
INSTITUTO SUPERIOR DE AGRONOMIA
ESTATÍSTICA E DELINEAMENTO – 2016-17
Resoluções dos exercícios de Regressão Linear Múltipla

1. Proceda como indicado no enunciado para ter disponível a *data frame* `vinho.RLM`.

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

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

Trata-se dum péssimo ajustamento, o que não surpreende, tendo em conta a nuvem de pontos deste par de variáveis, obtida na alínea anterior. O coeficiente de determinação é quase nulo: $R^2 = 0.05608$ e menos de 6% da variabilidade no teor de flavonóides é explicado pela regressão sobre o teor alcoólico.

No entanto, a hipótese nula do teste de ajustamento global ($H_0 : \mathcal{R}^2 = 0$ ou, alternativamente, $H_0 : \beta_1 = 0$) é rejeitada: o seu *p-value* é apenas $p = 0.00146$ (valor que tanto pode ser lido na última linha da listagem produzida pelo comando `summary` como na linha do teste-*t* à hipótese $\beta_1 = 0$). Ou seja, um coeficiente de determinação tão baixo quanto $R^2 = 0.05608$ é considerado significativamente diferente de zero (em boa parte, devido à grande dimensão da amostra). Mas isso não é sinónimo de um bom ajustamento do modelo. Como sempre, a Soma de Quadrados Total é o numerador da variância amostral dos valores observados da variável resposta. Ora,

```
> var(vinho.RLM$V8)
[1] 0.9977187
> dim(vinho.RLM)
[1] 178  13
> 177*var(vinho.RLM$V8)
[1] 176.5962
> 177*var(fitted(lm(V8 ~ V2 , data=vinho.RLM)))
[1] 9.903747
> 177*var(residuals(lm(V8 ~ V2 , data=vinho.RLM)))
[1] 166.6925
```

pelo que $SQT = (n - 1) s_y^2 = 176.5962$; $SQR = (n - 1) s_y^2 = 9.903747$; e $SQRE = (n - 1) s_e^2 = 166.6925$.

NOTA: Há outras maneiras possíveis de determinar estas Somas de Quadrados. Por exemplo, $SQR = R^2 \times SQT = 0.05608 \times 176.5962 = 9.903515$ (com um pequeno erro de arredondamento) e $SQRE = SQT - SQR = 176.5962 - 9.903515 = 166.6927$.

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

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

Analisando a coluna (ou linha) relativa à variável resposta **V8**, observa-se que a variável com a qual esta se encontra mais correlacionada (em módulo) é **V7** ($r_{7,8} = 0.86$), o que confirma a inspeção visual feita na alínea 1a. Assim, o coeficiente de determinação numa regressão de **V8** sobre **V7** é $R^2 = 0.8645635^2 = 0.74747$, ou seja, o conhecimento do índice de fenóis totais permite, através da regressão ajustada, explicar cerca de 75% da variabilidade total do teor de flavonóides. O valor de $SQT = 176.5962$ é igual ao obtido na alínea anterior, uma vez que diz apenas respeito à variabilidade da variável resposta (não dependendo do modelo de regressão ajustado). Já o valor de SQR vem alterado e é agora: $SQR = R^2 \cdot SQT = 132.0004$, sendo $SQRE = SQT - SQR = 176.5962 - 132.0004 = 44.5958$.

(d) O modelo pedido no enunciado é:

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

> summary(lm(V8 ~ V4 + V5 + V11 + V12 + V13 , data=vinho.RLM))
(...)
Multiple R-squared:  0.7144
(...)
```

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

(e) Ajustando a mesma variável resposta **V8** sobre a totalidade das restantes variáveis obtêm-se os seguintes resultados:

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

Call:
lm(formula = V8 ~ . , data = vinho.RLM)

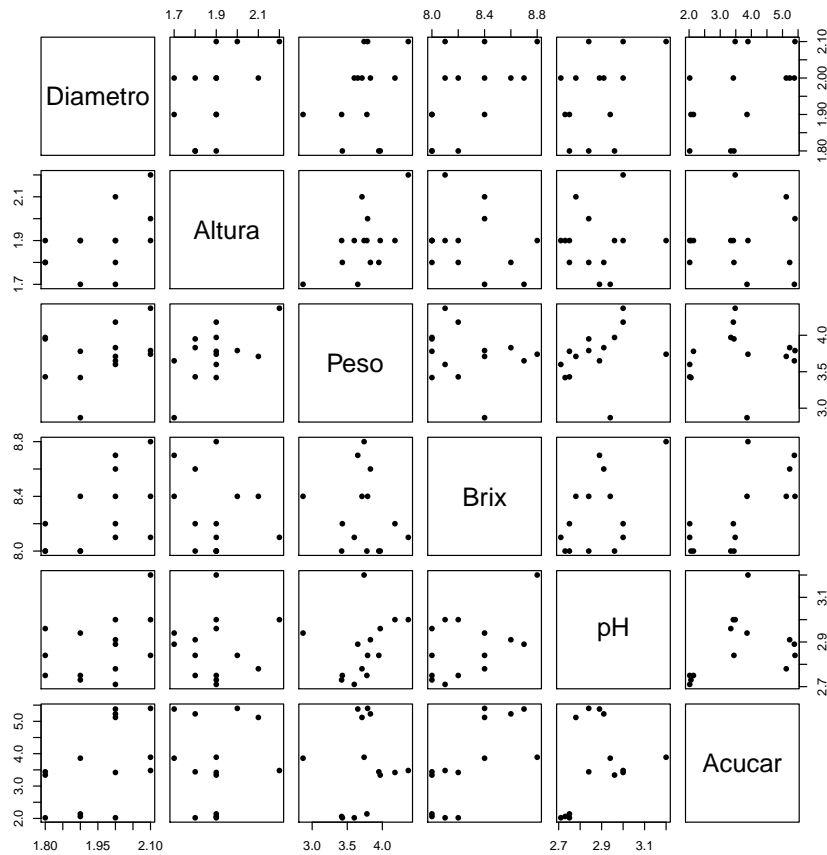
Coefficients:
(Intercept)          V2          V3          V4          V5          V6          V7
  -1.333e+00   4.835e-03  -4.215e-02   4.931e-01  -2.325e-02  -3.559e-03   7.058e-01
          V9          V10          V11          V12          V13          V14
```

-1.000e+00 2.840e-01 1.068e-04 4.387e-01 3.208e-01 9.557e-05

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

- i. De novo, o valor da Soma de Quadrados Total já é conhecido das alíneas anteriores: não depende do modelo ajustado, mas apenas da variância dos valores observados de Y ($V8$, neste exercício), que não se alteraram. Logo, $SQT = 176.5962$. Como se pode deduzir da listagem acima, $SQR = (n-1) \cdot s_y^2 = 151.4666$ e $SQRE = (n-1) \cdot s_e^2 = 25.12269$. 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*.

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



```
> round(cor(brix),d=3)
      Diametro Altura  Peso  Brix  pH Acucar
Diametro  1.000  0.488  0.302  0.557  0.411  0.492
Altura    0.488  1.000  0.587 -0.247  0.048  0.023
Peso      0.302  0.587  1.000 -0.198  0.308  0.118
Brix      0.557 -0.247 -0.198  1.000  0.509  0.714
pH        0.411  0.048  0.308  0.509  1.000  0.353
Acucar    0.492  0.023  0.118  0.714  0.353  1.000
```

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

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

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

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

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

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

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

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

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

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

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

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

- (f) Num contexto descritivo, a discussão da qualidade deste ajustamento faz-se com base no coeficiente de determinação $R^2 = \frac{SQR}{SQT}$. Pode calcular-se a Soma de Quadrados Total como o numerador da variância dos valores observados y_i de teor brix: $SQT = (n - 1) s_y^2 = 13 \times 0.07565934 = 0.9835714$. A Soma de Quadrados da Regressão é calculada de forma análoga à anterior, mas com base na variância dos valores ajustados \hat{y}_i , obtidos a partir da regressão ajustada: $SQR = (n - 1) s_{\hat{y}}^2 = 13 \times 0.06417822 = 0.8343169$. 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 predictoras.

O vector $\vec{\mathbf{b}}$ dos $p+1$ parâmetros ajustados é dado pelo produto matricial do enunciado: $\vec{\mathbf{b}} = (\mathbf{X}^t \mathbf{X})^{-1} (\mathbf{X}^t \vec{\mathbf{y}})$. 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 $\vec{\mathbf{b}}$ 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`.

3. Começemos por recordar alguns resultados já previamente discutidos:

- Viu-se no Exercício 3b) da Regressão Linear Simples que, para qualquer conjunto de n pares de observações se tem: $(n-1) \text{cov}_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})y_i$. Distribuindo y_i e o somatório pela diferença, tem-se:

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

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

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

- O produto de matrizes AB só é possível quando o número de colunas da matriz A for 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 : $(AB)_{ij} = (a_{i1} \ a_{i2} \ \dots \ a_{iq}) \begin{pmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{qj} \end{pmatrix} = \sum_{k=1}^q a_{ik} b_{kj}$.

- O produto interno de dois vectores n -dimensionais \vec{x} e \vec{y} é dado por $\vec{x}^t \vec{y} = \sum_{i=1}^n x_i y_i$. No caso de um dos vectores ser o vector de n uns, $\vec{\mathbf{1}}_n$, o produto interno resulta na soma dos elementos do outro vector, ou seja, em n vezes a média dos elementos do outro vector:

$$\vec{\mathbf{1}}_n^t \vec{x} = \sum_{i=1}^n x_i = n\bar{x}.$$

- A *matriz inversa* dum 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:

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

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 dum regressão linear simples ($p=1$), significa $n \times 2$. Tem uma primeira coluna de uns (o vector $\vec{\mathbf{1}}_n$) e uma segunda coluna com os n valores observados da variável preditora x , coluna essa que designamos pelo vector

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

(d) Multiplicando a matriz $(\mathbf{X}^t \mathbf{X})^{-1}$ pela variância σ^2 dos erros aleatórios obtém-se a matriz

$$\sigma^2 (\mathbf{X}^t \mathbf{X})^{-1} = \begin{bmatrix} \sigma^2 \frac{(n-1)s_x^2 + n\bar{x}^2}{n(n-1)s_x^2} & \frac{-n\bar{x}\sigma^2}{n(n-1)s_x^2} \\ \frac{-n\bar{x}\sigma^2}{n(n-1)s_x^2} & \frac{n\sigma^2}{n(n-1)s_x^2} \end{bmatrix} = \begin{bmatrix} \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right] & \frac{-\bar{x}\sigma^2}{(n-1)s_x^2} \\ \frac{-\bar{x}\sigma^2}{(n-1)s_x^2} & \frac{\sigma^2}{(n-1)s_x^2} \end{bmatrix}$$

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 $\text{Cov}[\hat{\beta}_0, \hat{\beta}_1] = \text{Cov}[\hat{\beta}_1, \hat{\beta}_0]$.

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

$$\begin{aligned} (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \vec{y} &= \frac{1}{n(n-1)s_x^2} \begin{bmatrix} (n-1)s_x^2 + n\bar{x}^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \begin{bmatrix} n\bar{y} \\ (n-1)\text{cov}_{xy} + n\bar{x}\bar{y} \end{bmatrix} \\ &= \frac{1}{n(n-1)s_x^2} \begin{bmatrix} (n-1)s_x^2 n\bar{y} + n^2 \bar{x}^2 \bar{y} - n\bar{x}(n-1)\text{cov}_{xy} - n^2 \bar{x}^2 \bar{y} \\ -n^2 \bar{x} \bar{y} + n(n-1)\text{cov}_{xy} + n^2 \bar{x} \bar{y} \end{bmatrix} \\ &= \begin{bmatrix} \frac{n(n-1)s_x^2 \bar{y}}{n(n-1)s_x^2} - \frac{n(n-1)\text{cov}_{xy} \bar{x}}{n(n-1)s_x^2} \\ \frac{n(n-1)\text{cov}_{xy}}{n(n-1)s_x^2} \end{bmatrix} = \begin{bmatrix} \bar{y} - b_1 \bar{x} \\ b_1 \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}. \end{aligned}$$

4. 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 $\mathbf{X}^t \mathbf{X}$:

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

A simetria resulta de três propriedades conhecidas de matrizes: a transposta de uma matriz transposta é a matriz original ($(\mathbf{A}^t)^t = \mathbf{A}$); a transposta de um produto de matrizes é o produto das correspondentes transpostas, pela ordem inversa ($(\mathbf{AB})^t = \mathbf{B}^t \mathbf{A}^t$); e a transposta de uma 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 teóricas, 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 $\mathbf{X}\mathbf{a}$, 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

$$\mathbf{HX}\mathbf{a} = \mathbf{X}(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t (\mathbf{X}\mathbf{a}) = \mathbf{X}(\mathbf{X}^t \mathbf{X})^{-1} (\mathbf{X}^t \mathbf{X}) \mathbf{a} = \mathbf{X}\mathbf{a} .$$

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

- (c) Por definição, o vector dos valores ajustados é dado por $\hat{\mathbf{y}} = \mathbf{H}\vec{\mathbf{y}}$. Ora, a média desses valores ajustados, que podemos representar por $\bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^n \hat{y}_i$, pode ser calculado tomando o produto interno do vector $\vec{\mathbf{1}}_n$ de n uns com o vector $\hat{\mathbf{y}}$, uma vez que esse produto interno devolve a soma dos elementos de $\hat{\mathbf{y}}$. Assim, a média dos valores ajustados é $\bar{\hat{y}} = \frac{1}{n} \vec{\mathbf{1}}_n^t \hat{\mathbf{y}} = \frac{1}{n} \vec{\mathbf{1}}_n^t \mathbf{H}\vec{\mathbf{y}} = \frac{1}{n} (\mathbf{H}\vec{\mathbf{1}}_n)^t \vec{\mathbf{y}} = \frac{1}{n} \vec{\mathbf{1}}_n^t \vec{\mathbf{y}}$, uma vez que $\mathbf{H}\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n$. Esta última afirmação decorre directamente da alínea anterior, uma vez que o vector $\vec{\mathbf{1}}_n$ pertence ao subespaço das colunas de \mathbf{X} , sendo a primeira das colunas dessa matriz (usando a notação da alínea anterior, tem-se $\vec{\mathbf{1}}_n = \mathbf{X}\mathbf{a}$ com $\mathbf{a} = (1, 0, 0, \dots, 0)^t$). Mas a expressão final obtida, $\frac{1}{n} \vec{\mathbf{1}}_n^t \vec{\mathbf{y}}$, é a média \bar{y} dos valores observados de Y (já que $\vec{\mathbf{1}}_n^t \vec{\mathbf{y}}$ devolve a soma dos elementos do vector dessas observações, $\vec{\mathbf{y}}$). Assim, também na regressão linear múltipla, valores observados de Y e correspondentes valores ajustados partilham o mesmo valor médio.

