ajustados, tem-se $\hat{Y}^*_{ijk} = \hat{Y}_{ijk}/1000$. Assim, as novas Somas de Quadrados resultam de dividir as suas congéneres originais por 1000^2, ou seja, por um milhão. De facto, $SQT^* = \sum_i \sum_j \sum_k (Y^*_{ijk} - \overline{Y}^*_{...})^2 = \sum_i \sum_j \sum_k (Y_{ijk}/1000 - \overline{Y}_{...}/1000)^2 = SQT/(1000^2)$. Também $SQRE^* = \sum_i \sum_j \sum_k (Y^*_{ijk} - \hat{Y}^*_{ijk})^2 = \sum_i \sum_j \sum_k (Y_{ijk}/1000 - \hat{Y}_{ijk}/1000)^2 = SQRE/(1000^2)$. De forma análoga, e utilizando as fórmulas para delineamentos equilibrados,

$$SQA^* = bn_c \sum_{i=1}^{a} (\overline{Y}^*_{i..} - \overline{Y}^*_{...})^2 = bn_c \sum_{i=1}^{a} (\overline{Y}_{i..}/1000 - \overline{Y}_{...}/1000)^2 = SQA/(1000^2)$$

$$SQB^* = an_c \sum_{j=1}^{b} (\overline{Y}^*_{.j.} - \overline{Y}^*_{...})^2 = an_c \sum_{j=1}^{b} (\overline{Y}_{.j.}/1000 - \overline{Y}_{...}/1000)^2 = SQB/(1000^2) \ .$$

Por diferença, tem igualmente de verificar-se $SQAB^* = SQAB/(1000^2)$. Assim, toda a coluna de Somas de Quadrados na tabela será dividida por um milhão. Essa mesma transformação aplica-se à coluna de Quadrados Médios (que resulta de dividir Somas de Quadrados por graus de liberdade). Mas na coluna final, correspondente aos valores calculados das estatísticas F, o quociente de Quadrados Médios mantém-se inalterado (a transformação multiplicativa de numerador e denominador é igual). Logo, as conclusões de todos os testes (incluíndo os respectivos *p-values*) mantêm-se inalterados.

v. **[Material Complementar]** Os dois gráficos de interacção reflectem a mesma informação, embora de formas diferentes. No gráfico da esquerda, as quatro localidades definem posições no eixo horizontal. Por cima de cada localidade encontram-se nove pontos, associados às nove cultivares. A ordenada de cada um desses nove pontos é dada pelo rendimento médio das parcelas correspondentes a essa combinação de localidade e cultivar. Os segmentos de recta unem os pontos correspondentes a cada cultivar (segundo a legenda indicada no gráfico). Embora haja algum paralelismo nas nove curvas seccionalmente lineares, para as três primeiras localidades, os rendimentos na Revilheira sugerem a existência de efeitos de interacção. Por exemplo, a cultivar *TE9110*, que regista o rendimento mais baixo em Elvas (facto que se pode confirmar na tabela de médias dada na alínea c) tem o segundo mais elevado rendimento na Revilheira. Também a cultivar *Celta*, cujo rendimento em Benavila é o terceiro mais baixo, regista o segundo maior rendimento em Elvas. Assim, há cultivares que manifestam "preferências" ou "aversões" por diferentes localidades, reflectindo efeitos de interacção. O teste à interacção efectuado na alínea anterior confirma que esses efeitos são significativos, ao nível $\alpha = 0.01$.

O gráfico da direita dá, como se disse, uma perspectiva diferente sobre a mesma informação. Agora, são as cultivares que definem nove posições no eixo horizontal. Por cima de cada uma dessas posições (cultivares) há quatro pontos, com ordenadas dadas pelos rendimentos médios da referida cultivar, nas quatro localidades consideradas no ensaio. Segmentos de recta unem os pontos correspondentes a uma mesma localidade. Neste gráfico torna-se evidente que os rendimentos são sempre bastante superiores em Elvas (no gráfico da esquerda, esse facto reflectia-se no "pico" por cima de Elvas). Essa será a principal razão pela clara rejeição da hipótese nula no teste à existência de efeitos principais de localidade. Por outro lado, os efeitos de interacção reflectem-se na mais visível ausência de paralelismo, nomeadamente nos traços correspondentes a Elvas e Revilheira, que para várias cultivares parecem ter comportamentos quase antagónicos.

8. (a) Trata-se dum delineamento factorial a dois factores: *Temperatura de conservação* (Factor A), com $a = 2$ níveis, e *Tempo de armazenamento* (Factor B), com $b = 4$ níveis. Para modelar a variável resposta Y (alterações no conteúdo em taninos das polpas de sapoti), utiliza-se um modelo ANOVA a dois factores, com interacção. É possível estudar a interacção devido à presença de repetições nas $2 \times 4 = 8$ células. Sempre que possível, é desejável considerar este modelo para delineamentos factoriais a dois factores, deixando que sejam os dados a sugerir se se deve admitir a existência desse tipo de efeitos. O delineamento é equilibrado, uma vez que todas as células têm o mesmo número de repetições: $n_{ij} = 4 = n_c$ $(\forall\, i, j)$, para um total de $n = 8 \times 4 = 32$ observações. O modelo é dado por:

   i. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},\ \ \forall\, i = 1, 2\ ,\ j = 1, 2, 3, 4\ ,\ k = 1, 2, 3, 4,$
   com $\alpha_1 = 0$, $\beta_1 = 0$, $(\alpha\beta)_{1j} = 0$ para qualquer j, e $(\alpha\beta)_{i1} = 0$ para qualquer i, onde
      - Y_{ijk} indica a k-ésima observação (repetição) na célula definida pelo nível i do Factor A e o nível j do Factor B;
      - μ_{11} indica a média (populacional) das observações na célula $(1, 1)$, ou seja, com temperatura alta e 0 dias de armazenamento;
      - α_i indica o efeito do nível i do Factor A (*Temperatura*);
      - β_j indica o efeito do nível j do Factor B (*Tempo de armazenamento*);
      - $(\alpha\beta)_{ij}$ indica o efeito de interacção na célula (i, j); e
      - ϵ_{ijk} indica o erro aleatório associado à observação Y_{ijk}.

   ii. $\epsilon_{ijk} \cap \mathcal{N}(0, \sigma^2)\ ,\ \ \forall\, i, j, k.$

   iii. $\{\epsilon_{ijk}\}_{i,j,k}$ constituem um conjunto de variáveis aleatórias independentes.

   (b) A tabela-resumo desta ANOVA terá três linhas associadas a cada tipo de efeitos previsto no modelo (ou seja, efeitos principais do Factor A, efeitos principais do Factor B e efeitos de interacção) e ainda uma linha para o residual (podendo também incluir-se a linha associada à variabilidade Total). Como em qualquer modelo ANOVA, a tabela-resumo tem as seguintes colunas: Somas de Quadrados, graus de liberdade correspondentes, Quadrados Médios e estatísticas F. Os graus de liberdade são dados por:

   - Factor A: $a - 1 = 1$;
   - Factor B: $b - 1 = 3$;
   - Interacção: $(a - 1)(b - 1) = 3$;
   - Residual: $n - ab = 32 - 8 = 24$.

