INSTITUTO SUPERIOR DE AGRONOMIA

Modelos Matemáticos e Aplicações – 2022-23
Resoluções dos exercícios de Regressão Linear Múltipla

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

   (a) A "matriz de nuvens de pontos" produzida pelo comando `plot(vinhos)` tem as nuvens de pontos associadas a cada possível par de variáveis do conjunto de dados. Repare-se como os niveis da variável qualitativa (factor de nome `V1`, na primeira coluna) são convertidos em números inteiros na construção dos respectivos gráficos. Na linha de gráficos 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=vinhos))
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(vinhos$V8)
[1] 0.9977187
> dim(vinhos)
[1] 178  14
> 177*var(vinhos$V8)
[1] 176.5962
> 177*var(fitted(lm(V8 ~ V2 , data=vinhos)))
[1] 9.903747
> 177*var(residuals(lm(V8 ~ V2 , data=vinhos)))
[1] 166.6925
```

pelo que $SQT = (n-1)\,s_y^2 = 176.5962$; $SQR = (n-1)\,s_{\hat{y}}^2 = 9.903747$; e $SQRE = (n-1)\,s_e^2 = 166.6925$.

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

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

```
> round(cor(vinhos[,-1]), 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
```

Note-se que, neste caso, é necessário excluir a primeira coluna (o factor `V1`), dado tratar-se duma variável não numérica, que não é aceite pelo `R` no cálculo duma correlação. Analisando a coluna (ou linha) relativa à variável resposta `V8`, observa-se que a variável com a qual esta se encontra mais correlacionada (em módulo) é `V7` ($r_{7,8} = 0.86$), o que confirma a inspecção visual feita na alínea 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=vinhos)
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=vinhos))
(...)
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`. O facto de o valor de R^2 ser agora inferior ao valor de R^2 na regressão linear simples de `V8` sobre `V7` não contradiz o facto de submodelos não poderem ter valores do coeficiente de determinação superiores, uma vez que o preditor `V7` não faz parte do grupo de cinco preditores agora considerado (ou seja, o modelo da alínea anterior *não é um submodelo* do que foi considerado nesta alínea).

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

```
> lm(V8 ~ . , data=vinhos[,-1])

Call: lm(formula = V8 ~ ., data = vinhos[,-1])
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=vinhos[,-1])))
[1] 151.4735
> 177*var(residuals(lm(V8 ~ . , data=vinhos[,-1])))
[1] 25.12269
```

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

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

(f) O algoritmo de exclusão sequencial solicitado pode ser aplicado com base nos testes t aos coeficientes β_j, ou utilizando a função `step` do `R`, que automatiza o mesmo algoritmo mas utilizando o Critério de Informação de Akaike (AIC). A explicação detalhada destas duas metodologias encontra-se na resolução do ex. 5a). Utilizando o comando

`> step(lm(V8 . , data=vinhos[,-1]))`

obtem-se uma listagem que, na parte final (que se mostra de seguida), apresenta o submodelo final produzido por esta metodologia:

```
(...)
Call:
lm(formula = V8 ~ V4 + V5 + V7 + V9 + V10 + V12 + V13, data = vinhos[,-1])

Coefficients:
(Intercept)        V4         V5         V7         V9        V10        V12      V13
-1.65205      0.45247   -0.02662    0.72642   -0.93935    0.26873    0.52973  0.33218
```

O submodelo final escolhido pelo algoritmo tem sete preditores,

```
> vinho_step.lm <- lm(V8 ~ V4 + V5 + V7 + V9 + V10 + V12 + V13, data = vinhos[, -1])
> summary(vinho_step.lm)

Call:
lm(formula = V8 ~ V4 + V5 + V7 + V9 + V10 + V12 + V13, data = vinhos[,-1])
(...)
```

```
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
```

Comparando a qualidade do modelo completo,