5. Seja $\vec{\mathbf{W}} = (W_1, W_2, \dots, W_k)^t$. 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)

$$E[\alpha \vec{\mathbf{W}}] = \begin{bmatrix} E[\alpha W_1] \\ E[\alpha W_2] \\ \vdots \\ E[\alpha W_k] \end{bmatrix} = \begin{bmatrix} \alpha E[W_1] \\ \alpha E[W_2] \\ \vdots \\ \alpha E[W_k] \end{bmatrix} = \alpha E[\vec{\mathbf{W}}] .$$

- (b)

$$E[\vec{\mathbf{W}} + \mathbf{a}] = \begin{bmatrix} E[W_1 + a_1] \\ E[W_2 + a_2] \\ \vdots \\ E[W_k + a_k] \end{bmatrix} = \begin{bmatrix} E[W_1] + a_1 \\ E[W_2] + a_2 \\ \vdots \\ E[W_k] + a_k \end{bmatrix} = \begin{bmatrix} E[W_1] \\ E[W_2] \\ \vdots \\ E[W_k] \end{bmatrix} + \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix} = E[\vec{\mathbf{W}}] + \mathbf{a} .$$

(c)

$$\begin{aligned} V[\alpha \vec{W}] &= \begin{bmatrix} V[\alpha W_1] & Cov[\alpha W_1, \alpha W_2] & \cdots & Cov[\alpha W_1, \alpha W_k] \\ Cov[\alpha W_2, \alpha W_1] & V[\alpha W_2] & \cdots & Cov[\alpha W_2, \alpha W_k] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[\alpha W_k, \alpha W_1] & Cov[\alpha W_k, \alpha W_2] & \cdots & V[\alpha W_k] \end{bmatrix} \\ &= \begin{bmatrix} \alpha^2 V[W_1] & \alpha^2 Cov[W_1, W_2] & \cdots & \alpha^2 Cov[W_1, W_k] \\ \alpha^2 Cov[W_2, W_1] & \alpha^2 V[W_2] & \cdots & \alpha^2 Cov[W_2, W_k] \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^2 Cov[W_k, W_1] & \alpha^2 Cov[W_k, W_2] & \cdots & \alpha^2 V[W_k] \end{bmatrix} = \alpha^2 V[\vec{W}] \end{aligned}$$

(d)

$$\begin{aligned} V[\vec{W} + \vec{a}] &= \begin{bmatrix} V[W_1 + a_1] & Cov[W_1 + a_1, W_2 + a_2] & \cdots & Cov[W_1 + a_1, W_k + a_k] \\ Cov[W_2 + a_2, W_1 + a_1] & V[W_2 + a_2] & \cdots & Cov[W_2 + a_2, W_k + a_k] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[W_k + a_k, W_1 + a_1] & Cov[W_k + a_k, W_2 + a_2] & \cdots & V[W_k + a_k] \end{bmatrix} \\ &= \begin{bmatrix} V[W_1] & Cov[W_1, W_2] & \cdots & Cov[W_1, W_k] \\ Cov[W_2, W_1] & V[W_2] & \cdots & Cov[W_2, W_k] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[W_k, W_1] & Cov[W_k, W_2] & \cdots & V[W_k] \end{bmatrix} = V[\vec{W}] \end{aligned}$$

(e)

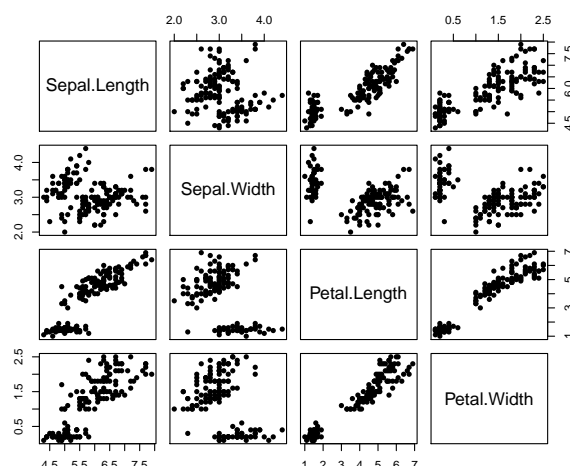
$$E[\vec{W} + \vec{U}] = \begin{bmatrix} E[W_1 + U_1] \\ E[W_2 + U_2] \\ \vdots \\ E[W_k + U_k] \end{bmatrix} = \begin{bmatrix} E[W_1] + E[U_1] \\ E[W_2] + E[U_2] \\ \vdots \\ E[W_k] + E[U_k] \end{bmatrix} = \begin{bmatrix} E[W_1] \\ E[W_2] \\ \vdots \\ E[W_k] \end{bmatrix} + \begin{bmatrix} E[U_1] \\ E[U_2] \\ \vdots \\ E[U_k] \end{bmatrix} = E[\vec{W}] + E[\vec{U}].$$

6. Trata-se dum conjunto de dados utilizado para ilustrar vários acetatos das aulas teóricas. Algumas alíneas desta pergunta encontram-se resolvidas nesses acetatos.

A *data frame* `iris` tem, na sua quinta e última coluna, o factor com o nome da espécie de lírio a que cada observação diz respeito. Neste exercício essa informação não é utilizada.

(a) O comando a usar no R para produzir as nuvens de pontos pedidas é:

```
> plot(iris[, -5], pch=16)
```



A relação linear da variável resposta `Petal.Width` com o preditor `Petal.Length` é (como sabemos do estudo deste conjunto de dados nos exercícios de regressão linear simples) bastante forte. Não parece existir uma relação linear tão forte da largura da pétala com qualquer das medições relativas às sépalas (embora a relação linear com o comprimento das sépalas não seja de desprezar). Isso não significa, só por si, que a introdução desses dois novos preditores não possa melhorar consideravelmente o ajustamento.

(b) Tem-se:

```
> iris2.lm <- lm(Petal.Width ~ Petal.Length + Sepal.Length + Sepal.Width , data=iris)
> iris2.lm
(...)
Coefficients:
 (Intercept) Petal.Length Sepal.Length Sepal.Width
      -0.2403      0.5241      -0.2073      0.2228

> summary(iris2.lm)
(...)
Multiple R-squared: 0.9379
(...)
```

Assim, o hiperplano ajustado tem equação $y = -0.2403 + 0.5241 x_1 - 0.2073 x_2 + 0.2228 x_3$, onde y indica a largura da pétala, x_1 indica o respectivo comprimento, x_2 indica o comprimento da sépala e x_3 a respectiva largura.

Já sabíamos (Exercícios 8 e 14 da RLS) que o coeficiente de determinação da regressão linear simples da largura das pétalas sobre o seu comprimento era $R^2 = 0.9271$. O novo valor $R^2 = 0.9379$ é superior, como teria de obrigatoriamente ser num modelo em que se acrescentaram preditores, mas não muito superior. Trata-se, de qualquer forma, dum valor muito elevado, sugerindo que se trata dum bom modelo linear.

(c) Qualquer coeficiente ajustado b_j , associado a uma variável preditora X_j , pode ser interpretado como a variação média na variável resposta Y , correspondente a aumentar X_j em uma unidade e mantendo os restantes preditores constantes. Assim, e tendo em conta os valores de b_1 , b_2 e b_3 obtidos na alínea anterior, a variação média na largura da pétala dum lírio, mantendo as restantes variáveis constantes, será:

- um acréscimo de 0.5241 cm por cada 1 cm a mais no comprimento da pétala;

- um decréscimo de 0.2073 cm por cada 1 cm a mais no comprimento da sépala;
- um acréscimo de 0.2228 cm por cada 1 cm a mais na largura da sépala.

Em relação à constante aditiva $b_0 = -0.2403$, trata-se dum valor que neste exercício tem pouco interesse prático. Interpreta-se da seguinte forma: num lírio com comprimento de pétala nulo, e largura e comprimento de sépala igualmente nulos, a largura média da pétala seria -0.2403 cm. A impossibilidade física deste valor sublinha que não faria sentido tentar aplicar este modelo a esse conjunto de valores nulos dos preditores, não apenas porque se trata de valores fora da gama de valores observados no ajustamento do modelo, mas sobretudo porque não faria sentido tentar utilizar este modelo para essa situação biologicamente impossível. Neste caso, deve pensar-se no valor de b_0 apenas como um auxiliar para obter um melhor ajustamento do modelo na região de valores que foram efectivamente observados.

- (d) Olhando novamente para a nuvem de pontos de `Petal.Width` contra `Sepal.Length`, verificamos a existência duma relação linear crescente (embora não muito forte). Como tal, a recta de regressão ajustada de largura da pétala sobre comprimento da sépala terá de ter um declive positivo. No entanto, o coeficiente associado ao preditor `Sepal.Length` na regressão linear múltipla agora ajustada é negativo: $b_2 = -0.2073$. Não se trata duma contradição. O modelo de regressão linear múltipla contém, além do preditor comprimento da sépala, outros dois preditores (largura da sépala e comprimento da pétala), que contribuem para a formação do valor ajustado de y . Na presença desses dois preditores, a contribuição do comprimento da sépala deve ter um sinal negativo. Esta aparente contradição sublinha uma ideia importante: *a introdução (ou retirada) de preditores numa regressão linear têm efeitos sobre todos os parâmetros, não sendo possível prever qual será a equação ajustada sem refazer as contas do ajustamento*. Em particular, repare-se que, embora a equação ajustada com os três preditores seja $PW = -0.2403 + 0.5241 PL - 0.2073 SL + 0.2228 SW$ (sendo as variáveis indicadas pelas iniciais dos seus nomes na *data frame iris*), *não é verdade* que a recta de regressão, apenas com o preditor comprimento da sépala, tenha equação $PW = -0.2403 - 0.2073 SL$ (nem tal faria sentido, pois desta forma todas as larguras de pétala ajustadas seriam negativas!). Ajustando directamente a regressão linear simples de largura da pétala sobre comprimento da sépala verifica-se que essa equação é bastante diferente: $PW = -3.2002 + 0.7529SL$.
- (e) Sabemos que a expressão genérica para os IC a $(1 - \alpha) \times 100\%$ para qualquer parâmetro β_j ($j = 0, 1, 2, \dots, p$) é:

$$\left[b_j - t_{\alpha/2[n-(p+1)]} \cdot \hat{\sigma}_{\hat{\beta}_j} \quad , \quad b_j + t_{\alpha/2[n-(p+1)]} \cdot \hat{\sigma}_{\hat{\beta}_j} \quad \right] .$$

Os valores estimados b_j e os erros padrões associados, $\hat{\sigma}_{\hat{\beta}_j}$, obtêm-se a partir das primeira e segunda colunas da tabela do ajustamento produzida pelo R:

```
> summary(iris2.lm)
(...)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.24031    0.17837  -1.347    0.18
Petal.Length  0.52408    0.02449  21.399 < 2e-16 ***
Sepal.Length -0.20727    0.04751  -4.363 2.41e-05 ***
Sepal.Width  0.22283    0.04894   4.553 1.10e-05 ***
---
Residual standard error: 0.192 on 146 degrees of freedom
```

Multiple R-squared: 0.9379, Adjusted R-squared: 0.9366
F-statistic: 734.4 on 3 and 146 DF, p-value: < 2.2e-16

Para intervalos de confiança a 95% precisamos do valor $t_{0.025(146)} = 1.976346$. Assim, o intervalo de confiança para β_1 é dado por:

$$] 0.52408 - 1.976346 \times 0.02449, 0.52408 + 1.976346 \times 0.02449 [=] 0.4756793, 0.5724807 [.$$

Analogamente, o intervalo a 95% de confiança para β_2 é dado por:

$$] -0.20727 - 1.976346 \times 0.04751, -0.20727 + 1.976346 \times 0.04751 [=] -0.3011662, -0.1133738 [.$$

Finalmente, o intervalo a 95% de confiança para β_3 é dado por:

$$] 0.22283 - 1.976346 \times 0.04894, 0.22283 + 1.976346 \times 0.04894 [=] 0.1261076, 0.3195524 [.$$

Com o auxílio do comando `confint` do R, podemos obter estes intervalos de confiança duma só assentada (as pequenas diferenças devem-se aos arredondamento usados acima):

```
> confint(iris2.lm)
              2.5 %      97.5 %
(Intercept) -0.5928277  0.1122129
Petal.Length  0.4756798  0.5724865
Sepal.Length -0.3011547 -0.1133775
Sepal.Width   0.1261101  0.3195470
```

Trata-se, no geral, de intervalos razoavelmente precisos (de pequena amplitude), para 95% de confiança. A interpretação do primeiro destes intervalos faz-se da seguinte forma: temos 95% de confiança em como o verdadeiro valor de β_1 está compreendido entre 0.4757 e 0.5725. Os outros dois intervalos interpretam-se de forma análoga.

- (f) A frase do enunciado traduz-se por: “teste se é admissível considerar que $\beta_2 < 0$ ”. Trata-se dum teste de hipóteses do tipo unilateral. Coloca-se a questão de saber se damos, ou não, o benefício da dúvida a esta hipótese. Se optarmos por exigir o ónus da prova a esta hipótese, teremos o seguinte teste:

Hipóteses: $H_0 : \beta_2 \geq 0$ vs. $H_1 : \beta_2 < 0$

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

Nível de significância: $\alpha = 0.05$.

Região Crítica: (Unilateral esquerda) Rejeitar H_0 se $T_{\text{calc}} < -t_{0.05(146)} \approx -1.6554$.

Conclusões: Tem-se $T_{\text{calc}} = \frac{b_2 - 0}{\hat{\sigma}_{\hat{\beta}_2}} = \frac{-0.20727 - 0}{0.04751} = -4.363 < -1.6554$. Assim, rejeita-se a hipótese nula (apesar de ter o benefício da dúvida) em favor de H_1 , ao nível de significância de 0.05, isto é, existe evidência experimental para considerar que a largura média das pétalas diminui, quando se aumenta o comprimento das sépalas, mantendo comprimento das pétalas e largura das sépalas constantes.