   Para calcular as Somas de Quadrados, registamos que no enunciado é dada a Soma de Quadrados Residual $SQRE = 20.72$. É igualmente dado o Quadrado Médio do Factor

B, e multiplicando pelos respectivos graus de liberdade obtém-se $SQB = QMB\,(b-1) = 96.01 \times 3 = 288.03$. A Soma de Quadrados Total também pode ser calculada facilmente, uma vez que no enunciado á dada a variância da totalidade das observações de Y, $s_y^2 = 47.83222$, e $SQT = (n-1)\,s_y^2 = 31 \times 47.83222 = 1482.799$. Assim, faltam as duas Somas de Quadrados relativas aos efeitos principais do factor A (SQA) e aos efeitos de interacção ($SQAB$). Utilizando a expressão para SQA, no caso de delineamentos equilibrados (disponível no formulário) e os valores das médias de nível do factor A e da média geral (disponíveis no enunciado), tem-se $SQA = bn_c \sum_{i=1}^{a} (\overline{Y}_{i..} - \overline{Y}_{...})^2 = 16\,[(24.681 - 22.14375)^2 + (19.606 - 22.14375)^2] = 16 \times 12.87781 = 206.045$. A última Soma de Quadrados em falta ($SQAB$) pode ser calculada a partir das restantes quatro: $SQAB = SQT - (SQA + SQB + SQRE) = 1482.799 - (206.045 + 288.03 + 20.72) = 968.004$. Assim,

Variação	g.l.	SQs	QMs	F_{calc}
Factor A	1	206.045	$QMA = \frac{SQA}{a-1} = 206.045$	$F = \frac{QMA}{QMRE} = 238.6622$
Factor B	3	288.03	$QMB = \frac{SQB}{b-1} = 96.01$	$F = \frac{QMB}{QMRE} = 111.2085$
Interacção	3	968.004	$QMAB = \frac{SQAB}{(a-1)(b-1)} = 322.668$	$F = \frac{QMAB}{QMRE} = 373.7467$
Residual	24	20.72	$QMRE = \frac{SQRE}{n-ab} = 0.8633333$	–
Total	31	1482.799	–	–

(c) De acordo com o modelo, a influência do Factor B nos valores da variável resposta pode resultar de dois tipos de efeitos: os efeitos principais do Factor B (os β_j) ou os efeitos de interacção (os $(\alpha\beta)_{ij}$). Efectuaremos estes dois testes, começando pelo dos efeitos de interacção. Neste exemplo, e como o Factor A apenas tem dois níveis, o índice i nos efeitos de interacção apenas toma o valor $i = 2$.

Hipóteses: $H_0 : (\alpha\beta)_{2j} = 0\ ,\ \forall\ j = 2,3,4$ vs. $H_1 : \exists\, j = 2,3,4$ tal que $(\alpha\beta)_{2j} \neq 0$.
Estatística do teste: $F = \frac{QMAB}{QMRE} \ \cap\ F_{[(a-1)(b-1)\,,\,n-ab]}$, sob H_0.
Nível de significância: $\alpha = 0.05$.
Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(3,24)} = 3.01$.
Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 373.7467$. É um valor claramente significativo e rejeita-se H_0 a favor da hipótese alternativa de que existem efeitos de interacção.

Já é possível responder afirmativamente: o Factor B tem efeitos sobre os valores médios de Y. No entanto, efectuaremos também o teste aos efeitos principais do Factor B:

Hipóteses: $H_0 : \beta_j = 0\ ,\ \forall\ j = 2,3,4$ vs. $H_1 : \exists\, j = 2,3,4$ tal que $\beta_j \neq 0$.
Estatística do teste: $F = \frac{QMB}{QMRE} \ \cap\ F_{(b-1\,,\,n-ab)}$, sob H_0.
Nível de significância: $\alpha = 0.05$.
Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(3,24)} = 3.01$.
Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 111.2085$. É um valor claramente significativo e rejeita-se H_0 a favor da hipótese de que existem efeitos principais do Factor B.

Assim, quer pela via dos efeitos principais, quer pela via dos efeitos de interacção, o Factor B (*tempo de armazenamento*) afecta os conteúdos médios de taninos nos sapotis.

9. (a) Trata-se dum delineamento a dois factores, o factor *casta* (factor A), e o factor *genótipo* (factor B). O objectivo do estudo é avaliar os eventuais efeitos destes factores sobre a variável

resposta (rendimento). Pela própria natureza dos factores em questão, o delineamento deve ser considerado *hierarquizado*, com genótipos subordinados a castas. Não faria sentido considerar o delineamento factorial: não há cruzamentos entre cada um dos oito genótipos e cada uma das duas castas, já que um genótipo apenas faz sentido quando referido à sua casta.

Assim, temos $a=2$ castas (níveis do factor A) e, para o factor subordinado genótipos, há $b_1=4$ genótipos para a casta 1 (Antão Vaz) e $b_2=4$ genótipos para a casta 2 (Malvasia Fina). Ao todo há $b_1+b_2=8$ situações experimentais, e $n_c=8$ repetições em cada uma das situações experimentais, num total de $n=64$ observações. O modelo mais adequado será o modelo hierarquizado:

- $Y_{ijk} = \mu_{11} + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$, $\forall\, i,j,k$, onde Y_{ijk} indica o rendimento da repetição k ($k = 1,2,...,8$) do genótipo j ($j = 1,2,3,4$) da casta i ($i = 1,2$). Impõem-se as restrições $\alpha_1 = 0$, $\beta_{1(i)} = 0$ para $i = 1,2$. Com estas restrições, o parâmetro μ_{11} é o rendimento médio populacional do primeiro genótipo da casta 1, isto é, do genótipo AN105 da casta Antão Vaz; α_2 é o efeito da casta Malvasia Fina; $\beta_{j(i)}$ ($j = 2,3,4$) é o efeito do genótipo j na casta $i = 1,2$, e ϵ_{ijk} é o erro aleatório associado à observação Y_{ijk}, que corresponde à variabilidade não explicada pelos efeitos previstos no modelo.
- $\epsilon_{ijk} \cap \mathcal{N}(0,\sigma^2)$, para qualquer i,j,k.
- Os erros aleatórios ϵ_{ijk} são independentes.

(b) Sabemos que os graus de liberdade na tabela-resumo da ANOVA são dados por: $a-1=1$ para o efeitos de castas; $(b_1-1)+(b_2-1) = 6$ para os efeitos do factor subordinado, genótipos; e $n-(b_1+b_2) = 64-8 = 56$ para o residual. Por outro lado, conhecemos a partir do enunciado a Soma de Quadrados do Factor A (castas), $SQA = 79.73597$ e o Quadrado Médio Residual, $QMRE = \frac{SQRE}{n-(b_1+b_2)} = 2.873782$, de onde é possível obter a Soma de Quadrados Residual $SQRE = 2.873782 \times 56 = 160.9318$. A Soma de Quadrados associada ao factor subordinado (genótipos) pode ser obtida pela diferença da soma das outras SQs já calculadas em relação à Soma de Quadrados Total, que sai do conhecimento da variância amostral da totalidade das 64 observações. Assim, $SQT = (n-1)s_y^2 = 63 \times 5.389415 = 339.5331$, logo $SQB(A) = SQT - (SQA + SQRE) = 339.5331 - (79.73597 + 160.9318) = 98.86533$. Os Quadrados Médios restantes obtêm-se dividindo Somas de Quadrados pelos respectivos graus de liberdade e os valores das duas estatísticas F resultam de dividir o correspondente quadrado médio pelo $QMRE$. Os valores resultantes são sintetizados na tabela em baixo.

Variação	g.l.	SQs	QMs	F
Casta (A)	1	79.73597	79.73597	$F_A = \frac{79.73597}{2.873782} = 27.74601$
Genótipo [B(A)]	6	98.86533	16.47755	$F_{B(A)} = \frac{16.47755}{2.873782} = 5.733751$
Residual	56	160.9318	2.873782	–
Total	63	339.5331	5.389415	–
	$(n-1)$	(SQT)	(s_y^2)	–

(c) Para responder será necessário efectuar um teste F aos efeitos do factor subordinado (genótipos), cuja hipótese nula corresponde à inexistência desse tipo de efeitos.