```
> vinhos.lm <- lm(V8~., data=vinhos[,-1])
> summary(vinhos.lm)
(...)
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
```

com a deste submodelo, verifica-se que, quer o coeficiente de determinação habitual, quer o modificado, pouco diminuiram, o que sugere não ter havido perda significativa de qualidade.

(g) O teste F parcial pedido permite confirmar a suspeita anterior, isto é, de modelo e submodelo não diferirem significativamente. No `R` este teste pode ser obtido através do comando `anova`, com o modelo completo ajustado guardado no objecto `vinhos.lm` e o submodelo no objecto `vinho_step.lm`:

```
> anova(vinho_step.lm,vinhos.lm)
Analysis of Variance Table
Model 1: V8 ~ V4 + V5 + V7 + V9 + V10 + V12 + V13
Model 2: V8 ~ 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
```

Como o valor de prova associado a este teste é muito elevado e superior a qualquer um dos habituais níveis de significância ($p - value = 0.5513$), não se rejeita a hipótese de modelo e submodelo serem equivalentes, tal como se anteviu.

O teste F parcial também pode ser feito de forma pormenorizada, utilizando a informação disponível para cada um dos modelos. Um exemplo dos passos e cálculos necessários pode ser consultado no ex. 5e).

2. (a) O procedimento mais fácil será criar um ficheiro de texto com os dados (por exemplo, seleccionando e copiando para uma folha de cálculo o quadro de dados e guardando-o, na pasta de trabalho, como o ficheiro Brix.txt) que depois poderá será lido no R. A leitura/importação deste ficheiro pode ser feita via RStudio (separador Environment, Import Dataset → From text (base)...), ou através do comando

`brix <- read.table("Brix.txt",header=TRUE)`

Em qualquer dos casos, o objecto criado é uma *data frame*, como pode ser verificado com o comando `str(brix)`.

(b) A nuvem de pontos e a matriz de correlações pedidas podem ser obtidas através dos comandos:

```
> plot(brix)
> 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.

(c) 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).

(d) 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
```

(e) 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.

(f) 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...).

(g) 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).

(h) Tem-se:

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

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

O vector $\vec{\mathbf{b}}$ 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. 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.

   O coeficiente de determinação da regressão linear simples da largura das pétalas sobre o seu comprimento é de $R^2 = 0.9271$ (verifique!). 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 $(j = 0, 1, 2, ..., p)$ é:

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

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:

$$]\,-0.20727-1.976346\times0.04751\,,\ -0.20727+1.976346\times0.04751\,[=]\,-0.3011662\,,\ -0.1133738\,[\,.$$

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

$$]\,0.22283-1.976346\times0.04894\,,\ 0.22283+1.976346\times0.04894\,[=]\,0.1261076\,,\ 0.3195524\,[\,.$$

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 comprendido 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: $H_0 : \beta_2 \geq 0 \quad$ vs. $\quad H_1 : \beta_2 < 0$

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:

i. 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;

ii. 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.

4. Na *data frame* `videiras`, a coluna indicando a casta é ignorada neste exercício.

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

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

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

(b) Tem-se:

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

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

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

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

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