Duas notas sobre o teste acabado de efectuar:

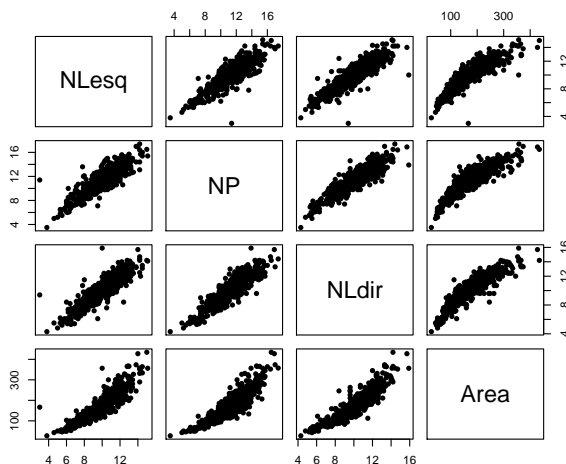
- Como o valor estimado de β_2 é negativo ($b_2 = -0.20727$) caso se tivesse dado o benefício da dúvida à hipótese $\beta_2 < 0$, nunca se poderia rejeitar essa hipótese;
- o valor da estatística é o indicado na terceira coluna da tabela produzida pelo R, mas o respectivo valor de prova não o é, uma vez que o *p-value* indicado na tabela corresponde a um teste bilateral. Para um teste unilateral esquerdo como o nosso, o valor de prova correspondente é dado por $p = P[t_{146} < -4.363] \approx 1.206 \times 10^{-5}$. Este valor é metade do *p-value* indicado na tabela.

7. 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 1.000000 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 $i = 1, 2, \dots, n$;
- 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: $H_0 : \beta_1 = 7$ vs. $H_1 : \beta_1 \neq 7$

Estatística do Teste: $T = \frac{\hat{\beta}_1 - 7}{\hat{\sigma}_{\beta_1}} \cap t_{(n-(p+1))}$, sob H_0 .

Nível de significância: $\alpha = 0.01$.

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

Conclusões: Tem-se $T_{calc} = \frac{b_1 - 0}{\hat{\sigma}_{\beta_1}} = \frac{9.987 - 7}{1.192} = 2.506 < 2.584$. 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 $|T_{calc}| > t_{0.025(596)} \approx 1.9640$. O valor da estatística de teste não se altera

($T_{\text{calc}} = 2.506$), 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 teóricas, é um caso particular dum teste a uma combinação linear dos parâmetros do modelo. Mais em pormenor, tem-se:

Hipóteses: $H_0 : \beta_2 - \beta_3 = 0$ vs. $H_1 : \beta_2 - \beta_3 \neq 0$

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 teóricas, 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 $\widehat{cov}[\hat{\beta}_2, \hat{\beta}_3]$, valores estes que surgem na matriz de (co)variâncias do estimador $\hat{\beta}$, que é estimada por $\hat{V}[\hat{\beta}] = QMRE(\mathbf{X}^t\mathbf{X})^{-1}$. 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 7d, obtêm-se os seguintes valores estimados:

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

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 $\hat{\sigma}_{\hat{\mu}_{Y|X}} = \sqrt{QMRE \cdot \vec{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{a}}$ onde os vectores \vec{a} são os vectores da forma $\vec{a} = (1, x_1, x_2, x_3)$. 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 \vec{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 $\hat{\sigma}_{indiv} = \sqrt{QMRE \cdot [1 + \vec{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{a}]}$. De novo, recorreremos 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
```

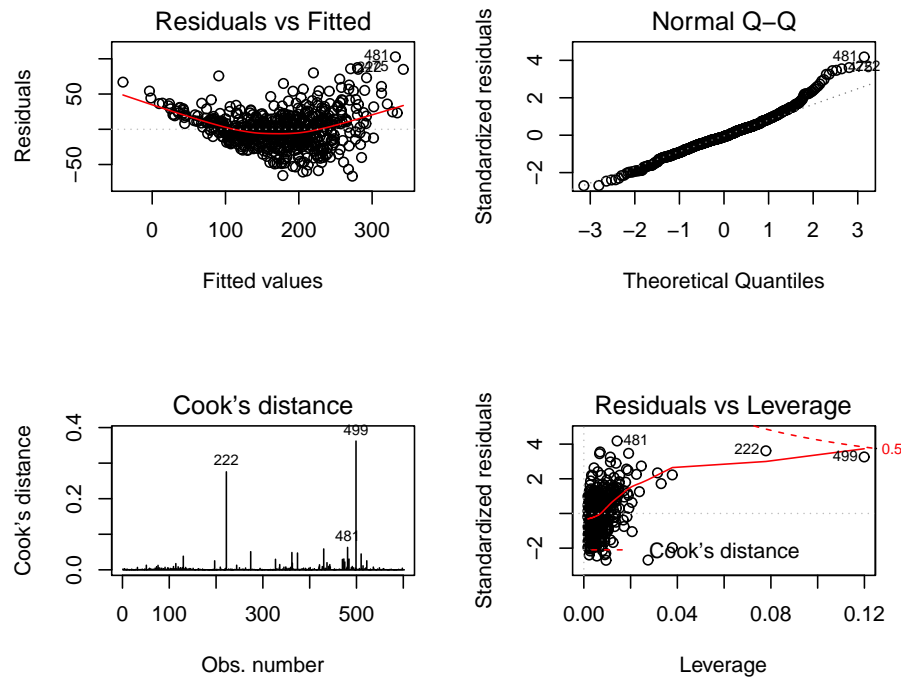
Assim, têm-se os seguintes intervalos de predição a 95% para os três valores de Y :

- *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 do hiperplano.

- (h) Recorreremos 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 $\bar{h} = \frac{4}{600} = 0.006667$ 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.*

8. (a) Os valores em falta são:

- $g.l.(QMRE) = n - (p + 1) = 39 - (3 + 1) = 35$
- $R_{\text{mod}}^2 = 1 - \frac{QMRE}{QMT} = 1 - \frac{\hat{\sigma}^2}{s_y^2} = 1 - \frac{13.62291^2}{502.7085} = 0.630824$
- $F_{\text{calc}} = \frac{n - (p + 1)}{p} \cdot \frac{R^2}{1 - R^2} = \frac{35}{3} \times \frac{0.66}{1 - 0.66} = 22.64706$
com $p = 3$ e $n - (p + 1) = 35$ graus de liberdade.

(b) O coeficiente de determinação deste modelo de regressão linear múltipla ($R^2 = 0.66$) não é muito bom: só 66% da variabilidade total do comprimento do élitro (variável resposta) é explicada por esta regressão.

Formalizando o teste de ajustamento global, relativo a parâmetros da equação de base do modelo, $Elytra_i = \beta_0 + \beta_1 TG_i + \beta_2 \text{Second.Antenna}_i + \beta_3 \text{Third.Antenna}_i + \epsilon_i$, vem:

Hipóteses: $H_0 : \beta_j = 0, \forall j \in \{1, 2, 3\}$ vs. $H_1 : \exists j \in \{1, 2, 3\}$, tal que $\beta_j \neq 0$.
[Modelo inútil] [Modelo não inútil]

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

Nível de significância: $\alpha = 0.05$.

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

Conclusões: Como $F_{\text{calc}} = 22.64706 > f_{0.05(3, 35)} \approx 2.88$, rejeita-se H_0 , isto é, rejeita-se a hipótese do modelo ser inútil, ao nível de significância de 0.05. Em alternativa, podia ter-se obtido esta mesma conclusão pela análise do valor de prova deste teste, presente no enunciado: como o p -value = $2.513 \times 10^{-08} < 0.05 = \alpha$, rejeita-se H_0 .

A validade deste teste depende da validade dos pressupostos adicionais admitidos no modelo de regressão linear múltipla, e que dizem respeito aos erros aleatórios. Concretamente, admite-se que esses erros aleatórios ϵ_i são Normais, de média zero, variância constante σ^2 e independentes.

(c) No contexto do problema em estudo e tendo em conta as unidades das várias variáveis, a estimativa do coeficiente associado à variável TG , $b_1 = 0.4874$, significa que se estima que o comprimento médio do élitro aumente $4.874 \mu m$ (0.4874 centésimas de milímetro) quando

a distância do sulco transversal à borda posterior do pró-torax aumenta 1 μm , mantendo-se os restantes preditores fixos.

- (d) Considerando que 10 $\mu m = 1$ centésima de milímetro, o que se pretende testar é se é admissível considerar que $\beta_2 < 1$. Dando o ónus da prova a esta hipótese, temos que:

Hipóteses: $H_0 : \beta_2 \geq 1$ vs. $H_1 : \beta_2 < 1$

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

Nível de significância: $\alpha = 0.05$.

Região Crítica: (Unilateral esquerda) Rejeitar H_0 se $T_{\text{calc}} < -t_{\alpha(n-(p+1))}$.

Conclusões: Como $T_{\text{calc}} = \frac{b_2 - 1}{\hat{\sigma}_{\hat{\beta}_2}} = \frac{0.9703 - 1}{0.1879} = -0.15806 > -t_{0.05(35)} \approx -1.6906$, não se rejeita H_0 ao nível de significância de 0.05, isto é, não há evidência experimental para considerar que, para cada micrómetro adicional no segundo segmento de antena, o comprimento do élitro aumenta, em média, menos de 10 μm .

- (e) É pedido um teste para averiguar a variação no valor médio da variável resposta **Elytra**, $\Delta E[Y]$, associada a aumentar simultaneamente cada uma das variáveis predictoras **Second.Antenna** (X_2) e **Third.Antenna** (X_3) numa unidade (1 μm). Ora,

$$E[Y | X_1 = x_1, X_2 = x_2 + 1, X_3 = x_3 + 1] = \beta_0 + \beta_1 x_1 + \beta_2(x_2 + 1) + \beta_3(x_3 + 1)$$

$$- E[Y | X_1 = x_1, X_2 = x_2, X_3 = x_3] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$$\Delta E[Y | \Delta X_1 = 0, \Delta X_2 = 1, \Delta X_3 = 1] = \beta_2 + \beta_3$$

Assim, o que se pretende testar é se $\beta_2 + \beta_3 = 1$. Formalizando, temos:

Hipóteses: $H_0 : \beta_2 + \beta_3 = 1$ vs. $H_1 : \beta_2 + \beta_3 \neq 1$

Estatística do Teste: $T = \frac{(\hat{\beta}_2 + \hat{\beta}_3) - 1}{\hat{\sigma}_{\hat{\beta}_2 + \hat{\beta}_3}} \cap t_{(n-(p+1))}$, 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: De acordo com os dados do enunciado,

$$\hat{\sigma}_{\hat{\beta}_2 + \hat{\beta}_3} = \sqrt{\hat{V}[\hat{\beta}_2 + \hat{\beta}_3]} = \sqrt{\hat{V}[\hat{\beta}_2] + \hat{V}[\hat{\beta}_3] + 2\widehat{Cov}[\hat{\beta}_2, \hat{\beta}_3]}$$

$$= \sqrt{0.035306802 + 0.0218088275 + 2 \times (-0.0162119398)} = 0.1571361,$$

pelo que $T_{\text{calc}} = \frac{b_2 + b_3 - 1}{\hat{\sigma}_{\hat{\beta}_2 + \hat{\beta}_3}} = \frac{0.9703 + 0.2923 - 1}{0.1571361} = 1.671163$. Como $|T_{\text{calc}}| < t_{0.025(35)} \approx 2.03$, não se rejeita H_0 ao nível de significância de 0.05, isto é, podemos admitir que o aumento simultâneo de 1 μm em cada um dos dois segmentos de antena, provoca um aumento de 10 μm no comprimento médio do élitro.

- (f) Pede-se um teste F parcial para comparar o

$$\text{Modelo completo (C)} \quad Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

com o

$$\text{Submodelo (S)} \quad Y = \beta_0 + \beta_2 x_2.$$

Hipóteses: $H_0 : \beta_1 = \beta_3 = 0$ vs. $H_1 : \beta_1 \neq 0 \vee \beta_3 \neq 0$
 [Submodelo OK] [Submodelo pior]

ou, de forma equivalente,

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

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

Nível de significância: $\alpha = 0.05$

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

Conclusões: Temos $n = 39$, $p = 3$, $k = 1$, $R_c^2 = 0.66$ e $R_s^2 = 0.7265072^2 = 0.5278127$ pois o submodelo considerado é a regressão linear simples de Elytra (Y) sobre Second. Antenna (X_2) pelo que o coeficiente de determinação do submodelo é igual ao quadrado do coeficiente de correlação entre estas duas variáveis, $R_s^2 = (r_{x_2y})^2$.

Logo, $F_{\text{calc}} = \frac{35}{3-1} \times \frac{0.66-0.5278127}{1-0.66} = 6.803758 > f_{0.05(2,35)} \approx 3.27$. Assim, rejeita-se H_0 , ou seja, modelo e submodelo diferem significativamente ao nível 0.05, pelo que é preferível trabalhar com o modelo com 3 preditores.

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

```
> summary(lm(y ~ . , data=milho))
[...]
```

	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: $R_{mod}^2 = 0.6488$. Este coeficiente modificado é definido como $R_{mod}^2 = 1 - \frac{QMRE}{QMT} = 1 - \frac{SQRE}{SQT} \cdot \frac{n-1}{n-(p+1)} = 1 - (1 - R^2) \cdot \frac{n-1}{n-(p+1)}$. 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_{mod}^2 , 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 $R_{mod}^2 = 1 - 0.2524 \times 1.3913 = 1 - 0.3512 = 0.6488$. Em geral, o R_{mod}^2 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_{mod}^2 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))
[...]
```