Hipóteses: $H_0 : \beta_{j(i)} = 0\ ,\ \forall\ i,j$ vs. $H_1 : \exists\, i,j$ tal que $\beta_{j(i)} \neq 0$.

Estatística do Teste: $F_{B(A)} = \frac{QMB(A)}{QMRE} \cap F_{[(b_1-1)+(b_2-1),n-(b_1+b_2)]}$, sob H_0.

Nível de significância: O enunciado pede o nível $\alpha=0.05$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(6,56)}$ que, pelas tabelas é um valor entre os valores tabelados 2.25 e 2.34.

Conclusões: Como $F_{calc} = 5.733751 > 2.34$, rejeita-se H_0, o que corresponde a admitir a existência de efeitos de genótipos.

Assim, foi importante prever este tipo de efeitos. Ignorar a existência de efeitos de genótipos iria inflaccionar a Soma de Quadrados Residual, o que poderia mascarar a existência de efeitos do outro factor (casta), mesmo que eles existam.

(d) Um teste análogo, mas aos efeitos do factor dominante (casta) terá como hipóteses $H_0 : \alpha_2 = 0$ (uma vez que apenas existem duas castas e impôs-se a restrição $\alpha_1 = 0$) *vs.* $H_1 : \alpha_2 \neq 0$. A região crítica deste teste (igualmente unilateral direita) é $f_{0.05(1,56)}$, um valor entre os valores tabelados 4.00 e 4.08. Como $F_{calc} = 27.746 > 4.08$, rejeita-se a hipótese nula. Assim, conclui-se (ao nível de significância $\alpha = 0.05$) que o efeito $\alpha_2 \neq 0$, ou seja que, para além de existirem efeitos de genótipos, há um efeito significativo de casta, e havendo apenas duas castas, pode-se afirmar que os rendimentos da casta Malvasia Fina são significativamente diferentes dos da casta Antão Vaz.

(e) **[Material Complementar]** O genótipo MF201 referido no enunciado tem o maior rendimento médio amostral $\overline{y}_{2,4} = 7.678$ (ordenando os genótipos como o `R`). Pretende-se saber que outras médias amostrais $\overline{y}_{ij}$ diferem significativamente de $\overline{y}_{2,4}$. Utilizaremos as comparações múltiplas de Tukey ao nível global $\alpha = 0.05$. O termo de comparação correspondente é $q_{\alpha(b_1+b_2,n-(b_1+b_2))}\sqrt{\frac{QMRE}{n_c}} = q_{0.05(8,56)}\sqrt{\frac{2.873782}{8}} \approx 4.45 \times 0.5993519 = 2.667$. Qualquer média amostral de rendimento de genótipo inferior a $7.678 - 2.667 = 5.011$ deverá assim ser considerada significativamente diferente da média do genótipo MF201. Há apenas dois genótipos que não têm rendimentos significativamente diferentes, ambos da casta Malvasia Fina: MF1420 e MF1426. Assim, não se rejeitam as hipóteses $\mu_{MF201} = \mu_{MF1420}$ e $\mu_{MF201} = \mu_{MF1426}$. Os três genótipos em questão são da casta Malvasia Fina, o que é coerente com a conclusão da alínea anterior: para além de efeitos de genótipo, é possível falar de efeitos de casta, sendo os rendimentos da casta Malvasia Fina globalmente superiores.

10. (a) Pede-se para mostrar que a soma dos n_i resíduos e_{ij}, correspondentes ao nível i do Factor $(i = 1, 2, ..., k)$, numa ANOVA a 1 Factor, é nula. Sabemos que, neste tipo de delineamento, os valores ajustados de cada observação correspondem à média amostral das n_i observações no nível i do Factor em que essa observação foi efectuada. Assim,

$$\sum_{j=1}^{n_i} e_{ij} \;=\; \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij}) \;=\; \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i.}) \;=\; 0 \;,$$

uma vez que se trata duma soma de desvios dum conjunto de observações em relação à sua média (ou seja, do tipo $\sum_{i=1}^{n}(x_i - \overline{x})$, estudada no Exercício 4a da Regressão Linear) que tem sempre soma zero.

(b) Trata-se duma situação análoga à da alínea anterior. Num modelo ANOVA a dois factores, com efeitos de interacção, sabemos que os valores ajustados $\hat{y}_{ijk}$ correspondem às médias $\overline{y}_{ij.}$ das observações da célula da referida observação. Assim, a soma dos resíduos das n_{ij} observações efectuadas na célula (i, j) é dada por:

$$\sum_{k=1}^{n_{ij}} e_{ijk} \;=\; \sum_{k=1}^{n_{ij}} (y_{ijk} - \hat{y}_{ijk}) \;=\; \sum_{k=1}^{n_{ij}} (y_{ijk} - \overline{y}_{ij.}) \;=\; 0 \;.$$

11. Tendo em conta que, no contexto duma ANOVA a um factor, a tradicional Soma de Quadrados associada ao ajustamento do modelo (que na regressão linear se designa SQR) é chamada SQF, tem-se $R^2 = \frac{SQF}{SQT}$.

    (a) A condição $R^2 = 0$ equivale a $SQF = 0$. Ora, no contexto ANOVA a um factor tem-se (ver formulário e tendo em conta que o delineamento é equilibrado):

    $$SQF = \sum_{i=1}^{k} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2 = n_c \sum_{i=1}^{k}(\overline{Y}_{i.} - \overline{Y}_{..})^2 = 0 \ .$$

    Ora, uma soma de quadrados só se pode anular se *todas* as suas parcelas se anulam o que, neste contexto, significa que $\overline{Y}_{i.} = \overline{Y}_{..}$, para todo o i. Por outras palavras, $R^2 = 0$ se e só se todas as médias amostrais de nível forem iguais à média amostral da totalidade das observações (e portanto iguais entre si). Assim, a informação proveniente da amostra aponta de forma clara em abono da hipótese de igualdade de todas as médias populacionais de nível ($\mu_1 = \mu_2 = ... = \mu_k$), que é a hipótese nula no teste F duma ANOVA a um único factor. Este resultado é inteiramente coerente com a não rejeição da hipótese nula do teste que resulta do facto de $R^2 = 0 \Leftrightarrow F_{calc} = 0$. Repare-se ainda que a condição $SQF = 0$ é equivalente a dizer que $SQT = SQF + SQRE = SQRE$, ou seja, toda a variabilidade de Y é residual, ou seja, interna aos níveis do factor.

    (b) A condição $R^2 = 1$ equivale a $SQF = SQT$, ou seja, $SQRE = 0$. Ora, no contexto ANOVA a um factor tem-se (ver formulário e para um delineamento equilibrado):

    $$SQRE = \sum_{i=1}^{k}(n_i - 1)S_i^2 = (n_c - 1)\sum_{i=1}^{k} S_i^2 = 0 \ .$$

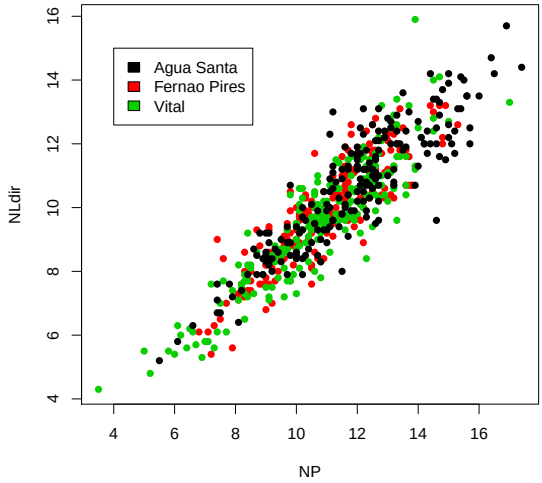
    De novo, uma soma de quadrados só pode ser nula se *todas* as suas parcelas forem nulas, pelo que $SQRE = 0$ equivale a $S_i^2 = 0$, para todo o nível i, ou seja, não existe variabilidade das observações de Y no seio dum mesmo nível do factor. Neste caso tem-se também $QMRE = \frac{SQRE}{n-k} = 0$. Embora não seja possível construir a estatística do teste $F = \frac{QMF}{QMRE}$, a divisão por zero sugere um valor limite infinito, que corresponderia sempre à rejeição da hipótese nula de igualdade das médias populacionais de nível μ_i, o que é coerente com o referido facto de, neste caso, toda a variabilidade nas observações de Y corresponder à mudança entre níveis do factor.