```
> videiras.lm <- lm(Area ~ NP + NLesq + NLdir, data=videiras)
> summary(videiras.lm)
(...)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -168.111      5.619 -29.919  < 2e-16 ***
NP             9.987      1.192   8.380  3.8e-16 ***
NLesq         11.078      1.256   8.817  < 2e-16 ***
NLdir         11.895      1.370   8.683  < 2e-16 ***
---
Residual standard error: 24.76 on 596 degrees of freedom
Multiple R-squared: 0.8649,Adjusted R-squared: 0.8642
F-statistic:  1272 on 3 and 596 DF,  p-value: < 2.2e-16
```

A equação do hiperplano ajustado é assim

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

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

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

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

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

Nível de significância: $\alpha = 0.05$.

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

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

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

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

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

Nível de significância: $\alpha = 0.01$.

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

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

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

(f) É pedido um teste à igualdade de dois coeficientes do modelo, concretamente $\beta_2 = \beta_3 \Leftrightarrow \beta_2 - \beta_3 = 0$. Trata-se dum teste à diferença de dois parâmetros, que como foi visto nas aulas 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 $\hat{cov}[\hat{\beta}_2, \hat{\beta}_3]$, valores estes que surgem na matriz de (co)variâncias do estimador $\hat{\vec{\boldsymbol{\beta}}}$, que é estimada por $\hat{V}[\hat{\vec{\boldsymbol{\beta}}}] = QMRE\,(\mathbf{X}^t\mathbf{X})^{-1}$. Esta matriz pode ser calculada no `R` da seguinte forma:

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

Assim,

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

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

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

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

O gráfico do canto superior esquerdo é o gráfico dos resíduos usuais (e_i) vs. valores ajustados ($\hat{y}_i$). Neste gráfico são visíveis dois problemas: uma tendência para a curvatura (já detectado nos gráficos da variável resposta contra cada preditor individual), que indica que o modelo linear pode não ser a melhor forma de relacionar área foliar com os comprimentos das nervuras; e uma forma em funil que sugere que a hipótese de homogeneidade das variâncias dos erros aleatórios pode não ser a mais adequada. Este gráfico foi usado no acetato 179 das aulas teóricas. O gráfico no canto superior direito é um *qq-plot*, de quantis empíricos vs. quantis teóricos duma Normal reduzida. A ser verdade a hipótese de Normalidade dos erros aleatórios, seria de esperar uma disposição linear dos pontos neste gráfico. É visível, sobretudo na parte direita do gráfico, um afastamento relativamente forte de muitas observações a esta linearidade, sugerindo problemas com a hipótese de Normalidade. O gráfico do canto inferior esquerdo é um diagrama de barras com as distâncias de Cook de cada observação. Embora nenhuma observação ultrapasse o limiar de guarda $D_i > 0.5$, duas observações têm um valor considerável da distância de Cook: a observação 499, com D_{499} próximo de 0.4 e a observação 222, com distância de Cook próxima de 0.3. Estas duas observações merecem especial atenção para procurar identificar as causas de tão forte influência no ajustamento. Finalmente, o gráfico do canto inferior direito relaciona resíduos (internamente) estandardizados (eixo vertical) com valor do efeito alavanca (eixo horizontal) e também com as distâncias de Cook (sendo traçadas automaticamente pelo `R` linhas de igual distância de Cook, para alguns valores particularmente elevados, como 0.5 ou 1). Este gráfico ilustra que as duas observações com maior distância de Cook (499 e 222) têm valores relativamente elevados, quer dos resíduos estandardizados, quer do efeito alavanca. Saliente-se que o efeito alavanca médio, neste ajustamento de $n = 600$ observações a um modelo com $p + 1 = 4$ parâmetros é $\overline{h} = \frac{4}{600} = 0.006667$ e as duas observações referidas têm os maiores efeitos alavanca das $n = 600$ observações com valores, respectivamente, próximos de 0.12 e 0.08. Já a observação 481, igualmente identificada no gráfico, tem o maior

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

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

(h) O pedido de logaritmizar previamente as variáveis envolvidas no estudo faz sentido, tendo em conta a curvilinearidade sugerida pelo gráfico de resíduos da alínea anterior (4g). Eis o resultado do ajustamento pedido:

```
> videiraslog.lm <- lm(log(Area) ~ log(NP) + log(NLesq) + log(NLdir), data=videiras)
> summary(videiraslog.lm)