	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

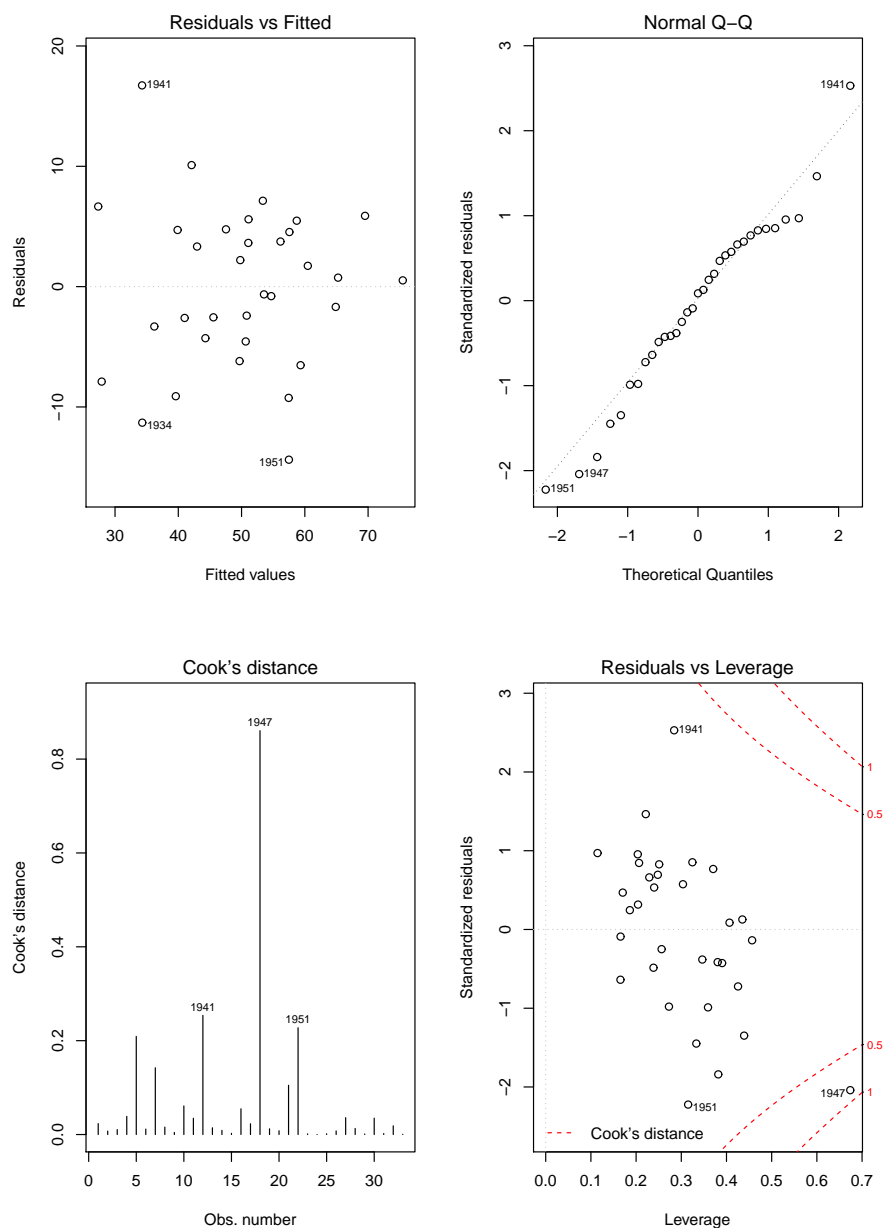
```
---
Residual standard error: 10.69 on 24 degrees of freedom
Multiple R-squared: 0.5074, Adjusted R-squared: 0.3432
F-statistic: 3.091 on 8 and 24 DF, p-value: 0.01524
```

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 $R_{mod}^2 = 0.3432$ 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 existir 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 homogeneidade de variâncias, admitidos no modelo RLM. Analogamente, no *qq-plot* comparando quantis teóricos duma Normal reduzida e quantis empíricos (canto superior direito), existe linearidade aproximada dos pontos, pelo que a hipótese de Normalidade dos erros aleatórios também parece admissível. Já no diagrama de barras das distâncias de Cook (canto inferior esquerdo) há um facto digno de registo: a observação correspondente ao ano 1947 tem um

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

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

- (d) Se no ajustamento do modelo com todos os preditores (x_1 a x_9) se efectuar um teste t às hipóteses $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$, estaremos a testar se é possível considerar equivalentes os dois modelos das alíneas anteriores, uma vez que esses modelos apenas diferem no preditor x_1 . A descrição pormenorizada dum tal teste já foi feita em resoluções de exercícios anteriores (por exemplo, no exercício 7e). Resumidamente, e observando o valor de prova que é dado na listagem referente a este teste, no modelo completo ($p = 0.000104$, associado ao valor calculado da estatística $t_{calc} = 4.678$), conclui-se pela rejeição de $H_0 : \beta_1 = 0$, para os níveis de significância usuais. Assim (e de forma nada surpreendente) conclui-se que modelo (com x_1) e submodelo (sem x_1) têm ajustamentos significativamente diferentes.

O mesmo problema de comparar modelo e submodelo pode ser abordado pela via dum teste F parcial. Neste contexto, temos:

Hipóteses: $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$
 [modelos equivalentes] [modelos diferentes]
 ou, de forma equivalente,
 $H_0 : \mathcal{R}_c^2 = \mathcal{R}_s^2$ vs. $H_1 : \mathcal{R}_c^2 > \mathcal{R}_s^2$

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

Nível de significância: $\alpha = 0.05$

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

Conclusões: Temos $n = 33$, $p = 9$, $k = 8$, $R_c^2 = 0.7476$ e $R_s^2 = 0.5074$.

Logo, $F_{calc} = \frac{23}{1} \times \frac{0.7476 - 0.5074}{1 - 0.7476} = 21.8827 > f_{0.05(1,23)} = 4.28$. Assim, rejeita-se H_0 , ou seja, modelo e submodelo diferem significativamente ao nível 0.05, pelo que é preferível trabalhar com o modelo com todos os preditores.

Este teste F parcial pode ser obtido no R através do comando `anova`, com o modelo completo ajustado guardado no objecto `milho.lm` e o submodelo sem x_1 no objecto `milhosx1.lm`:

```
> anova(milhosx1.lm, milho.lm)
Analysis of Variance Table
Model 1: y ~ x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
Model 2: y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
  Res.Df  RSS Df Sum of Sq    F    Pr(>F)
1      24 2741.2
2      23 1404.7  1    1336.5 21.883 0.0001039 ***
```

Além de se confirmar o valor calculado da estatística $F_{calc} = 21.883$, obtemos o valor de prova que lhe está associado: $p = 0.0001039$. Trata-se do mesmo p -value obtido no teste t considerado antes. Este facto não é uma coincidência. Quando modelo e submodelo diferem numa única variável, a estatística do teste F parcial é o quadrado da estatística t no teste a que $\beta_j = 0$ (tendo-se, no nosso caso, $t_{calc}^2 = (4.678)^2 = 21.88368 = F_{calc}$, aparte os erros de arredondamento). Os respectivos p -values têm de ser iguais pois (resultado estudado na disciplina de Estatística dos primeiros ciclos do ISA) se $T \cap t_\nu$, então $T^2 \cap F_{(1,\nu)}$. Trata-se de duas estatísticas de teste essencialmente equivalentes.

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

ou alternativamente,

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

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

Nível de significância: $\alpha = 0.05$

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

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

Logo, $F_{calc} = \frac{23}{4} \times \frac{0.7476 - 0.6435}{1 - 0.7476} = 2.371533 < f_{0.05(4,23)} = 2.78$. 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.

- (f) 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 9a), 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 mantém-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 (x_7 , x_8 e x_4) 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 (x_1 , x_2 , x_6 e x_9), 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 teóricas, 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, cujas características são discutidas mais em pormenor na resolução do Exercício 10, 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 teóricas, 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 predictoras. 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 x_1 , x_2 , x_6 e x_9 , 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.

(g) 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))
[...]
```

	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 9f), 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 predictoras (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.

10. (a) Os dados para este exercício estão na *data frame* de nome `trigo`, que estará disponível na sessão após executar-se a instrução `load` descrita no “Aviso” no início dos enunciados dos exercícios de regressão linear múltipla. Nesse caso, o modelo de regressão linear múltipla pedido é invocado, no R, através dos comandos:

```
> trigo.lm <- lm(Y ~ x1 + x2 + x3 + x4, data=trigo)
> summary(trigo.lm)
```

O ajustamento obtido é o seguinte:

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.7897576	0.7904169	2.264	0.0642 .
x1	-0.0020205	0.0010862	-1.860	0.1122
x2	0.0453405	0.0465081	0.975	0.3673
x3	-0.0010773	0.0021529	-0.500	0.6346
x4	0.0003520	0.0004524	0.778	0.4661

```
---
Residual standard error: 0.124 on 6 degrees of freedom
Multiple R-squared: 0.7949, Adjusted R-squared: 0.6582
F-statistic: 5.815 on 4 and 6 DF, p-value: 0.02919
```

O Coeficiente de Determinação é elevado ($R^2 = 0.7949$), com cerca de 79.5% da variabilidade dos valores observados de Y explicada pela regressão. Eis o teste de ajustamento global:

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 $F_{calc} > f_{\alpha(p,n-(p+1))} = f_{0.05(4,6)} = 4.53$.

Conclusões: O valor calculado da estatística é dado na listagem produzida pelo R, mas pode também ser calculado a partir do valor do coeficiente de determinação: $F_{calc} = \frac{6}{4} \times \frac{0.7949}{1-0.7949} \approx 5.8135$. Logo, rejeita-se (por pouco) 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: $p = 0.02919$ corresponde a uma rejeição ao nível de significância $\alpha = 0.05$, mas (apesar do valor bastante elevado de R^2) a uma não rejeição de H_0 ao nível $\alpha = 0.01$. Temos aqui uma situação que está no extremo oposto ao encontrado no Exercício 1b) da regressão linear múltipla. Nesse caso, um valor de $R^2 = 0.056$ era significativamente diferente de zero, ao nível de significância $\alpha = 0.01$. Neste caso, um valor de R^2 tão elevado quanto $R^2 = 0.7949$ não difere significativamente de zero, ao mesmo nível $\alpha = 0.01$. Uma parte importante da explicação para tal aparente contradição reside na relação entre o número de observações n usadas no ajustamento do modelo, e o número $p + 1$ de parâmetros do mesmo. Quando (como neste Exercício) $n \approx p + 1$ há pouca informação e o benefício da dúvida resultante dessa escassez de informação protege H_0 (excepto para valores de R^2 muito próximos de 1 - veja a expressão da estatística F acima). Pelo contrário, quando $n \gg p + 1$, a informação é muita e permite mais facilmente rejeitar a hipótese nula $\mathcal{R}^2 = 0$.

- (b) Pede-se um intervalo de confiança para β_2 , o parâmetro que multiplica a variável **x2** (temperatura média em Julho). Sabemos que esse intervalo de confiança, a $(1 - \alpha) \times 100\%$, é da forma:

$$\left[b_2 - t_{\frac{\alpha}{2}(n-(p+1))} \cdot \hat{\sigma}_{\hat{\beta}_2}, \quad b_2 + t_{\frac{\alpha}{2}(n-(p+1))} \cdot \hat{\sigma}_{\hat{\beta}_2} \right],$$

onde (tendo em conta a listagem anterior) $b_2 = 0.0453405$ e $\hat{\sigma}_{\hat{\beta}_2} = 0.0465081$. Uma vez que se pede um IC a 95% de confiança, tem-se ainda $t_{\frac{\alpha}{2}(n-(p+1))} = t_{0.025(6)} = 2.446912$. Assim, o IC pedido é: $]-0.06846, 0.15914[$. Tal como numa regressão linear simples, este intervalo de confiança (e os dos outros parâmetros β_j) podem ser obtidos no R, através do comando `confint`:

```
> confint(trigo.lm)
                2.5 %      97.5 %
(Intercept) -0.1443229969 3.7238381639
x1           -0.0046784122 0.0006374421
x2           -0.0684608295 0.1591417618
x3           -0.0063453276 0.0041907752
x4           -0.0007551014 0.0014590854
```

- (c) Qualquer variável cujo coeficiente associado, β_j , possa ser nulo é uma variável candidata a sair do modelo sem que essa exclusão afecte de forma significativa o ajustamento do mesmo. Por exemplo, na alínea anterior viu-se que o intervalo a 95% de confiança para β_2 contém o valor zero, pelo que é admissível que $\beta_2 = 0$. Desta forma, a variável **x2** é candidata a sair do modelo, para o nível de significância $\alpha = 0.05$, sem que com isso o ajustamento seja afectado de forma significativa. Na listagem produzida pelo R, no ajustamento de regressões lineares, as duas últimas colunas da tabela *Coefficients* contém informação relativa a testes às hipóteses $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ (para cada $j = 1, 2, 3, 4$). No nosso caso, é fácil concluir que, para o nível de significância $\alpha = 0.05$, não se rejeita a hipótese de qualquer dos β_j ser nulo. Assim, todas as variáveis predictoras são candidatas a sair sem que isso afecte de forma significativa o ajustamento. No entanto, apenas se pode excluir *uma* variável predictor (os testes dizem respeito ao modelo com todas as variáveis predictoras, não sendo possível prever os efeitos de excluir mais do que uma variável simultaneamente). Nesse caso, no algoritmo de exclusão sequencial exclui-se a variável cujo **p-value** (do teste bilateral a

$H_0 : \beta_j = 0$) seja maior (equivalentemente, cujo valor calculado da estatística, T_{calc} , esteja mais próximo de zero). No nosso caso, essa variável é x_3 . A exclusão dessa variável obriga a reajustar o modelo, apenas com os três preditores restantes:

```
> summary(lm(Y ~ x1 + x2 + x4, data=trigo))
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.5456031  0.5875900   2.630  0.0339 *
x1           -0.0023114  0.0008670  -2.666  0.0322 *
x2            0.0491501  0.0433545   1.134  0.2943
x4            0.0004975  0.0003276   1.518  0.1727
---
Residual standard error: 0.1171 on 7 degrees of freedom
Multiple R-squared: 0.7864, Adjusted R-squared: 0.6948
F-statistic: 8.589 on 3 and 7 DF, p-value: 0.009578
```

Repare-se como o ajustamento do submodelo sem o preditor x_3 introduz várias alterações importantes: os coeficientes das variáveis sobrantas (e a ordenada na origem b_0) mudaram. Para uma dessas variáveis predictoras (x_1) o coeficiente estimado $b_1 = -0.0023114$ é agora significativamente diferente de zero ao nível $\alpha = 0.05$ (com um valor de prova $p = 0.0322$), o que não acontecia para qualquer variável no modelo completo. Este facto também ilustra porque razão não se deve excluir mais do que uma variável de cada vez. No entanto, a qualidade global do ajustamento pouco mudou: o valor do coeficiente de determinação $R^2 = 0.7864$ é pouco diferente do registado para o modelo completo.

Neste novo ajustamento, ainda existem variáveis candidatas à exclusão, uma vez que há variáveis para as quais não se rejeita a hipótese nula de que o respectivo coeficiente seja nulo. Em particular, não se rejeita a hipótese $H_0 : \beta_2 = 0$, nem a hipótese $H_0 : \beta_4 = 0$. Destas duas candidatas a sair, excluimos a variável x_2 , cujo valor de prova é mais elevado ($p = 0.2943$). O novo submodelo ajustado é:

```
> summary(lm(Y ~ x1 + x4, data=trigo))
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.1410034  0.2681594   7.984 4.43e-05 ***
x1           -0.0021761  0.0008740  -2.490  0.03753 *
x4            0.0007731  0.0002235   3.459  0.00858 **
---
Residual standard error: 0.1192 on 8 degrees of freedom
Multiple R-squared: 0.7471, Adjusted R-squared: 0.6839
F-statistic: 11.82 on 2 and 8 DF, p-value: 0.004087
```

A exclusão da variável x_2 fez com a variável x_4 (que até agora era candidata a sair) tenha passado a ter um coeficiente significativamente diferente de zero (quer ao nível $\alpha = 0.05$, quer ao nível $\alpha = 0.01$), facto que volta a ilustrar que não é possível excluir vários preditores simultaneamente sem correr o risco de deitar fora preditores importantes. A qualidade do submodelo agora ajustado baixou um pouco ($R^2 = 0.7471$). O submodelo agora ajustado é o modelo final, no algoritmo de exclusão sequencial com nível de significância $\alpha = 0.05$, uma vez que para todos os preditores sobrantas se rejeita a hipótese $H_0 : \beta_j = 0$. Note-se que o submodelo final seria o mesmo caso se tivesse optado por trabalhar, ao longo dos vários passos do algoritmo, com um nível de significância $\alpha = 0.10$.

No R, o comando `step` corre algoritmos de selecção de submodelos, embora utilizando um diferente critério de exclusão de variáveis, baseado no Critério de Informação de Akaike (AIC). Em relação ao algoritmo baseado nos testes- t aos parâmetros β_j , acima ilustrado,

apenas pode diferir no ponto de paragem. No entanto, no caso deste exercício há total identidade de resultados, como se pode verificar:

```
> step(trigo.lm, direction="backward") <-- Comando do R
Start: AIC=-42.6 <-- AIC do modelo completo inicial (AIC
Y ~ x1 + x2 + x3 + x4 <-- pode ser negativo; quanto menor, melhor)
```

[O R ordena o modelo inicial, bem como os possíveis submodelos resultantes de excluir uma das variáveis preditoras, por ordem crescente de AIC. Nas listagens produzidas pelo R, ‘RSS’ indica a Soma de Quadrados Residual (SQRE) do modelo correspondente e "Sum of Sq" indica a diferença nessa Soma de Quadrados associada a cada possível exclusão de um preditor:]

	Df	Sum of Sq	RSS	AIC	
- x3	1	0.003847	0.096043	-44.149	<-- exclusão de x3 ; melhor AIC
- x4	1	0.009300	0.101496	-43.542	<-- exclusão de x4
- x2	1	0.014604	0.106800	-42.982	<-- exclusão de x2
<none>			0.092196	-42.599	<-- modelo inicial
- x1	1	0.053164	0.145360	-39.591	<-- exclusão de x1: AIC pior

[Excluída a variável x3, inicia-se novo passo, onde se ensaia a exclusão de cada uma das variáveis preditoras ainda presentes:]

```
Step: AIC=-44.15 <-- AIC do modelo actual
Y ~ x1 + x2 + x4 <-- modelo actual
```

	Df	Sum of Sq	RSS	AIC	
- x2	1	0.017634	0.113677	-44.295	<- exclusão de x2 melhora AIC
<none>			0.096043	-44.149	<- modelo actual
- x4	1	0.031636	0.127679	-43.017	<- exclusão de x4: AIC pior
- x1	1	0.097504	0.193547	-38.441	<- exclusão de x1: AIC pior

[Ajusta-se o novo modelo resultante da exclusão da variável x2, e inicia-se novo passo para estudar o efeito de excluir um dos dois preditores sobrantess:]

```
Step: AIC=-44.3 <-- AIC do modelo actual
Y ~ x1 + x4 <-- modelo actual
```

	Df	Sum of Sq	RSS	AIC	
<none>			0.11368	-44.295	<-- modelo actual: a melhor opção
- x1	1	0.088092	0.20177	-39.984	<-- exclusão de x1 aumenta AIC
- x4	1	0.170018	0.28369	-36.235	<-- exclusão de x4 aumenta AIC

[Submodelo final escolhido:]

```
Call: lm(formula = Y ~ x1 + x4, data = trigo)
```

Coefficients:

(Intercept)	x1	x4
2.1410034	-0.0021761	0.0007731

- (d) O melhor modelo de regressão linear simples será aquele que tenha, como único preditor a variável cujo coeficiente de correlação com Y fôr, em valor absoluto, maior. Analisemos a matriz de correlações entre todas as variáveis (arredondada a três casas decimais, para facilitar a leitura):

```
> round(cor(trigo),d=3)
      x1      x2      x3      x4      Y
```

```

x1  1.000 -0.087  0.490 -0.239 -0.607
x2 -0.087  1.000 -0.624  0.739  0.639
x3  0.490 -0.624  1.000 -0.808 -0.798
x4 -0.239  0.739 -0.808  1.000  0.742
Y   -0.607  0.639 -0.798  0.742  1.000

```

A variável mais correlacionada (em módulo) com Y é a variável x_3 . Curiosamente, foi a primeira variável a ser excluída pelo algoritmo de exclusão sequencial (este tipo de comportamentos estranhos são mais frequentes para conjuntos de dados em que - como é o nosso caso - o número de observações n é pequeno face ao número $p+1$ de parâmetros do modelo). O coeficiente de determinação do modelo de regressão linear simples com x_3 como único preditor é (aproximadamente) $R^2 = (-0.798)^2 = 0.637$. A título de curiosidade, registre-se que, sendo o melhor preditor individual, modelos com dois preditores dos quais um é x_3 nunca alcançam, neste exemplo, os valores do coeficiente de determinação do submodelo escolhido pelo algoritmo. Para o submodelo com preditores x_1 e x_3 tem-se $R^2 = 0.6980$; para o submodelo com preditores x_2 e x_3 tem-se $R^2 = 0.6695$; e para o submodelo com preditores x_4 e x_3 tem-se $R^2 = 0.6639$ ¹.

11. (a) Utilizamos o R para ajustar o modelo de regressão da variável V_8 sobre todas as restantes variáveis. Podemos utilizar um atalho na fórmula do R que consiste em substituir por um ponto a listagem de todas as restantes variáveis na *data frame* como preditores. Para executar o algoritmo de exclusão sequencial, vamos optar por um nível de significância $\alpha = 0.10$. A análise da tabela de resultados associada ao ajustamento do modelo completo sugere que há várias variáveis que podem ser excluídas do modelo sem afectar de forma significativa a qualidade do ajustamento, com destaque para a variável V_{11} , cujo p -value, no teste bilateral a $H_0: \beta_{11} = 0$, é enorme ($p = 0.99614$):

```

> summary(lm(V8 ~ . , data=vinho.RLM))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.333e+00  7.558e-01  -1.764  0.07956 .
V2           4.835e-03  5.667e-02   0.085  0.93211
V3          -4.215e-02  3.363e-02  -1.253  0.21191
V4           4.931e-01  1.533e-01   3.216  0.00156 **
V5          -2.325e-02  1.302e-02  -1.786  0.07591 .
V6          -3.559e-03  2.429e-03  -1.465  0.14487
V7           7.058e-01  8.062e-02   8.755 2.33e-15 ***
V9          -1.000e+00  3.061e-01  -3.267  0.00132 **
V10          2.840e-01  6.855e-02   4.143 5.47e-05 ***
V11          1.068e-04  2.201e-02   0.005  0.99614
V12          4.387e-01  2.021e-01   2.171  0.03137 *
V13          3.208e-01  7.639e-02   4.199 4.37e-05 ***
V14          9.557e-05  1.563e-04   0.611  0.54182
---
Residual standard error: 0.3902 on 165 degrees of freedom
Multiple R-squared:  0.8577, Adjusted R-squared:  0.8474
F-statistic:  82.9 on 12 and 165 DF,  p-value: < 2.2e-16

```

Procedemos agora a ajustar o submodelo resultante da exclusão do preditor V_{11} . O comando do R pode ainda aproveitar o atalho na fórmula, indicando (precedida de um sinal negativo) o nome da variável que queremos excluir:

¹Até poderia acontecer que algum destes modelos tivesse R^2 superior. Não há, à partida, garantias de que o algoritmo de exclusão sequencial seleccione o modelo de k preditores com melhor R^2 .

```
> summary(lm(V8 ~ . - V11 , data=vinho.RLM))
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.334e+00  7.427e-01  -1.796  0.07428 .
V2           4.951e-03  5.128e-02   0.097  0.92320
V3          -4.217e-02  3.315e-02  -1.272  0.20507
V4           4.931e-01  1.519e-01   3.246  0.00141 **
V5          -2.325e-02  1.297e-02  -1.792  0.07492 .
V6          -3.559e-03  2.422e-03  -1.469  0.14363
V7           7.059e-01  7.951e-02   8.878 1.06e-15 ***
V9          -1.000e+00  3.050e-01  -3.279  0.00127 **
V10          2.840e-01  6.735e-02   4.217 4.05e-05 ***
V12          4.383e-01  1.790e-01   2.449  0.01538 *
V13          3.206e-01  6.843e-02   4.686 5.78e-06 ***
V14          9.571e-05  1.531e-04   0.625  0.53264
---
Residual standard error: 0.389 on 166 degrees of freedom
Multiple R-squared: 0.8577, Adjusted R-squared: 0.8483
F-statistic: 90.99 on 11 and 166 DF, p-value: < 2.2e-16
```

Como se pode verificar, a exclusão do preditor V11 não afectou o coeficiente de determinação $R^2 = 0.8577$ (até à quarta casa decimal). Este facto reflecte o enorme valor de prova ($p = 0.99614$) da variável excluída. Este facto traduz-se também num *aumento* do coeficiente R^2 modificado ($R^2_{adjusted}$), algo que nunca pode acontecer ao coeficiente de determinação usual quando se passa dum modelo para um seu submodelo. Continuam a existir vários preditores candidatos à exclusão, com destaque para V2, cujo p -value no teste bilateral a $H_0 : \beta_2 = 0$ é também muito grande: $p = 0.9232$. Ajustando o modelo com menos esse preditor, obtemos:

```
> summary(lm(V8 ~ . - V11 - V2 , data=vinho.RLM))
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.2732556  0.3929674  -3.240  0.00144 **
V3          -0.0416084  0.0325333  -1.279  0.20269
V4           0.4947522  0.1505264   3.287  0.00124 **
V5          -0.0234559  0.0127473  -1.840  0.06753 .
V6          -0.0035598  0.0024149  -1.474  0.14233
V7           0.7067612  0.0787511   8.975 5.69e-16 ***
V9          -1.0009149  0.3039571  -3.293  0.00121 **
V10          0.2836573  0.0670345   4.232 3.82e-05 ***
V12          0.4354675  0.1760931   2.473  0.01440 *
V13          0.3201306  0.0680210   4.706 5.27e-06 ***
V14          0.0001031  0.0001320   0.781  0.43566
---
Residual standard error: 0.3879 on 167 degrees of freedom
Multiple R-squared: 0.8577, Adjusted R-squared: 0.8492
F-statistic: 100.7 on 10 and 167 DF, p-value: < 2.2e-16
```

Mais uma vez, o valor de R^2 não se altera, confirmando a dispensabilidade da variável agora excluída. Os valores de prova das variáveis presentes no modelo, embora já não tão elevados como em casos anteriores, sugerem que há ainda variáveis dispensáveis, com destaque para V14. Ajustando o novo submodelo resultante de excluir essa variável, obtem-se:

```
> summary(lm(V8 ~ . - V11 - V2 - V14, data=vinho.RLM))
Coefficients:
```

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.266062   0.392404  -3.226 0.001507 **
V3           -0.041829   0.032494  -1.287 0.199767
V4            0.540169   0.138693   3.895 0.000142 ***
V5           -0.028327   0.011107  -2.550 0.011652 *
V6           -0.003183   0.002363  -1.347 0.179889
V7            0.718705   0.077164   9.314 < 2e-16 ***
V9           -1.018104   0.302809  -3.362 0.000957 ***
V10           0.286785   0.066837   4.291 3.00e-05 ***
V12           0.439005   0.175831   2.497 0.013497 *
V13           0.316561   0.067789   4.670 6.14e-06 ***

```

```

Residual standard error: 0.3874 on 168 degrees of freedom
Multiple R-squared: 0.8572, Adjusted R-squared: 0.8496
F-statistic: 112.1 on 9 and 168 DF, p-value: < 2.2e-16

```

O coeficiente de determinação baixou, mas de forma muitíssimo ligeira, confirmando a dispensabilidade do preditor V14. Repare-se como o R^2 modificado continua a aumentar (embora de forma ligeira). O leque de preditores dispensáveis neste submodelo está reduzido a dois: V3 e V6. Excluimos a primeira, cujo valor de prova é superior, obtendo novo ajustamento:

```
> summary(lm(V8 ~ . - V11 -V2 - V14 - V3, data=vinho.RLM))
```

Coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.403877   0.378251  -3.711 0.000279 ***
V4            0.522565   0.138285   3.779 0.000218 ***
V5           -0.029014   0.011115  -2.610 0.009860 **
V6           -0.003151   0.002368  -1.331 0.185019
V7            0.727268   0.077026   9.442 < 2e-16 ***
V9           -1.056654   0.301909  -3.500 0.000595 ***
V10           0.285137   0.066955   4.259 3.41e-05 ***
V12           0.540246   0.157567   3.429 0.000762 ***
V13           0.313494   0.067878   4.618 7.63e-06 ***

```

```

Residual standard error: 0.3882 on 169 degrees of freedom
Multiple R-squared: 0.8558, Adjusted R-squared: 0.849
F-statistic: 125.4 on 8 and 169 DF, p-value: < 2.2e-16

```

De novo, há uma muito ligeira alteração no valor de R^2 . No submodelo agora ajustado há uma única variável candidata a sair: a variável V6. O modelo resultante de excluir esse preditor é:

```
> summary(lm(V8 ~ . - V11 -V2 - V14 - V3 - V6, data=vinho.RLM))
```

Coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.65205   0.32984  -5.009 1.36e-06 ***
V4            0.45247   0.12815   3.531 0.000533 ***
V5           -0.02662   0.01099  -2.422 0.016503 *
V7            0.72642   0.07720   9.410 < 2e-16 ***
V9           -0.93935   0.28941  -3.246 0.001411 **
V10           0.26873   0.06596   4.074 7.07e-05 ***
V12           0.52973   0.15772   3.359 0.000967 ***
V13           0.33218   0.06656   4.991 1.48e-06 ***

```

Residual standard error: 0.3891 on 170 degrees of freedom
 Multiple R-squared: 0.8543, Adjusted R-squared: 0.8483
 F-statistic: 142.4 on 7 and 170 DF, p-value: < 2.2e-16

Este novo submodelo já não tem variáveis candidatas a sair e é, portanto, o *submodelo final*, escolhido pelo algoritmo de exclusão sequencial (ao nível $\alpha = 0.10$, mas os resultados seriam neste exemplo idênticos se tivesse sido escolhido um nível de significância $\alpha = 0.05$). Trata-se dum modelo com sete variáveis preditoras. Repare-se como foi possível excluir quase metade dos preditores iniciais, baixando o valor original de R^2 (0.8577) para um valor apenas ligeiramente inferior: $R^2 = 0.8543$. O submodelo final é bem mais parcimonioso, sem perda substancial de qualidade.