3 Análise de Covariância

1. Neste exercício consideram-se os dados da *data frame* `videiras`. A variável resposta é, em todas as alíneas, o comprimento da nervura lateral direita (`NLdir`) e o preditor, o comprimento da nervura principal (`NP`).

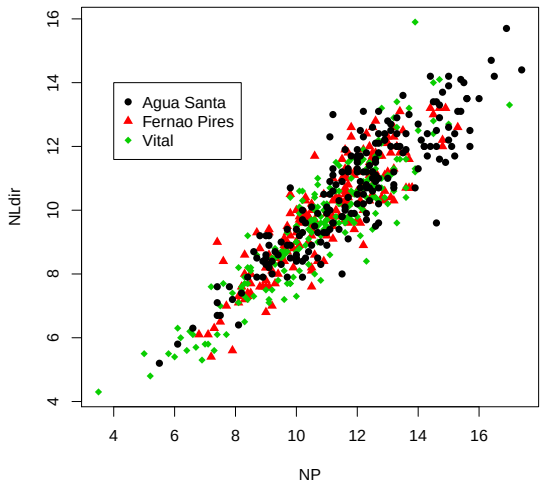
    (a) Os comandos `R` para obter a nuvem de pontos pedida, e o respectivo resultado, são:

    ```
    > plot(NLdir ~ NP, col=Casta, data=videiras, pch=16)
    > legend(4,15,legend=levels(videiras$Casta), fill=1:3)
    ```



Alternativamente, podemos também querer construir um gráfico com, não apenas cores diferentes, mas também símbolos diferentes para cada casta. Eis uma forma possível de construir um tal gráfico no R, usando os símbolos a que correspondem os códigos 16 (círculos), 17 (triângulos) e 18 (losangos), como indicado na legenda.

```
> plot(NLdir ~ NP, col=as.numeric(Casta), pch=as.numeric(Casta)+15, data=videiras)
> legend(4,14,levels(videiras$Casta),col=1:3, pch=16:18)
```



A nuvem de pontos sugere a existência duma relação linear bastante intensa, que poderá ser a mesma nas três castas consideradas. A nuvem sugere também que poderá haver dispersões maiores das observações, em torno da recta de fundo, para as folhas de maior dimensão.

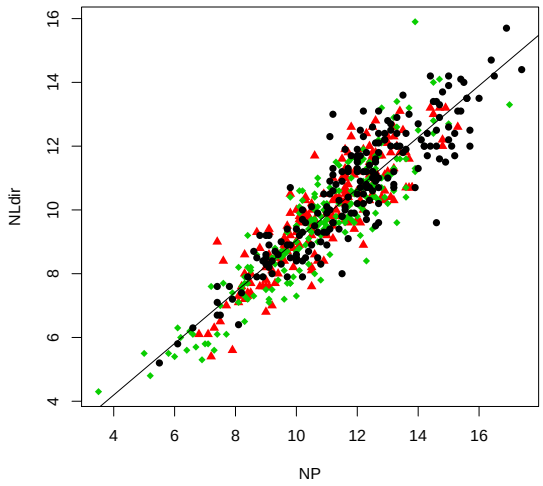
(b) Eis os comandos R necessários, e os resultados numéricos correspondentes:

```
> videirasN.lm <- lm(NLdir ~ NP, data=videiras)
> summary(videirasN.lm)
[...]
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.96218    0.18309   5.255 2.06e-07 ***
NP           0.80841    0.01607  50.314  < 2e-16 ***
---
Residual standard error: 0.8339 on 598 degrees of freedom
Multiple R-squared:  0.8089,Adjusted R-squared:  0.8086
F-statistic:  2532 on 1 and 598 DF,  p-value: < 2.2e-16
```

Assim, a recta de regressão $y = 0.96218 + 0.80841\,x$ explica cerca de 81% da variabilidade observada nas nervuras laterais direitas, para o conjunto das $n = 600$ observações. Trata-se duma aproximação razoavelmente boa (como se pode constatar no gráfico), que explica cerca de 81% da variabilidade observada nas nervuras laterais direitas. Como seria de esperar, o modelo ajustado difere significativamente do modelo nulo, tendo a estatística calculada no teste F de ajustamento global um valor $F_{calc} = 2532$, cuja significância (*p-value*) correspondente é inferior à precisão de máquina, logo indistinguível de zero.

```
> abline(videirasN.lm)
```



(c) Eis os comandos R necessários, e os resultados numéricos correspondentes ao modelo ANCOVA pedido:

```
> videirasNCasta.lm <- lm(NLdir ~ NP*Casta, data=videiras)
> summary(videirasNCasta.lm)
[...]
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.39812    0.32102   4.355 1.57e-05 ***
NP                    0.77780    0.02654  29.305  < 2e-16 ***
CastaFernao Pires    -0.43069    0.48897  -0.881    0.379
CastaVital           -0.66120    0.43788  -1.510    0.132
NP:CastaFernao Pires  0.03395    0.04253   0.798    0.425
NP:CastaVital         0.04100    0.03798   1.079    0.281
---
Residual standard error: 0.8316 on 594 degrees of freedom
Multiple R-squared:  0.8112,Adjusted R-squared:  0.8096
```

```
F-statistic: 510.5 on 5 and 594 DF,  p-value: < 2.2e-16
```

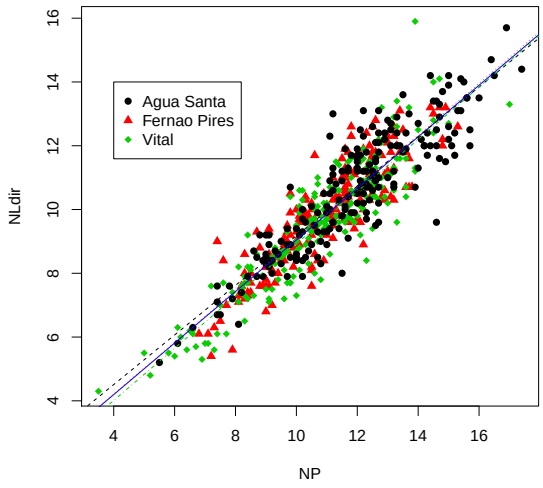
A recta para a casta Água Santa (a casta correspondente ao primeiro nível do factor, o nível de referência, logo não explicitada na listagem de resultados) tem equação $y = 1.39812 + 0.77780\,x$. Para obter a equação correspondente à casta Fernão Pires, será necessário acrescentar à ordenada na origem o acréscimo estimado $\hat{\alpha}_{0:2} = -0.43069$ e ao declive, o respectivo acréscimo estimado, $\hat{\alpha}_{1:2} = 0.03395$. De forma análoga, obtém-se a recta ajustada para a casta Vital. Eis as equações das três rectas ajustadas:

Casta Água Santa	$y = 1.39812 + 0.77780\,x$
Casta Fernão Pires	$y = (1.39812 - 0.43069) + (0.77780 + 0.03395)\,x = 0.96743 + 0.81175\,x$
Casta Vital	$y = (1.39812 - 0.66120) + (0.77780 + 0.04100)\,x = 0.73692 + 0.81880\,x$

Para traçar as rectas de cada casta na nuvem de pontos já criada, podem usar-se os seguintes comandos:

```
> coefVidCasta <- coef(videirasNCasta.lm)
> coefVidCasta
(Intercept)         NP  CastaFernao Pires  CastaVital  NP:CastaFernao Pires   NP:CastaVital
 1.39811600  0.77779606        -0.43068514  -0.66119902           0.03394865      0.04100268

> abline(coefVidCasta[c(1,2)], col=1, lty=2)                      <-- recta casta Água Santa
> abline(coefVidCasta[c(1,2)]+coefVidCasta[c(3,5)],col=2,lty=3)   <-- recta casta Fernão Pires
> abline(coefVidCasta[c(1,2)]+coefVidCasta[c(4,6)],col=3,lty=4)   <-- recta casta Vital
```

Apesar das equações diferentes, as quatro rectas são difíceis de distinguir no gráfico.