Call: lm(formula = log(Area) ~ log(NP) + log(NLesq) + log(NLdir), data = videiras)
[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.40983    0.06136   6.679 5.52e-11 ***
log(NP)      0.72660    0.06574  11.052  < 2e-16 ***
log(NLesq)   0.57049    0.05649  10.100  < 2e-16 ***
log(NLdir)   0.71077    0.06780  10.484  < 2e-16 ***
---
Residual standard error: 0.1259 on 596 degrees of freedom
Multiple R-squared:  0.9081,Adjusted R-squared:  0.9076
F-statistic:  1963 on 3 and 596 DF,  p-value: < 2.2e-16
```

Não é legítimo procurar comparar directamente o coeficiente de determinação deste modelo, $R^2 = 0.9081$, e o coeficiente de determinação do modelo análogo sem a logaritmização (alínea 4d), $R^2 = 0.8649$), uma vez que a escala onde são medidos os resíduos são diferentes, nos dois casos. Apenas podemos afirmar que o modelo agora ajustado explica mais de 90% da variância dos valores observados *das log-áreas* foliares. A equação do hiperplano ajustado é da forma $\ln(y) = b_0 + b_1\, ln(x_1) + b_2\, \ln(x_2) + b_3\, \ln(x_3)$, sendo y a `Area`, x_1 a variável `NP`, x_2 a variável `NLesq`, e x_3 a variável `NLdir`, e tendo $b_0 = 0.40983$, $b_1 = 0.72660$, $b_2 = 0.57049$ e $b_3 = 0.71077$. Em termos das variáveis originais esta relação corresponde a:

$$\begin{aligned}
\ln(y) = b_0 + b_1\, ln(x_1) + b_2\, \ln(x_2) + b_3\, \ln(x_3) \quad &\Leftrightarrow \quad y = \exp^{b_0 + b_1\, ln(x_1) + b_2\, \ln(x_2) + b_3\, \ln(x_3)} \\
&\Leftrightarrow \quad y = \mathrm{e}^{b_0}\, \mathrm{e}^{b_1\, \ln(x_1)}\, \mathrm{e}^{b_2\, \ln(x_2)}\, \mathrm{e}^{b_3\, \ln(x_3)} \\
&\Leftrightarrow \quad y = \mathrm{e}^{b_0}\, \mathrm{e}^{\ln(x_1^{b_1})}\, \mathrm{e}^{\ln(x_2^{b_2})}\, \mathrm{e}^{\ln(x_3^{b_3})} \\
&\Leftrightarrow \quad y = \mathrm{e}^{b_0}\, x_1^{b_1}\, x_2^{b_2}\, x_3^{b_3}
\end{aligned}$$

Logo o modelo ajustado tem a seguinte equação:

$$Area = 1.506562 \ \mathrm{NP}^{0.72660} \ \mathrm{NLesq}^{0.57049} \ \mathrm{NLdir}^{0.71077} \ .$$

(i) Procede-se como na alínea 4g) e obtém-se:

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

Repare-se como todos os problemas identificados na alínea 4g) foram em boa medida corrigidos. O gráfico de resíduos usuais e_i contra valores ajustados $\hat{y}_i$ mostra agora uma dispersão dos pontos numa banda horizontal em torno do valor médio zero, tendo praticamente desaparecido, quer a curvatura, quer a forma em funil. Assim, a linearidade da relação entre as variáveis logaritmizadas, bem como a homogeneidade das variâncias dos respectivos erros aleatórios parecem pressupostos admissíveis. Da mesma forma, o *qq-plot* do canto superior direito mostra que (à excepção da observação 499) tem-se uma boa linearidade, sustentando o pressuposto de Normalidade dos erros aleatórios. No canto inferior esquerdo, a observação 499 surge de novo destacada, com uma enormíssima distância de Cook, superior a 4, e portanto muito superior ao limiar de guarda 0.5. Assim, esta observação tem uma enorme influência na regressão ajustada, e a sua exclusão provocaria alterações importantes, quer nos coeficientes ajustados b_j, quer nos valores resultantes de $\hat{y}_i$. Essa mesma indicação é dada no quarto e último gráfico, onde (graças ao elevadíssimo valor de D_{499}) são visíveis dois pares de isolinhas de Cook, correspondentes aos limiares 0.5 e 1. Registe-se ainda, nos dois gráficos da direita, como a observação 499 tem um enorme resíduo (internamente) estandardizado, com $R_{499} > 6$, bem como um efeito alavanca razoavelmente elevado (que nenhuma outra observação acompanha). Assim, a logaritmização das quatro variáveis revelou ser uma opção adequada, e por várias razões em simultâneo.

A discordante observação 499 (que é, simultaneamente uma observação atípica, influente e de valor razoavelmente elevado do efeito alavanca) já foi discutida anteriormente. Tratando-se de uma folha com uma muito evidente assimetria (possivelmente correspondente a um erro de medição/registo, ou então danificada por alguma razão), haverá espaço para discutir a sua eventual exclusão do modelo, podendo argumentar-se que o modelo destina-se a ser usado com folhas de videira não danificadas ou excessivamente irregulares. A título de curiosidade, registe-se o resultado de reajustar o modelo, apenas com as 599 folhas restantes:

```
> summary(lm(log(Area) ~ log(NP) + log(NLesq) + log(NLdir), data=videiras[-499,]))

Call: lm(formula = log(Area) ~ log(NP) + log(NLesq) + log(NLdir), data = videiras[-499, ])
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.38758    0.05912   6.555  1.2e-10 ***
log(NP)      0.60695    0.06553   9.262  < 2e-16 ***
log(NLesq)   0.80654    0.06399  12.604  < 2e-16 ***
log(NLdir)   0.60978    0.06681   9.127  < 2e-16 ***
---
Residual standard error: 0.1211 on 595 degrees of freedom
Multiple R-squared:  0.9151,Adjusted R-squared:  0.9146
F-statistic:  2137 on 3 and 595 DF,  p-value: < 2.2e-16
```

Repare-se na alteração substancial dos valores estimados dos quatro parâmetros, e em especial dos coeficientes dos log-comprimentos das nervuras, uma alteração que confirma que a observação 499 era muito influente.

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

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

Não sendo um ajustamento excelente, apesar de tudo as variáveis preditoras explicam quase 75% da variabilidade nos rendimentos. Um teste de ajustamento global rejeita a hipótese nula (inutilidade do modelo) com um valor de prova de $p=0.00004349$.
Quanto ao estudo dos resíduos, devem-se analisar os gráficos produzidos pelos comandos