O comando `step` do R escolhe o mesmo submodelo final com sete variáveis preditoras, como se pode confirmar executando o comando

```
> step(lm(V8 ~ . , data=vinho.RLM), direction="backward")
```

- (b) Embora os resultados da alínea anterior deixem antever o resultado do teste pedido nesta alínea, executaremos o *teste F parcial* para comparar o modelo completo original (com todas as variáveis preditoras) e o submodelo final. Nesta resolução, escrevemos os parâmetros do modelo completo com uma numeração não sequencial, de forma a que cada β_j tenha o mesmo índice que a variável V_j ($j = 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14$) que esse coeficiente multiplica:

$$V8 = \beta_0 + \beta_2 V2 + \beta_3 V3 + \beta_4 V4 + \beta_5 V5 + \beta_6 V6 + \beta_7 V7 + \beta_9 V9 + \beta_{10} V10 \\ + \beta_{11} V11 + \beta_{12} V12 + \beta_{13} V13 + \beta_{14} V14 + \epsilon$$

O submodelo resulta de excluir as variáveis preditoras **V2, V3, V6, V11 e V14**:

$$V8 = \beta_0 + \beta_4 V4 + \beta_5 V5 + \beta_7 V7 + \beta_9 V9 + \beta_{10} V10 + \beta_{12} V12 + \beta_{13} V13 + \epsilon$$

Os dois modelos serão idênticos caso os coeficientes das variáveis excluídas do submodelo sejam todos nulos: $\beta_2 = \beta_3 = \beta_6 = \beta_{11} = \beta_{14} = 0$. Essa é a hipótese nula do teste.

Hipóteses: $H_0 : \beta_j = 0, \forall j \in \{2, 3, 6, 11, 14\}$ vs. $H_1 : \exists j \in \{2, 3, 6, 11, 14\}$ tal que $\beta_j \neq 0$.

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

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral direita): Rejeitar H_0 se $F_{calc} > f_{\alpha(p-k, n-(p+1))} = f_{0.05(5, 165)} \approx 2.27$ (entre 2.21 e 2.29, nas tabelas).

Conclusões: No nosso caso, o valor calculado da estatística é $F_{calc} = \frac{165}{5} \frac{0.8577 - 0.8543}{1 - 0.8577} \approx 0.788$. Assim, *não se rejeita* H_0 , que é a hipótese de igualdade entre modelo e submodelo. Esta conclusão está de acordo com as expectativas discutidas na sequência da alínea anterior.

É possível efectuar este teste *F* parcial a modelos encaixados, recorrendo ao comando `anova` do R, indicando como argumentos os referidos modelos ajustados. No nosso caso:

```
> vinho.lm.cmp <- lm(V8 ~ . , data=vinho.RLM)
> vinho.lm.sub <- lm(V8 ~ V4 + V5 + V7 + V9 + V10 + V12 + V13 , data=vinho.RLM)
> anova(vinho.lm.sub, vinho.lm.cmp)
Analysis of Variance Table
```

Model 1: $V8 \sim V4 + V5 + V7 + V9 + V10 + V12 + V13$
 Model 2: $V8 \sim V2 + V3 + V4 + V5 + V6 + V7 + V9 + V10 + V11 + V12 + V13 + V14$

	Res. Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	170	25.732				
2	165	25.123	5	0.60889	0.7998	0.5513

O valor da estatística calculada (sem os arredondamentos introduzidos pelos nossos cálculos) é $F_{calc} = 0.7998$ e o respectivo valor de prova é $p = 0.5513$, indicando uma não rejeição de H_0 . Na listagem de resultados são ainda indicados os valores das Somas de Quadrados Residuais de modelo completo e submodelo (na coluna RSS), o que serve para calcular a estatística através da sua expressão alternativa: $F = \frac{n-(p+1)}{p-k} \frac{SQRE_s - SQRE_c}{SQRE_c}$.

Um comentário final: o algoritmo de exclusão sequencial permite assegurar que dois submodelos obtidos em passos *consecutivos* do algoritmo não diferem significativamente. No entanto, após uma sequência de várias exclusões, poderia dar-se o caso de o submodelo final já diferir de forma significativa do modelo inicial. Assim, e embora não seja o caso neste exercício, seria possível que o resultado dum teste análogo a este, noutro conjunto de dados, fosse diferente.

12. Pede-se para considerar o modelo de regressão linear múltipla de equação

$$v = \beta_0 + \beta_1 \text{cl} + \beta_2 \text{dl} + \beta_3 \text{r} + \beta_4 \text{b} + \epsilon.$$

- (a) Em geral, coeficientes de determinação tomam valores no intervalo $[0, 1]$. No entanto, a matriz de correlações entre cada par de variáveis é disponibilizada no enunciado. Assim, sabemos qual o coeficiente de determinação associado a todas as possíveis regressões lineares *simples* que tenham a variável v como variável resposta. O maior desses coeficientes de determinação corresponde à escolha do preditor b , e seria $R_b^2 = (0.9555627)^2 = 0.9131001$. Uma vez que, acrescentando novos preditores, o coeficiente de determinação R^2 apenas pode crescer, sabemos que para a regressão múltipla indicada o coeficiente de determinação tem de estar no intervalo $]0.9131, 1]$. Trata-se dum informação que faz antever um modelo útil.
- (b) A qualidade do ajustamento do modelo é indicada de duas formas complementares: (i) um teste F de ajustamento global do modelo; e (ii) a análise do valor do coeficiente de determinação. Pelos resultados disponíveis no enunciado, este último é muito elevado: $R^2 = 0.9256$, sugerindo um bom modelo, que explica 92.56% da variabilidade total da variável resposta v . Este facto é confirmado pela claríssima rejeição da hipótese nula num teste de ajustamento global (veja-se a resolução do Exercício 10 para os pormenores dum teste deste tipo). De facto, o valor de prova associado à hipótese nula $H_0 : \mathcal{R}^2 = 0$ é inferior à precisão numérica do computador ($< 2.2 \times 10^{-16}$), ou seja, indistinguível de zero, pelo que não há dúvidas em rejeitar a hipótese nula (hipótese que indicaria um modelo inútil).
- (c) É pedido um intervalo a 95% de confiança para o coeficiente β_2 que, no modelo, multiplica a variável “distância ao solo dum cacho” (variável dl), a fim de avaliar a hipótese que esse coeficiente tenha o valor 0.005. Tem-se:

$$\left] b_2 - t_{\frac{\alpha}{2}(n-(p+1))} \cdot \hat{\sigma}_{\hat{\beta}_2}, b_2 + t_{\frac{\alpha}{2}(n-(p+1))} \cdot \hat{\sigma}_{\hat{\beta}_2} \right[$$

com $b_2 = 0.004121$, $\hat{\sigma}_{\hat{\beta}_2} = 0.002218$ e $t_{\frac{\alpha}{2}(n-(p+1))} = t_{0.025(59)} = 2.00$. Logo, o IC pedido é $] - 0.000315, 0.008557[$. O valor sugerido no enunciado (0.005) pertence a este intervalo, logo é um valor admissível, a 95% de confiança.

- (d) i. O modelo completo tem quatro preditores. O pedido é para indicar qual destes quatro preditores pode ser excluído do modelo com a menor perda (não significativa) de qualidade de ajustamento. Tendo em conta a listagem de resultados do ajustamento do modelo completo, verifica-se que a variável para a qual um teste bilateral a $H_0 : \beta_j = 0$ daria não rejeição, de forma mais clara, é a variável **c1**, cujo *p-value* nesse teste é o mais elevado de todos os preditores: $p = 0.9870$. A escolha deve recair sobre o modelo com preditores **d1**, **b** e **r**.
- ii. Sabemos que a estatística do teste F parcial, quando se compara um modelo e submodelo que diferem numa única variável \vec{x}_j , é o quadrado da estatística T que no modelo completo testa a hipótese $H_0 : \beta_j = 0$ (note que se trata do coeficiente da variável que foi excluída do modelo). Assim, a estatística do teste F parcial relevante será $F_{calc} = 0.016^2 = 0.000256$. Mas por outro lado, sabemos que esta estatística pode ser escrita como $F = \frac{n-(p+1)}{p-k} \cdot \frac{R_c^2 - R_s^2}{1 - R_c^2}$. Nesta expressão conhecemos todos os valores menos R_s^2 , que poderá assim ser calculado:

$$0.000256 = \frac{59}{1} \cdot \frac{0.9256 - R_s^2}{1 - 0.9256} \Leftrightarrow 3.228 \times 10^{-7} = 0.9256 - R_s^2 \Leftrightarrow R_s^2 \approx 0.9256 .$$

Assim, a quatro casas decimais não há alteração no valor de R^2 , resultante da exclusão de **c1**. A Soma de Quadrados Residual do submodelo pode ser obtida utilizando a expressão alternativa da estatística do teste F parcial: $F = \frac{n-(p+1)}{p-k} \cdot \frac{SQRE_s - SQRE_c}{SQRE_c}$. Para poder efectuar o mesmo raciocínio, é necessário determinar o valor de $SQRE_c$. Uma vez que o enunciado fornece o valor de $\sqrt{QMRE_c} = 2.087$, temos que $QMRE_c = \frac{SQRE_c}{n-(p+1)} = 2.087^2 = 4.3556$. Logo, $SQRE_c = 4.3556 * 59 = 256.979$. Assim,

$$\begin{aligned} 0.000256 &= \frac{59}{1} \times \frac{SQRE_s - 256.979}{256.979} \Leftrightarrow 0.001115 = SQRE_s - 256.979 \\ &\Leftrightarrow SQRE_s = 256.9801 . \end{aligned}$$

- iii. Prosseguimos no algoritmo de exclusão sequencial, a partir do submodelo com os três preditores **d1**, **b** e **r**. Como a única informação disponível para os submodelos de dois preditores é o valor do coeficiente de determinação, vamos utilizar os testes F parciais, em vez dos testes t na justificação dos submodelos a escolher. Sabemos que, em modelos que diferem numa única variável, estes dois testes são equivalentes.

Nenhum submodelo de dois preditores, entre os quais esteja a variável **c1** já excluída no passo anterior, pode resultar do passo seguinte do algoritmo de exclusão sequencial. Assim, a questão reside em saber se algum dos três submodelos correspondentes à segunda linha da tabela do enunciado merece ser escolhido. Duas questões se colocam: (i) qual o melhor dos submodelos de dois preditores admissíveis; e (ii) se esse melhor submodelo difere, ou não, significativamente do submodelo actual. A resposta à primeira pergunta é fácil: o melhor dos submodelos candidatos é aquele que tem o maior coeficiente de determinação, não apenas pelo valor em si, mas também porque a esse maior valor de R_s^2 corresponderá o menor valor da estatística do teste F parcial que compara o modelo de três preditores com cada um dos possíveis submodelos de dois preditores. Este facto tornar-se-á claro ao efectuar o teste, o que teremos de fazer para dar resposta à segunda questão acima referida (no caso de todos os submodelos com dois preditores serem significativamente piores que o modelo de três preditores, este último seria o modelo final). O melhor submodelo com dois preditores é o submodelo com as variáveis **d1** e **b**, cujo

coeficiente de determinação associado é $R_s^2 = 0.9229$. O valor muito próximo do valor do coeficiente de determinação do modelo com três preditores (que agora funciona como o modelo completo neste teste F parcial, e para o qual $R_c^2 = 0.9256$ e $p = 3$) sugere que a diferença não seja significativa. Mas façamos o teste F parcial, tendo em conta que a variável a excluir do modelo é a variável \mathbf{r} , cujo coeficiente é β_3 :

Hipóteses: $H_0 : \beta_3 = 0$ vs. $H_1 : \beta_3 \neq 0$.

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

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral direita): Rejeitar H_0 se $F_{calc} > f_{\alpha(p-k, n-(p+1))} = f_{0.05(1,60)} = 4.00$.

Conclusões: O valor calculado da estatística é $F_{calc} = \frac{60}{1} \times \frac{0.9256 - 0.9229}{1 - 0.9256} \approx 2.177$.

Assim, *não se rejeita* H_0 , que é a hipótese de igualdade entre modelo e submodelo.

Esta conclusão está de acordo com as expectativas e sugere que podemos simplificar o modelo sem perda significativa de qualidade de ajustamento.

Falta ainda verificar se este submodelo com dois preditores é o modelo final, ou se é possível simplificar ulteriormente o modelo. Neste caso, queremos comparar o modelo de dois preditores \mathbf{dl} e \mathbf{b} a que chegámos, com os modelos de regressão linear simples de \mathbf{v} com cada uma daquelas variáveis predictoras. Como se viu na alínea (a), a melhor destas duas variáveis predictoras, considerada isoladamente, é o preditor \mathbf{b} , cujo coeficiente de determinação associado é $R_s^2 = 0.9131$. Vejamos se o modelo de regressão linear simples agora referido difere, de forma significativa, do modelo com dois preditores resultante do passo anterior (que aqui toma o papel de modelo completo, com $p = 2$):

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

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

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral direita): Rejeitar H_0 se $F_{calc} > f_{\alpha(p-k, n-(p+1))} = f_{0.05(1,61)} = 3.9985$.

Conclusões: O valor calculado da estatística é $F_{calc} = \frac{61}{1} \times \frac{0.9229 - 0.9131}{1 - 0.9229} \approx 7.7536$.

Logo, *rejeita-se* H_0 , a hipótese de igualdade entre modelo e submodelo. Esta conclusão indica que o modelo de regressão linear simples será significativamente pior que o submodelo com os dois preditores \mathbf{b} e \mathbf{dl} . Este último será o submodelo final.