(d) A equação do modelo ANCOVA ajustado pode escrever-se da seguinte forma, utilizando a notação vectorial:

$$\vec{\mathbf{y}} \;=\; \beta_0 + \beta_1\,\vec{\mathbf{x}} + \alpha_{0:2}\,\vec{\boldsymbol{\mathcal{I}}}_2 + \alpha_{0:3}\,\vec{\boldsymbol{\mathcal{I}}}_3 + \alpha_{1:2}\,\vec{\boldsymbol{\mathcal{I}}}_2 \star \vec{\mathbf{x}} + \alpha_{1:3}\,\vec{\boldsymbol{\mathcal{I}}}_3 \star \vec{\mathbf{x}} + \vec{\boldsymbol{\epsilon}}\,,$$

sendo $\vec{\boldsymbol{\mathcal{I}}}_i$ a variável indicatriz das observações da casta $i = 2, 3$ (Fernão Pires e Vital, respectivamente) e $\alpha_{j:i}$ o acréscimo no parâmetro β_j (em relação à casta de referência, a Água Santa), resultante de estarmos na casta $i = 2, 3$. O símbolo $\star$ indica um produto elemento a elemento entre dois vectores de igual dimensão. O modelo linear ajustado acima

pode agora ser visto como um submodelo deste modelo ANCOVA, associado à hipótese $\alpha_{0:2} = \alpha_{0:3} = \alpha_{1:2} = \alpha_{1:3} = 0$. Vamos efectuar um teste F parcial para testar a equivalência de modelo e submodelo.

Hipóteses: $H_0 : \alpha_{j:i} = 0, \forall j = 0, 1; i = 2, 3$ vs. $H_1 : \exists j = 0, 1; i = 2, 3$ tal que $\alpha_{j:i} \neq 0$.

Estatística do Teste: (na forma mais adequada à informação disponível)

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

Nível de significância: $\alpha = 0.05$.

Região Crítica: (unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(4,594)} \approx 2.39$.

Conclusão: Temos $F_{calc} = \frac{0.8112 - 0.8089}{1 - 0.8112} \cdot \frac{594}{4} = 1.809$. Logo, não rejeitamos H_0, isto é, não se pode dizer que o modelo ANCOVA se ajuste de forma significativamente diferente do modelo RLS com uma única recta para as três castas. Assim, não se justifica abandonar o modelo RLS, que é mais parcimonioso e tem um ajustamento considerado adequado.

Este teste F parcial, comparando o modelo ANCOVA ajustado na alínea anterior com o submodelo ajustado na alínea 1b (recta única para a totalidade das observações) obtém-se no `R` com o comando `anova`:

```
> anova(videirasN.lm, videirasNCasta.lm)
Analysis of Variance Table
Model 1: NLdir ~ NP
Model 2: NLdir ~ NP * Casta
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    598 415.80
2    594 410.81  4    4.9948 1.8055 0.1262
```

NOTA: A pequena discrepância no valor calculado da estatística de teste resulta de, na nossa resolução anterior, terem sido usados valores de R^2 arredondados a 4 casas decimais.

(e) Eis os três ajustamentos "mono-casta" pedidos.

i. Tendo em atenção que as $n_1 = 200$ observações da casta Água Santa estão nas linhas 401 a 600 da *data frame*, tem-se:

```
> summary(lm(NLdir ~ NP, data=videiras[401:600,]))
[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.39812    0.33349   4.192 4.16e-05 ***
NP           0.77780    0.02757  28.210  < 2e-16 ***
---
Residual standard error: 0.8639 on 198 degrees of freedom
Multiple R-squared:  0.8008,Adjusted R-squared:  0.7998
F-statistic: 795.8 on 1 and 198 DF,  p-value: < 2.2e-16
```

A recta de regressão obtida ($y = 1.39812 + 0.77780\,x$) é a mesma que no modelo completo (modelo ANCOVA) considerado acima. O valor do coeficiente de determinação ($R^2 = 0.8008$) é muito próximo do valor obtido com a recta única para a totalidade das $n = 600$ observações, facto que não era possível prever a partir dos ajustamentos anteriores.

ii. As $n_2 = 200$ observações da casta Fernão Pires estão nas 200 primeiras linhas do objecto `videiras`. Assim,

```
> summary(lm(NLdir ~ NP, data=videiras[1:200,]))
[...]
Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.96743    0.34914   2.771  0.00612 **
NP           0.81174    0.03146  25.801  < 2e-16 ***
---
Residual standard error: 0.7872 on 198 degrees of freedom
Multiple R-squared:  0.7708,Adjusted R-squared:  0.7696
F-statistic: 665.7 on 1 and 198 DF,  p-value: < 2.2e-16
```

Também neste caso, e como teria de ser, a recta obtida ($y = 0.96743 + 0.81174\,x$) é, a menos de erros de arredondamento, a recta obtida ao ajustar o modelo ANCOVA. Também neste caso, o coeficiente de determinação $R^2 = 0.7708$ é próximo do valor obtido para a recta única, embora neste caso não tenha necessariamente de ser assim.

iii. Para as restantes observações, relativas à casta Vital, tem-se:

```
> summary(lm(NLdir ~ NP, data=videiras[201:400,]))
[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.73692    0.30147   2.444   0.0154 *
NP           0.81880    0.02751  29.769   <2e-16 ***
---
Residual standard error: 0.8418 on 198 degrees of freedom
Multiple R-squared:  0.8174,Adjusted R-squared:  0.8164
F-statistic: 886.2 on 1 and 198 DF,  p-value: < 2.2e-16
```

Confirma-se a recta de regressão $y = 0.73692 + 0.8188\,x$, e mais uma vez o valor $R^2 = 0.8174$ é próximo do obtido com uma única recta de regressão para as três castas, o que é, como para as outras castas, uma particularidade deste exemplo, associada ao facto de as três nuvens de pontos serem de configuração semelhante.