(e) Nesta alínea trabalha-se com a regressão linear simples com variável resposta \mathbf{v} e variável preditora \mathbf{b} , logo de equação $v_i = \beta_0 + \beta_1 b_i + \epsilon_i$ ($i = 1, \dots, n$).

- i. Como se trata duma regressão linear simples, podemos usar as fórmulas $b_1 = \frac{cov_{vb}}{s_b^2} = r_{vb} \frac{s_v}{s_b}$ e $b_0 = \bar{v} - b_1 \bar{b}$. As médias e variâncias de cada variável são dadas no enunciado, logo sabemos que $\bar{v} = 16.43750$, $\bar{b} = 17.53125$ e os desvios padrões são $s_v = \sqrt{54.85317} = 7.406293$ e $s_b = \sqrt{63.64980} = 7.978082$. Também conhecemos o coeficiente de correlação (igualmente disponibilizado no enunciado) $r_{vb} = 0.9555627$. Assim, $b_1 = 0.9555627 \cdot \frac{7.406293}{7.978082} = 0.8870774$ e $b_0 = 16.43750 - 0.8870774 \cdot 17.53125 = 0.8859243$. Logo, a equação da recta de regressão ajustada é $v = 0.8859 + 0.8871 b$. O declive ajustado indica que, por cada botão adicional por cacho, esperamos que vinguem, em média, mais 0.8871 frutos por cacho.
- ii. O investigador chama a atenção que, dada a natureza das variáveis, tem de verificar-se $v \leq b$. No gráfico de \mathbf{v} vs. \mathbf{b} , disponibilizado no enunciado, essa relação reflecte-se no

facto de todos os pontos estarem abaixo da bissectriz $v = b$. As observações que estão em cima dessa bissectriz correspondem a cachos em que todos os botões vingam.

A observação dos gráficos de resíduos do enunciado sugere que, apesar do valor bastante elevado de R^2 , existem alguns problemas com os pressupostos do modelo de regressão linear múltipla, nomeadamente para efeitos inferenciais. Assim, o primeiro gráfico indicia alguma tendência para um gráfico em forma de funil, o que levanta dúvidas sobre a validade do pressuposto de homogeneidade de variâncias. No segundo gráfico verifica-se que os quantis teóricos e empíricos estão longe de se dispor em linha recta, o que sugere erros aleatórios fortemente não-Normais. No terceiro gráfico é salientada uma observação influente, ou seja, cuja exclusão do conjunto de dados alteraria bastante a recta ajustada: a observação número 13, cuja distância de Cook excede 0.5. Este elevado valor da distância de Cook deve-se essencialmente ao elevado - em módulo - resíduo estandardizado, já que $R_{13} \approx -4$ (recorde-se que as distâncias de Cook podem ser escritas como $D_i = R_i^2 \cdot \left(\frac{h_{ii}}{1-h_{ii}}\right) \cdot \frac{1}{p+1}$ pelo que distâncias de Cook elevadas correspondem a grandes resíduos estandardizados $|R_i|$ e/ou a grandes valores do *leverage* h_{ii}). Note-se que o valor médio dos *leverages* é $\frac{p+1}{n} = \frac{2}{64} = 0.03125$, e a observação 13 tem um *leverage* próximo da média. Há duas observações com *leverage* algo elevado: as observações 62 ($h_{ii} \approx 0.20$) e 14 ($h_{ii} \approx 0.125$). Tratando-se duma regressão linear simples, sabemos que estas observações têm de ter valor da variável preditora \mathbf{b} mais distante da média dos valores dessa variável, ou seja, têm de ter um número de botões por cacho muito diferente de 17.53125. Este facto é confirmado pelo gráfico inicial, de v vs. \mathbf{b} , onde as duas observações referidas têm um número de botões por cacho muito elevado, próximo de 40.

Nos gráficos surgem três observações de resíduos negativos elevados (em módulo): as observações 41, 33 e 13. A partir do gráfico original verificamos que se trata das observações em que maior parece ser a discrepância entre botões e frutos vingados, ou seja, os cachos onde há maiores problemas no vingamento.

É natural que uma parte importante destes problemas detectados nos gráficos de resíduos resulte da já referida restrição a que os dados estão sujeitos: $v \leq \mathbf{b}$. Esta restrição não está incorporada no modelo de regressão linear múltipla. Como vimos, obriga a nuvem de pontos a estar no triângulo inferior direito do gráfico relacionando estas duas variáveis. Qualquer que seja a verdadeira equação da recta de regressão teórica ($v = \beta_0 + \beta_1 \mathbf{b}$), este facto torna impossível que os erros aleatórios ϵ_i tenham distribuição Normal, uma vez que a simetria da Normal em torno do seu ponto médio entra em conflito com a existência duma barreira física (associada à desigualdade $v \leq \mathbf{b}$) para além da qual o erro não pode tomar valores. É de esperar que a distribuição dos erros aleatórios seja assimétrica e de variâncias heterogéneas. Este facto condiciona o valor possível dos resíduos, a sua distribuição, etc.

Saliente-se ainda que, nos três gráficos de resíduos, são visíveis aglomerações de pontos que se distribuem em formas curiosas. Em particular, no primeiro gráfico existem resíduos que estão numa relação quase linear com os valores ajustados. É de supôr que se trata das observações para as quais $v = \mathbf{b}$, já discutidas a propósito do gráfico inicial relacionando estas duas variáveis. De facto, para as observações i em que $v_i = \mathbf{b}_i$, temos que a recta de regressão gera valores ajustados $\hat{v}_i = b_0 + b_1 \mathbf{b}_i = b_0 + b_1 v_i$, o que equivale a dizer $v_i = \frac{\hat{v}_i - b_0}{b_1}$. Logo, os resíduos correspondentes serão: $e_i = v_i - \hat{v}_i = \left(\frac{\hat{v}_i}{b_1} - \frac{b_0}{b_1}\right) - \hat{v}_i = -\frac{b_0}{b_1} + \left(\frac{1}{b_1} - 1\right) \hat{v}_i$. Ou seja, há uma relação linear exacta entre resíduos

e valores esperados da variável resposta, nas observações para as quais $v_i = b_i$. O grupo de pontos em linha recta no primeiro gráfico de resíduos será, assim, o grupo de pontos em cima da bissetriz no gráfico original de v vs. b .

13. (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 **Diâmetro** 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 **Diâmetro** 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 logaritimizadas, tem-se:

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

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

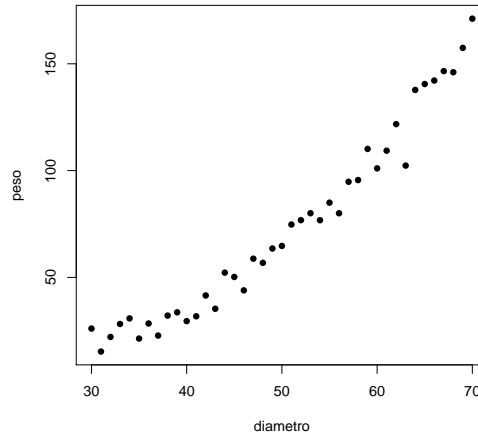
os valores disponíveis no enunciado, equação $y = e^{-6.63162} x_1^{1.98265} x_2^{1.11712}$, ou seja, $Volume = 0.001318 \text{ Diámetro}^{1.98265} \text{ Altura}^{1.11712}$.

- 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 mede 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 $Volume = \beta_0^* \left(\frac{Diámetro}{2}\right)^{\beta_1} \text{ Altura}^{\beta_2}$, 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 **Diámetro** 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.

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

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



É visível uma relação curvilínea, mas uma relação linear entre diâmetro e peso não seria totalmente disparatada, como primeira aproximação. A recta de regressão resultante é:

```
> lm(peso ~ diametro, data=ameixas)
[...]
```

Coefficients:	
(Intercept)	diâmetro
-106.618	3.615

O gráfico da recta $y = -106.618 + 3.615x$ é dado na alínea seguinte (em conjunto com o gráfico da parábola pedida nessa alínea).

(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 (**diâmetro**). De facto, e como foi visto nas aulas teóricas, 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 **
diâmetro	-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. O valor correspondente para o modelo linear ajustado na alínea anterior é $R^2 = 0.9406$.

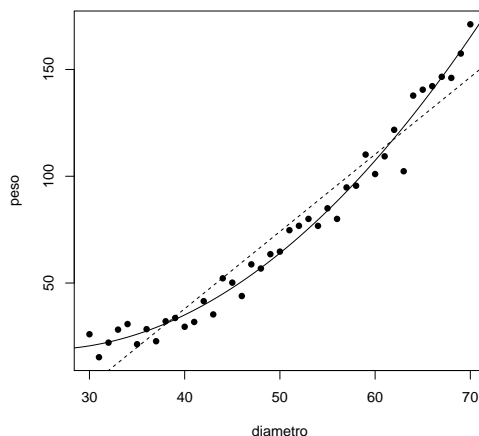
Os parâmetros do modelo (β_0 , β_1 e β_2) são estimados, respectivamente, por: $b_0 = 63.763698$, $b_1 = -3.604849$ e $b_2 = 0.072196$. Logo, a parábola ajustada tem a seguinte equação:

$$peso = 63.763698 - 3.604849 \text{ diametro} + 0.072196 \text{ diametro}^2 .$$

Deve salientar-se que a equação da recta de regressão obtida na alínea anterior (que corresponde a ajustar um polinómio de primeiro grau), **não** é a equação que resulta de deixar cair a parcela associada a x^2 na equação da parábola agora obtida.

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). Como pedido na alínea anterior, também se representa (a tracejado) a recta de regressão de peso sobre diâmetro, 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) Pode ser efectuado um teste de ajustamento global deste modelo de forma análoga ao de qualquer outra regressão linear múltipla. Dado o valor muito elevado do coeficiente de determinação R^2 , é de prever uma rejeição clara da hipótese nula (que corresponde à inutilidade do modelo, neste caso, do modelo quadrático).

Hipóteses: $H_0 : \beta_1 = \beta_2 = 0$ vs. $H_1 : (\beta_1 \neq 0)$ ou $(\beta_2 \neq 0)$.

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

Nível de significância: $\alpha = 0.05$.

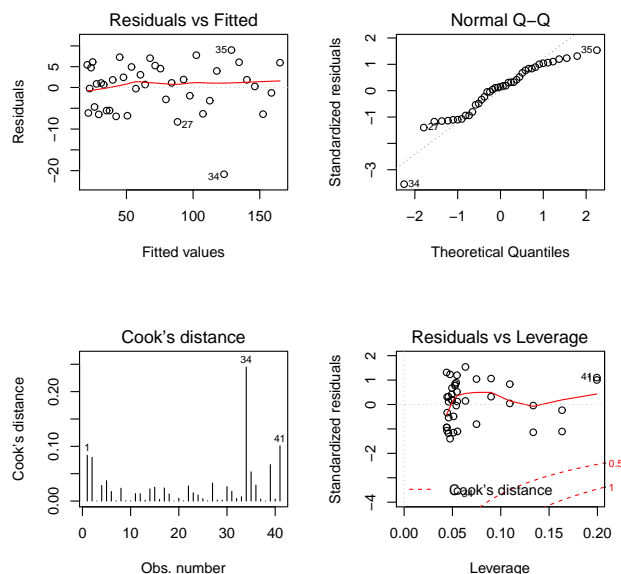
Região Crítica (Unilateral direita): Rej. H_0 se $F_{calc} > f_{\alpha(p,n-(p+1))} = f_{0.05(2,38)} \approx 3.245$.

Conclusões: O valor calculado da estatística é dado na listagem produzida pelo R, mas pode também ser calculado a partir do valor do coeficiente de determinação: $F_{calc} = \frac{38}{2} \times \frac{0.9826}{1-0.9826} = 1072.954$. Logo, rejeita-se (muito enfaticamente) a hipótese nula. Esta conclusão também resulta directamente da análise do valor de prova (p -value) associado à estatística de teste calculada. O modelo parabólico passa claramente este teste de ajustamento global.

Uma questão diferente que se podia colocar era se valia a pena passar do modelo linear para o modelo quadrático, ou seja, saber se o ajustamento da parábola é significativamente melhor do que o ajustamento duma recta de regressão. Para responder (afirmativamente) a esta pergunta, basta olhar para a tabela de resultados do ajustamento obtido com o modelo quadrático: a igualdade dos dois modelos corresponderia à hipótese $\beta_2 = 0$, e essa hipótese é claramente rejeitada ($p = 1.17 \times 10^{-11}$).

(d) 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.

(e) 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

$$peso = 71.27 - 4.089 \text{ diametro} + 0.08222 \text{ diametro}^2 - 0.0006682 \text{ diametro}^3 .$$

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. Mais formalmente, 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).

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 14b não é igual à que se obteria ignorando a última parcela do ajustamento cúbico agora efectuado.

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.

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

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

Assim, calculando o valor $\vec{\mathbf{a}}^t \vec{\mathbf{b}} = a_0 b_0 + a_1 b_1 + \dots + a_p b_p$ do estimador $\vec{\mathbf{a}}^t \hat{\boldsymbol{\beta}}$ e o erro padrão $\hat{\sigma}_{\vec{\mathbf{a}}^t \hat{\boldsymbol{\beta}}}$, para a nossa amostra, temos o intervalo a $(1 - \alpha) \times 100\%$ de confiança para $\vec{\mathbf{a}}^t \vec{\boldsymbol{\beta}} = a_0 \beta_0 + a_1 \beta_1 + \dots + a_p \beta_p$:

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

16. Parte-se dum 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:

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

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 $\left] -217.777, -202.8504 \right[$. 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.