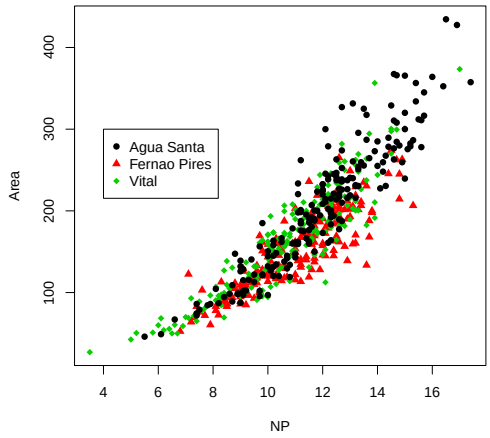
(f) O único modelo que não é de RLS é o modelo completo de ANCOVA, e será o único cuja matriz do modelo é aqui considerada. A fim de poupar no espaço, apenas se mostram as linhas correspondentes às três primeiras observações de cada casta. Recorde-se que à casta de referência (que, uma vez que o `R` ordena os níveis do factor por ordem alfabética, é a Água Santa) correspondem as últimas 200 linhas da matriz. As restantes castas estão indicadas nos nomes de coluna.

```
> model.matrix(videirasNCasta.lm)
    (Intercept)   NP CastaFernao Pires CastaVital NP:CastaFernao Pires NP:CastaVital
1             1 13.8                1          0                 13.8           0.0
2             1  9.1                1          0                  9.1           0.0
3             1 14.5                1          0                 14.5           0.0
[...]
201           1 11.7                0          1                  0.0          11.7
202           1 10.6                0          1                  0.0          10.6
203           1 11.0                0          1                  0.0          11.0
[...]
401           1 15.7                0          0                  0.0           0.0
402           1 11.7                0          0                  0.0           0.0
403           1 10.2                0          0                  0.0           0.0
[...]
```

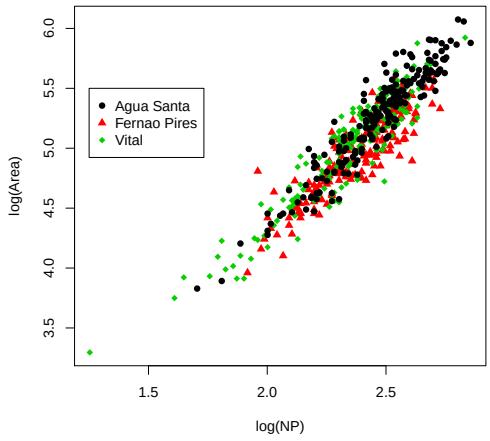
2. Neste exercício a variável resposta é `Area` e a variável preditora é `NP`.

(a) O gráfico (obtido de forma análoga ao que foi visto no Exercício 1a) torna evidente a existência duma curvatura na relação entre área foliar e comprimento da nervura principal. esta curvatura não é de estranhar, uma vez que a área é uma característica bi-dimensional,

enquanto que o comprimento é unidimensional, sugerindo que a área seja aproximadamente proporcional ao quadrado do comprimento da nervura.



(b) Com a dupla logaritmização pedida no enunciado obtém-se uma relação mais próxima da linearidade. Assim, a logaritmização de área foliar e de comprimento da nervura principal é uma boa transformação linearizante.



(c) O modelo pedido tem o seguinte ajustamento.

```
> vid.Anc2.lm <- lm(log(Area) ~ log(NP), data=videiras)
> summary(vid.Anc2.lm)
[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.57869    0.07703   7.513 2.12e-13 ***
log(NP)      1.87642    0.03203  58.579  < 2e-16 ***
---
Residual standard error: 0.1597 on 598 degrees of freedom
Multiple R-squared: 0.8516,Adjusted R-squared: 0.8513
F-statistic:  3431 on 1 and 598 DF,  p-value: < 2.2e-16
```

A recta ajustada, às variáveis logaritmizadas é $\ln(Area) = 0.57869 + 1.87642\,\ln(NP)$. Em termos das variáveis originais (não logaritmizadas), esta relação corresponde a uma relação potência $Area = \mathrm{e}^{0.57869}\,NP^{1.87642}$ (ver acetatos das aulas relativos às transformações linearizantes).

(d) O modelo ANCOVA agora pedido tem o seguinte ajustamento:

```
> vid.Anc2d <- lm(log(Area) ~ log(NP)*Casta, data=videiras)
> summary(vid.Anc2d)
[...]
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                0.33820    0.13050   2.592 0.009791 **
log(NP)                    1.99648    0.05294  37.711  < 2e-16 ***
CastaFernao Pires          0.62328    0.19914   3.130 0.001834 **
CastaVital                 0.39524    0.17007   2.324 0.020463 *
log(NP):CastaFernao Pires -0.31298    0.08232  -3.802 0.000158 ***
log(NP):CastaVital        -0.17654    0.07025  -2.513 0.012232 *
---
Residual standard error: 0.1482 on 594 degrees of freedom
Multiple R-squared: 0.8731,Adjusted R-squared: 0.872
F-statistic: 817.4 on 5 and 594 DF,  p-value: < 2.2e-16
```

O valor do coeficiente de determinação deste modelo ($R^2 = 0.8731$) é comparável com o do modelo de regressão linear simples ajustado na alínea anterior ($R^2 = 0.8516$), uma vez que em ambos os casos a escala da variável resposta é a de log-áreas. O coeficiente de determinação aumentou com o modelo ANCOVA (como tem de ser, uma vez que o modelo de uma única recta de regressão é um submodelo do modelo ANCOVA), mas o aumento não é muito acentuado (pouco mais de 2%), pelo que é legítima a dúvida se o aumento obtido com o modelo ANCOVA compensa a maior complexidade do modelo.

(e) Tendo em conta a natureza destes parâmetros estimados, resultam as seguintes relações para cada casta:

Água Santa	$\ln(Area) = 0.33820 + 1.99648\,\ln(NP)$	$\Leftrightarrow$	$Area = \mathrm{e}^{0.33820}\,NP^{1.99648}$
Fernão Pires	$\ln(Area) = 0.96148 + 1.6835\,\ln(NP)$	$\Leftrightarrow$	$Area = \mathrm{e}^{0.96148}\,NP^{1.6835}$
Vital	$\ln(Area) = 0.73344 + 1.81994\,\ln(NP)$	$\Leftrightarrow$	$Area = \mathrm{e}^{0.73344}\,NP^{1.81994}$

Em todos os casos, a área foliar é modelada como proporcional a uma potência do comprimento da nervura principal, potência essa que varia entre 1.68 e 2. Uma relação $Area = NP^2$ corresponderia a folhas de forma quadrada, com lado igual a NP. A forma irregular da folha justifica as potências menores que 2 e as constantes de proporcionalidade, que oscilam entre 1.40 (no caso da casta Água Santa) e 2.62 (casta Fernão Pires).

(f) Uma vez que os modelos das alíneas (c) e (d) são modelos encaixados, é possível usar um teste F parcial para estudar se o respectivo ajustamento é significativamente diferente. A equação do modelo ANCOVA é da forma

$$\vec{\boldsymbol{y}} \;=\; \beta_0 + \beta_1\,\vec{\mathbf{x}} + \alpha_{0:2}\,\vec{\boldsymbol{I}}_2 + \alpha_{0:3}\,\vec{\boldsymbol{I}}_3 + \alpha_{1:2}\,\vec{\boldsymbol{I}}_2 \star \vec{\mathbf{x}} + \alpha_{1:3}\,\vec{\boldsymbol{I}}_3 \star \vec{\mathbf{x}} + \vec{\boldsymbol{\epsilon}}\,,$$

sendo $\vec{\boldsymbol{I}}_i$ a variável indicatriz das observações da casta $i = 2, 3$ (Fernão Pires e Vital, respectivamente) e $\alpha_{j:i}$ o acréscimo no parâmetro β_j (em relação à casta de referência, a Água Santa), resultante de estarmos na casta $i = 2, 3$. O símbolo $\star$ indica um produto elemento a elemento entre dois vectores de igual dimensão. O modelo linear ajustado acima pode agora ser visto como um submodelo deste modelo ANCOVA, associado à hipótese $\alpha_{0:2} = \alpha_{0:3} = \alpha_{1:2} = \alpha_{1:3} = 0$.

Hipóteses: $H_0 : \alpha_{j:i} = 0, \forall j = 0, 1; i = 2, 3$ vs. $H_1 : \exists j = 0, 1; i = 2, 3$ tal que $\alpha_{j:i} \neq 0$.

Estatística do Teste: (na forma mais adequada à informação disponível)

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

Nível de significância: $\alpha = 0.05$.

Região Crítica: (unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(4,594)} \approx 2.39$.

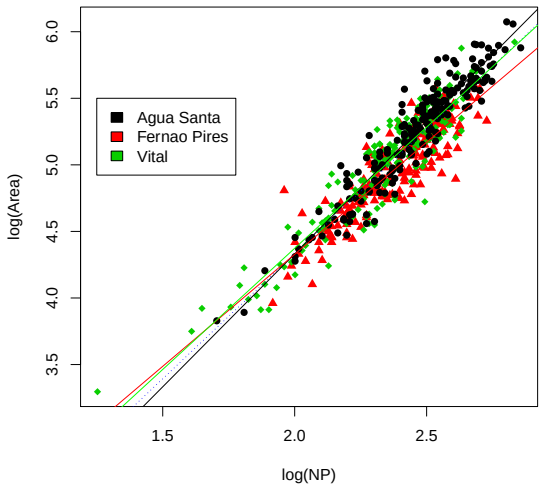
Conclusão: Temos $F_{calc} = \frac{0.8731 - 0.8516}{1 - 0.8731} \cdot \frac{594}{4} = 25.15957$. Logo, neste caso rejeita-se claramente H_0, isto é, conclui-se que o ajustamento do modelo ANCOVA é significativamente diferente do ajustamento do modelo RLS com uma única recta para as três castas. Assim, do ponto de vista estatístico justifica-se a utilização do modelo ANCOVA, com rectas/curvas diferentes para cada casta.

O recurso ao comando `anova` do `R` confirma o valor calculado da estatística (arredondamentos aparte) e o valor quase nulo do *p-value* correspondente.

```
> anova(vid.Anc2.lm, vid.Anc2d)
Analysis of Variance Table
Model 1: log(Area) ~ log(NP)
Model 2: log(Area) ~ log(NP) * Casta
Res.Df    RSS Df Sum of Sq      F    Pr(>F)
     1    598 15.248
     2    594 13.037  4    2.2102 25.174 < 2.2e-16 ***
```

(g) O gráfico pedido é indicado em baixo, sendo a recta única para a totalidade das $n = 600$ observações indicada a tracejado. O gráfico foi construído com os seguintes comandos do `R`:

```
> plot(log(Area)~log(NP),col=as.numeric(Casta),pch=as.numeric(Casta)+15,data=videiras)
> abline(vid.Anc2.lm, col="blue", lty="dotted")
> abline(0.33820, 1.99648, col="black")
> abline(0.33820+0.62328, 1.99648-0.31298, col="red")
> abline(0.33820+0.39524, 1.99648-0.17654, col="green")
> legend(1.25,5.5, levels(videiras$Casta), fill=1:3)
```

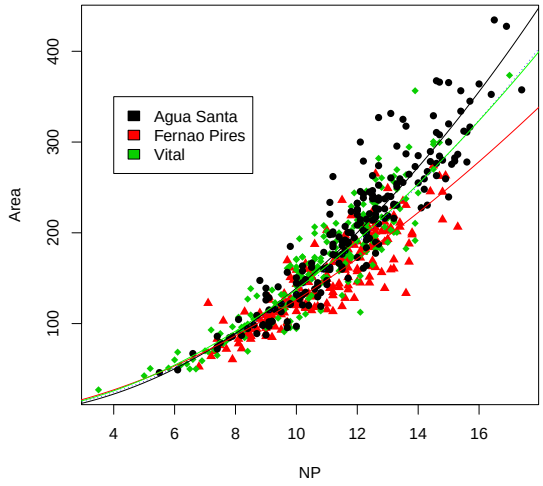


Confirma-se o maior declive da recta associada à casta Água Santa, e o menor associado à casta Fernão Pires. Em comparação com a relação análoga estudada no Exercício 1, é

visível uma maior distinção das três rectas ajustadas, que foi reflectida no facto de o teste F parcial ter considerado que o modelo ANCOVA e o modelo de regressão linear simples para as três castas em conjunto serem significativamente diferentes.

NOTA: Convém acrescentar que a signficância do teste F parcial resulta também do número bastante elevado de observações usado para ajustar estes modelos ($n = 600$). Quanto mais informação estiver disponível na amostra, mais facilmente as diferenças são consideradas significativas.

(h) O gráfico para as variáveis não logaritmizadas é o seguinte.



Foi produzido com os comandos:

```
> plot(Area ~ NP, col=as.numeric(Casta), pch=as.numeric(Casta)+15, data=videiras)
> curve(exp(0.5787)*x^(1.8764), from=0, to=18, col="blue", lty="dotted", add=TRUE)
> curve(exp(0.3382)*x^(1.9965), from=0, to=18, add=TRUE)
> curve(exp(0.3382+0.6233)*x^(1.9965-0.3130), from=0, to=18, col="red", add=TRUE)
> curve(exp(0.3382+0.3952)*x^(1.9965-0.1765), from=0, to=18, col="green", add=TRUE)
> legend(4,350, levels(videiras$Casta), fill=1:3)
```

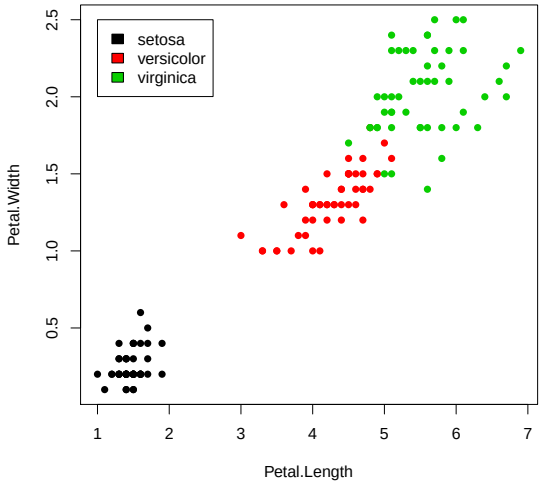
Nas escalas originais (não logaritmizadas) as diferenças entre as castas Água Santa e Fernão Pires é mais visível. A casta Vital tem um comportamento muito próximo do comportamento conjunto das três castas, sendo a sua curva ajustada quase indistinguível da curva única para as três castas (representada a ponteado).

3. FALTA

4. Neste exercício, consideram-se as $n = 150$ observações sobre lírios, com variável resposta dada pela largura das pétalas (variável `Petal.Width`) e preditor numérico comprimento das pétalas (`Petal.Length`). Será considerado também o factor espécie (`Species`), havendo $n_i = 50$ observações de cada espécie.

   (a) O gráfico pedido é obtido com os comandos seguintes. A nuvem é prometedora para uma relação linear global.

```
> plot(Petal.Width ~ Petal.Length, col=Species, data=iris, pch=16)
> legend(1,2.5, legend=levels(iris$Species), fill=1:3)
```



(b) Tem-se:

```
> iris.lm <- lm(Petal.Width ~ Petal.Length, data=iris)
> summary(iris.lm)
[...]
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 recta $y = -0.363076 + 0.415755\,x$ explica quase 93% da variabilidade observada nas larguras das pétalas, para o conjunto das três espécies de lírios.

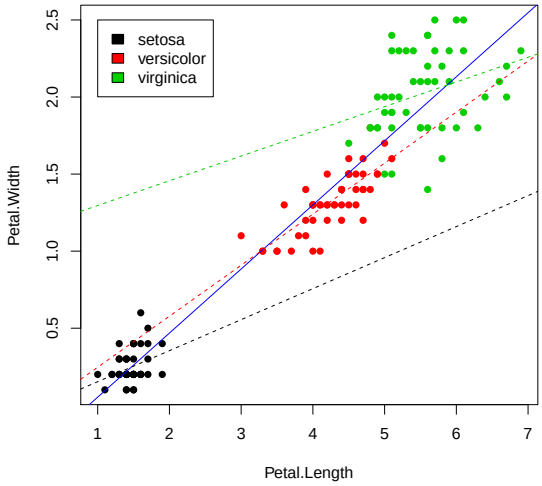
(c) O modelo completo, cruzando o preditor numérico `Petal.Length` com o factor `Species` é:

```
> irisSpecies.lm <- lm(Petal.Width ~ Petal.Length*Species, data=iris)
> summary(irisSpecies.lm)
[...]
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -0.04822    0.21472  -0.225 0.822627
Petal.Length                  0.20125    0.14586   1.380 0.169813
Speciesversicolor            -0.03607    0.31538  -0.114 0.909109
Speciesvirginica              1.18425    0.33417   3.544 0.000532 ***
Petal.Length:Speciesversicolor  0.12981    0.15550   0.835 0.405230
Petal.Length:Speciesvirginica  -0.04095    0.15291  -0.268 0.789244
---
Residual standard error: 0.1773 on 144 degrees of freedom
Multiple R-squared:  0.9477,Adjusted R-squared:  0.9459
F-statistic: 521.9 on 5 and 144 DF,  p-value: < 2.2e-16
```

i. As três rectas de regressão, para cada espécie individual são: $y=-0.04822+0.20125\,x$ para a espécie *setosa*, $y=-0.08429+0.33106\,x$ para a espécie *versicolor*, e $y=1.1360+0.1603\,x$ para a espécie *virginica*. O valor do coeficiente de determinação do modelo ANCOVA, $R^2=0.9477$ é naturalmente maior do que o R^2 do submodelo constituído por uma única recta de regressão. Mas o seu valor não é de interpretação imediata, como se viu nas aulas e como se verá nas alíneas seguintes. Para traçar estas três rectas por espécie individual em cima da nuvem de pontos já anteriormente obtida, podem dar-se os seguintes comandos:

```
> coefIrisSpecies <- coef(irisSpecies.lm)
> abline(coefIrisSpecies[c(1,2)], col=1, lty=2)
> abline(coefIrisSpecies[c(1,2)]+coefIrisSpecies[c(3,5)], col=2, lty=2)
> abline(coefIrisSpecies[c(1,2)]+coefIrisSpecies[c(4,6)], col=3, lty=2)
```

Os resultados obtidos, juntamente com a recta única obtida para a totalidade das $n=150$ observações (a azul, em traço contínuo), são indicados no gráfico seguinte.



Como se pode constatar, a situação é bem mais confusa do que no exercício 1, com duas das rectas (das espécies *setosa* e *virginica*) com declives bastante diferentes em relação aos da recta global e da recta da espécie *versicolor*. No entanto, as rectas das espécies *setosa* e *virginica* parecem ser aproximadamente paralelas, sendo os declives ajustados (0.20125 e 0.1603) próximos. No modelo completo discutido nas aulas, o declive da recta para a espécie de referência (*setosa*) é o parâmetro β_1. O declive da recta para a espécie *virginica* é a soma de β_1 com o acréscimo específico do declive da espécie *virginica*, ou seja, com o acréscimo $\alpha_{1:3}$. A hipótese de que essas duas rectas sejam paralelas corresponde à hipótese de $H_0:\ \alpha_{1:3}=0$. Esta hipótese corresponde a um teste a um parâmetro individual num modelo linear (ou seja, corresponde aos testes t usados na regressão linear para aferir possíveis valores de cada β_j). A informação necessária para efectuar esse teste está disponível na listagem de resultados obtida acima para o modelo `irisSpecies.lm`. Em particular, a estimativa desse acréscimo é -0.04095, com um erro padrão associado de $\hat{\sigma}_{\hat{\alpha}_{1:3}}=0.15291$. Tendo em conta a hipótese nula referida, a estatística t do teste também é dada na listagem e tem valor $T_{calc}=-0.268$, a que corresponde um valor de prova $p=0.789244$. Sendo assim, está-se muito longe de rejeitar a hipótese nula $H_0:\ \alpha_{1:3}=0$, para qualquer nível de significância usual. Assim, não se rejeita que essas duas rectas de espécie são paralelas.

(d) Os três modelos individuais de espécie, ajustados apenas usando as $n_i = 50$ observações de cada espécie têm os coeficientes de determinação indicados de seguida:

```
> irisSetosa.lm <- lm(Petal.Width ~ Petal.Length, data=iris[1:50,])
> irisVersi.lm <- lm(Petal.Width ~ Petal.Length, data=iris[51:100,])
> irisVirgi.lm <- lm(Petal.Width ~ Petal.Length, data=iris[101:150,])
> summary(irisSetosa.lm)$r.sq
[1] 0.1099785
> summary(irisVersi.lm)$r.sq
[1] 0.6188467
> summary(irisVirgi.lm)$r.sq
[1] 0.1037537
```

Assim, em todos os casos, estes R^2 por espécie individual são muito mais baixos que o R^2 global correspondente ao modelo ANCOVA completo. Como se discutiu nas aulas, tal facto corresponde a uma situação em que uma ANOVA da variável resposta `Petal.Width` sobre um único factor `Species` tem um valor elevado da Soma de Quadrados correspondente ao ajustamento do modelo, ou seja, SQF elevado. Por outras palavras, o valor elevado de $R^2 = 0.9477$ no modelo ANCOVA resulta do facto de ao factor espécie corresponderem larguras médias das pétalas bastante diferentes, e não tanto ao valor preditivo do preditor numérico `Petal.Length`. A tradução prática desse facto é visível na nuvem de pontos original, se repararmos que a forte relação linear global tem sobretudo a que ver com a separação entre os três grupos de observações correspondentes a cada espécie, e não tanto com relações lineares fortes entre as duas medições das pétalas no seio de cada espécie. Por outras palavras, a relação linear tão prometedora que parece existir entre largura e comprimento das pétalas, na nuvem da totalidade das $n = 150$ observações, é em certo sentido uma ilusão resultante de se ter considerado em conjunto as três espécies.

(e) Nas aulas foi vista a fórmula que relaciona o valor de R^2 global do modelo ANCOVA com os R^2 e as Somas de Quadrados Totais para cada subconjunto de observações (por espécie), bem como o valor de SQF na ANOVA a um factor relacionando `Petal.Width` e o factor `Species`. A fórmula é

$$R^2 \;=\; \frac{\sum_{i=1}^{s} R_i^2\, SQT_i \;+\; SQF}{\sum_{i=1}^{s} SQT_i \;+\; SQF}.$$

O valor de SQF pode obter-se da seguinte forma:

```
> summary(aov(Petal.Width ~ Species, data=iris))
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  80.41   40.21     960 <2e-16 ***
Residuals   147   6.16    0.04
```

Por outro lado, os valores de SQT_i podem ser obtidos como o numerador das variâncias dos valores observados das larguras de pétalas em cada espécie. Tem-se $SQT_1 = 49 \times s^2_{y_1} = 0.5442$; $SQT_2 = 49 \times s^2_{y_2} = 1.9162$ e $SQT_3 = 49 \times s^2_{y_3} = 3.6962$. Logo,

$$R^2 = \frac{(0.1099785 \times 0.5442) + (0.6188467 \times 1.9162) + (0.1037537 \times 3.6962) + 80.41}{(0.5442 + 1.9162 + 3.6962) + 80.41} = 0.9477001\ .$$

Como se pode constatar, o valor de SQF sobrepõe-se ao das restantes parcelas, quer no numerador, quer no denominador, gerando um valor muito elevado do coeficiente de determinação global do modelo ANCOVA, que não corresponde a valores elevados de R^2 em

nenhuma das regressões individuais de cada espécie. Confirma-se que a interpretação dos valores de R^2 em modelos ANCOVA deve ser feita com cuidado.