17. (a) A matriz de projecção ortogonal $\mathbf{P} = \vec{\mathbf{1}}_n (\vec{\mathbf{1}}_n^t \vec{\mathbf{1}}_n)^{-1} \vec{\mathbf{1}}_n^t$ é de dimensão $n \times n$ (confirme!), uma vez que o vector $\vec{\mathbf{1}}_n$ é $n \times 1$. Mas o seu cálculo é facilitado pelo facto de que $\vec{\mathbf{1}}_n^t \vec{\mathbf{1}}_n$ é, neste caso, um escalar. Concretamente, $\vec{\mathbf{1}}_n^t \vec{\mathbf{1}}_n = n$, pelo que $(\vec{\mathbf{1}}_n^t \vec{\mathbf{1}}_n)^{-1} = \frac{1}{n}$. Logo $\mathbf{P} = \frac{1}{n} \vec{\mathbf{1}}_n \vec{\mathbf{1}}_n^t$. O produto $\vec{\mathbf{1}}_n \vec{\mathbf{1}}_n^t$ resulta numa matriz $n \times n$ com todos os elementos iguais a 1 (não confundir com o produto pela ordem inversa, $\vec{\mathbf{1}}_n^t \vec{\mathbf{1}}_n$: 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) Tem-se $\mathbf{P}^t = \left[\frac{1}{n} \vec{\mathbf{1}}_n \vec{\mathbf{1}}_n^t \right]^t = \frac{1}{n} [\vec{\mathbf{1}}_n^t]^t [\vec{\mathbf{1}}_n]^t = \frac{1}{n} \vec{\mathbf{1}}_n \vec{\mathbf{1}}_n^t = \mathbf{P}$, logo \mathbf{P} é simétrica. Por outro lado, $\mathbf{P}\mathbf{P} = \vec{\mathbf{1}}_n (\vec{\mathbf{1}}_n^t \vec{\mathbf{1}}_n)^{-1} \vec{\mathbf{1}}_n^t \cdot \vec{\mathbf{1}}_n (\vec{\mathbf{1}}_n^t \vec{\mathbf{1}}_n)^{-1} \vec{\mathbf{1}}_n^t = \vec{\mathbf{1}}_n (\vec{\mathbf{1}}_n^t \vec{\mathbf{1}}_n)^{-1} \vec{\mathbf{1}}_n^t = \mathbf{P}$, logo \mathbf{P} é idempotente.
- (c) A projecção ortogonal do vector $\vec{\mathbf{x}} = (x_1, x_2, \dots, x_n)^t$ (cujos elementos serão por nós encarados como n observações duma variável X) sobre o subespaço gerado pelo vector dos uns $\vec{\mathbf{1}}_n$ é:

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

onde $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ 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 $\mathcal{C}(\vec{\mathbf{1}}_n)$, 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 $\vec{\mathbf{x}}$.

- (d) 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:

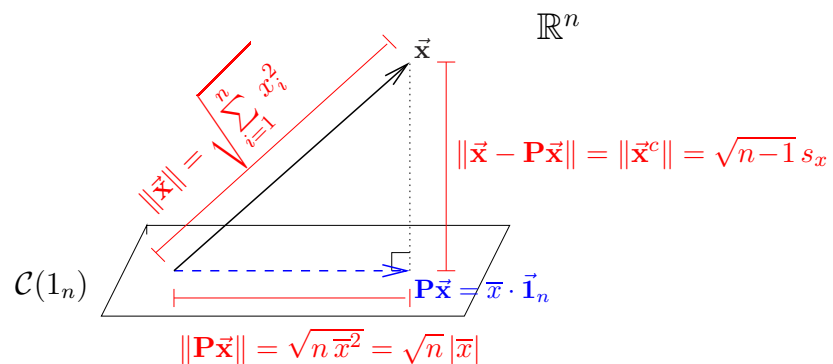
$$(\mathbf{I} - \mathbf{P})\vec{\mathbf{x}} = \mathbf{I}\vec{\mathbf{x}} - \mathbf{P}\vec{\mathbf{x}} = \vec{\mathbf{x}} - \mathbf{P}\vec{\mathbf{x}} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} - \begin{bmatrix} \bar{x} \\ \bar{x} \\ \bar{x} \\ \vdots \\ \bar{x} \end{bmatrix} =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

- (e) A norma do vector $\vec{\mathbf{x}}^c$ é, por definição, a raiz 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:

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

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

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



Nota: O subespaço $\mathcal{C}(\vec{\mathbf{1}}_n)$ é gerado por um único vector, $\vec{\mathbf{1}}_n$, 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 Teóricas, salientando que se trata do mesmo conceito de projecções ortogonais. Aplicando o Teorema de Pitágoras ao triângulo rectângulo indicado, tem-se:

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

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

18. (a) Note-se que a matriz $\mathbf{P}_{\vec{\mathbf{1}}_n}$ referida neste exercício (e que será representada apenas por \mathbf{P} no que se segue) é a mesma que foi discutida no Exercício 17. Assim, o vector $\vec{\mathbf{Y}} - \mathbf{P}\vec{\mathbf{Y}}$ é o vector centrado das observações de $\vec{\mathbf{Y}}$:

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

A norma ao quadrado dum qualquer vector $\vec{\mathbf{z}} = (z_1, z_2, \dots, z_n)^t$ pode ser escrita de duas formas equivalentes: por um lado, $\|\vec{\mathbf{z}}\|^2 = \sum_{i=1}^n z_i^2$, e por outro, $\|\vec{\mathbf{z}}\|^2 = \vec{\mathbf{z}}^t \vec{\mathbf{z}}$. Assim, tem-se

$\|\vec{\mathbf{Y}} - \mathbf{P}\vec{\mathbf{Y}}\|^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 = SQT$. Tendo em conta as propriedades relativas a matrizes, incluindo a simetria e idempotência das matrizes \mathbf{I} (de verificação trivial) e \mathbf{P} (ver Exercício 17), tem-se também:

$$\begin{aligned} SQT &= \|\vec{\mathbf{Y}} - \mathbf{P}\vec{\mathbf{Y}}\|^2 = \|(\mathbf{I} - \mathbf{P})\vec{\mathbf{Y}}\|^2 = [(\mathbf{I} - \mathbf{P})\vec{\mathbf{Y}}]^t (\mathbf{I} - \mathbf{P})\vec{\mathbf{Y}} = \vec{\mathbf{Y}}^t (\mathbf{I} - \mathbf{P})^t (\mathbf{I} - \mathbf{P})\vec{\mathbf{Y}} \\ &= \vec{\mathbf{Y}}^t (\mathbf{I}^t - \mathbf{P}^t) (\mathbf{I} - \mathbf{P})\vec{\mathbf{Y}} = \vec{\mathbf{Y}}^t (\mathbf{I} - \mathbf{P}) (\mathbf{I} - \mathbf{P})\vec{\mathbf{Y}} \\ &= \vec{\mathbf{Y}}^t (\mathbf{I}^2 - \mathbf{I}\mathbf{P} - \mathbf{P}\mathbf{I} + \mathbf{P}^2)\vec{\mathbf{Y}} = \vec{\mathbf{Y}}^t (\mathbf{I} - \mathbf{P} - \mathbf{P} + \mathbf{P})\vec{\mathbf{Y}} \\ &= \vec{\mathbf{Y}}^t (\mathbf{I} - \mathbf{P})\vec{\mathbf{Y}} \end{aligned}$$

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

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

pelo que o quadrado da sua norma é $SQR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$.

Uma expressão alternativa para SQR resulta de considerar (como no caso de SQT) a definição duma norma ao quadrado, e usar as propriedades de matrizes, incluindo a simetria e idempotência de \mathbf{P} e da matriz \mathbf{H} (ver Exercício 4), bem como a propriedade $\mathbf{HP} = \mathbf{PH} = \mathbf{P}$. Esta última passagem resulta do facto de $\mathbf{P} = \vec{\mathbf{1}}_n(\vec{\mathbf{1}}_n^t \vec{\mathbf{1}}_n)^{-1} \vec{\mathbf{1}}_n^t$, pelo que $\mathbf{HP} = \mathbf{H}\vec{\mathbf{1}}_n(\vec{\mathbf{1}}_n^t \vec{\mathbf{1}}_n)^{-1} \vec{\mathbf{1}}_n^t$. Mas, como se viu no Exercício 4, a projecção dum vector sobre um subespaço ao qual esse vector pertence, deixa o vector invariante. Ora \mathbf{H} projecta ortogonalmente sobre o espaço das colunas da matriz \mathbf{X} , $\mathcal{C}(\mathbf{X})$, e o vector $\vec{\mathbf{1}}_n$ pertence a esse subespaço, uma vez que é a primeira coluna da matriz \mathbf{X} . Logo, $\mathbf{H}\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n$, pelo que $\mathbf{HP} = \mathbf{H}\vec{\mathbf{1}}_n(\vec{\mathbf{1}}_n^t \vec{\mathbf{1}}_n)^{-1} \vec{\mathbf{1}}_n^t = \vec{\mathbf{1}}_n(\vec{\mathbf{1}}_n^t \vec{\mathbf{1}}_n)^{-1} \vec{\mathbf{1}}_n^t = \mathbf{P}$. Por outro lado, a simetria de \mathbf{P} (e de \mathbf{H}) implica que $\mathbf{P} = \mathbf{P}^t = (\mathbf{HP})^t = \mathbf{P}^t \mathbf{H}^t = \mathbf{PH}$. Logo,

$$\begin{aligned} SQR &= \|\mathbf{H}\vec{\mathbf{Y}} - \mathbf{P}\vec{\mathbf{Y}}\|^2 = \|(\mathbf{H} - \mathbf{P})\vec{\mathbf{Y}}\|^2 = \vec{\mathbf{Y}}^t(\mathbf{H} - \mathbf{P})^t(\mathbf{H} - \mathbf{P})\vec{\mathbf{Y}} \\ &= \vec{\mathbf{Y}}^t(\mathbf{H}^t - \mathbf{P}^t)(\mathbf{H} - \mathbf{P})\vec{\mathbf{Y}} = \vec{\mathbf{Y}}^t(\mathbf{H} - \mathbf{P})(\mathbf{H} - \mathbf{P})\vec{\mathbf{Y}} \\ &= \vec{\mathbf{Y}}^t(\mathbf{H}^2 - \mathbf{HP} - \mathbf{PH} + \mathbf{P}^2)\vec{\mathbf{Y}} = \vec{\mathbf{Y}}^t(\mathbf{H} - \mathbf{P} - \mathbf{P} + \mathbf{P})\vec{\mathbf{Y}} \\ &= \vec{\mathbf{Y}}^t(\mathbf{H} - \mathbf{P})\vec{\mathbf{Y}}. \end{aligned}$$

Finalmente, o vector $\vec{\mathbf{Y}} - \mathbf{H}\vec{\mathbf{Y}} = \vec{\mathbf{Y}} - \vec{\mathbf{Y}}$ é o vector dos resíduos, e a sua norma ao quadrado é $SQRE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$. Mas, ao mesmo tempo, tem-se:

$$\begin{aligned} SQRE &= \|\vec{\mathbf{Y}} - \mathbf{H}\vec{\mathbf{Y}}\|^2 = \|(\mathbf{I} - \mathbf{H})\vec{\mathbf{Y}}\|^2 = \vec{\mathbf{Y}}^t(\mathbf{I} - \mathbf{H})^t(\mathbf{I} - \mathbf{H})\vec{\mathbf{Y}} \\ &= \vec{\mathbf{Y}}^t(\mathbf{I}^t - \mathbf{H}^t)(\mathbf{I} - \mathbf{H})\vec{\mathbf{Y}} = \vec{\mathbf{Y}}^t(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})\vec{\mathbf{Y}} \\ &= \vec{\mathbf{Y}}^t(\mathbf{I}^2 - \mathbf{IH} - \mathbf{HI} + \mathbf{H}^2)\vec{\mathbf{Y}} = \vec{\mathbf{Y}}^t(\mathbf{I} - \mathbf{H} - \mathbf{H} + \mathbf{H})\vec{\mathbf{Y}} \\ &= \vec{\mathbf{Y}}^t(\mathbf{I} - \mathbf{H})\vec{\mathbf{Y}}. \end{aligned}$$

- (b) Dadas as expressões obtidas na alínea anterior, tem-se (pondo em evidência $\vec{\mathbf{Y}}^t$ à esquerda e $\vec{\mathbf{Y}}$ à direita),

$$\begin{aligned} SQRE + SQR &= \vec{\mathbf{Y}}^t(\mathbf{I} - \mathbf{H})\vec{\mathbf{Y}} + \vec{\mathbf{Y}}^t(\mathbf{H} - \mathbf{P})\vec{\mathbf{Y}} = \vec{\mathbf{Y}}^t[(\mathbf{I} - \mathbf{H}) + (\mathbf{H} - \mathbf{P})]\vec{\mathbf{Y}} \\ &= \vec{\mathbf{Y}}^t[\mathbf{I} - \mathbf{P}]\vec{\mathbf{Y}} = SQT. \end{aligned}$$

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

- (a) A fórmula (disponível no formulário) para o vector dos estimadores de mínimos quadrados do modelo linear contínua válida, pelo que $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{\mathbf{Y}}$. Como

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

temos que $(\mathbf{X}^T \mathbf{X})^{-1} = \frac{1}{n}$ e $\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^n Y_i = \bar{Y}$. 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 $\hat{\beta}_0 = \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{Y}}$ tem distribuição normal (multinormal com uma só componente), de média $E[\hat{\beta}] = \beta = \beta_0$ e variância $V[\hat{\beta}] = \frac{\sigma^2}{n}$ (como determinado na alínea b). Ou seja, $\hat{\beta}_0 \cap \mathcal{N}(\beta_0, \sigma^2/n)$.

(d) Por definição $SQR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$. Considerando o modelo em estudo e o resultado obtido na alínea a), $\hat{Y}_i = \hat{\beta}_0 = \bar{Y}, \forall i = 1, \dots, n$, pelo que $SQR = \sum_{i=1}^n (\bar{Y} - \bar{Y})^2 = 0$. 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 $\{Y_1, Y_2, \dots, Y_n\}$ 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), $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ é um estimador de μ e $\bar{Y} \cap \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

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 $\hat{\beta}_0 = \bar{Y}$ 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 predictoras.

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

$$\text{Modelo completo (C)} \quad Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

$$\text{Submodelo (S)} \quad Y = \beta_0$$

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

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.

20. 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 $\vec{\mathbf{Y}} = \mathbf{X}\vec{\beta} + \epsilon$, embora agora $\vec{\beta}$ 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, $\vec{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}}$. 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 $(\mathbf{X}^t\mathbf{X})^{-1} = \frac{1}{\sum_{j=1}^n x_j^2}$, ficando então $\vec{\beta} = \hat{\beta}_1 = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{j=1}^n x_j^2}$.
- (b) Com os pressupostos usuais no modelo de regressão linear, o vector das observações $\vec{\mathbf{Y}}$ tem distribuição Multinormal, com vector médio $\mathbf{X}\vec{\beta} = \beta_1\mathbf{X}$ e matriz de variâncias-covariâncias $\sigma^2\mathbf{I}_n$, como no caso usual. Também se mantém válido que $\vec{\beta} = \hat{\beta}_1 = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}}$ 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 $E[\vec{\beta}] = \vec{\beta} = \beta_1$ e variância $V[\vec{\beta}] = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1} = \frac{\sigma^2}{\sum_{j=1}^n x_j^2}$.