```
> par(mfrow=c(2,2))
> plot(lm(y ~ . , data=milho),which=c(1,2,4,5))
```

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

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

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

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

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

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

- o preditor x_1 funciona 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 de *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 para iguais condições meteorológicas, ou seja, mantendo constantes as restantes variáveis) 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.

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) Efectuar um teste t às hipóteses $H_0 : \beta_1 = 0$ *vs.* $H_1 : \beta_1 \neq 0$ no modelo com todos os preditores (x_1 a x_9) corresponde 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 4e). 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.

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

Hipóteses: $H_0 : \beta_1 = 0$ [modelos equivalentes] vs. $H_1 : \beta_1 \neq 0$ [modelos diferentes]

ou, de forma equivalente,

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

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

Nível de significância: $\alpha = 0.05$

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

Conclusões: Temos $n = 33$, $p = 9$, $k = 8$, $R^2_c = 0.7476$ e $R^2_s = 0.5074$.

Logo, $F_{\text{calc}} = \frac{23}{1} \times \frac{0.7476-0.5074}{1-0.7476} = 21.8827 > f_{0.05(1,23)} = 4.28$. 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, $t^2_{calc} = (4.678)^2 = 21.88368 = F_{calc}$, 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 $T \frown t_\nu$, então $T^2 \frown F_{(1,\nu)}$. Trata-se de duas estatísticas de teste essencialmente equivalentes.

(f) Com base na listagem de resultados obtidos na alínea 5a), pode identificar-se o preditor x_7 como aquele cuja exclusão do modelo menos prejudicaria a qualidade do modelo. De facto, as colunas relativas aos testes às hipóteses $\beta_j = 0$ mostram que é para essa variável preditora que a não rejeição de H_0 (ou seja, a admissibilidade da hipótese $\beta_7 = 0$) é mais clara, uma vez que o respectivo valor de prova (*p-value*) é o mais elevado de todos, e quase 1: $p = 0.934556$. Este *p-value* corresponde a um valor calculado da estatística T quase nulo: $T_{calc} = 0.083$. Mas, como se viu na alínea anterior, o quadrado deste valor T_{calc} é o valor calculado da estatística do teste F parcial comparando o modelo completo com o submodelo resultante da exclusão do preditor x_7. E tendo em conta a expressão dessa estatística do teste F parcial, onde comparecem os coeficientes de determinação do modelo completo (conhecido: $R^2_c = 0.7476$) e do submodelo (R_s, desconhecido), é possível escrever uma equação em que apenas R_s seja uma icógnita, assim permitindo calcular o seu valor. Logo, tem-se:

$$\begin{aligned} T^2_{calc} = 0.083^2 = 0.006889 &= F_{calc} = \frac{n-(p+1)}{p-k} \cdot \frac{R^2_c - R^2_s}{1-R^2_c} = \frac{23}{1} \cdot \frac{0.7476 - R^2_s}{1-0.7476} \\ \Leftrightarrow \frac{0.006889 \times 0.2524}{23} &= 0.7476 - R^2_s \\ R^2_s &= 0.7476 - 0.0000756 = 0.7475 \end{aligned}$$

Um ajustamento do submodelo sem o preditor x_7 permite confirmar este valor de R^2_s (arredondado a quatro casas decimais).

(g) 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,

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

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

Nível de significância: $\alpha = 0.05$

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

Conclusões: Temos $n = 33$, $p = 9$, $k = 5$, $R^2_c = 0.7476$ e $R^2_s = 0.6435$.
Logo, $F_{\text{calc}} = \frac{23}{4} \times \frac{0.7476-0.6435}{1-0.7476} = 2.371533 < f_{0.05(4,23)} = 2.78$. Assim, não se rejeita H_0, ou seja, o modelo e o submodelo não diferem significativamente ao nível 0.05.

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

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

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

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

(h) 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 5a), conclui-se que há várias variáveis candidatas a sair (os *p-values* correspondentes aos testes a $\beta_j = 0$ são superiores ao limiar acima indicado). De entre estas, é a variável x_7 que tem de longe o maior *p-value*, pelo que é a primeira variável a excluir.

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

```
> summary(lm(y ~ . - x7, data=milho))
[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  54.8704    70.6804   0.776   0.4451
x1            0.8693     0.1602   5.425 1.42e-05 ***
x2            0.7751     0.3983   1.946   0.0634 .
x3           -0.4590     0.4199  -1.093   0.2852
x4           -0.7982     0.9995  -0.799   0.4324
x5            0.4814     0.5613   0.858   0.3996
x6            2.5245     1.2687   1.990   0.0581 .
x8            0.4137     1.0074   0.411   0.6849
```

```
x9           -0.6426     0.6252  -1.028   0.3143
---
Residual standard error: 7.652 on 24 degrees of freedom
Multiple R-squared:  0.7475,Adjusted R-squared:  0.6633
F-statistic: 8.882 on 8 and 24 DF,  p-value: 1.38e-05
```

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

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

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

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

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

```
> summary(lm(y ~ . - x7 - x8 - x4 - x3, data=milho))
[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.3646    64.0755   0.614   0.5441
x1            0.8870     0.1544   5.747 4.13e-06 ***
x2            0.7562     0.3586   2.109   0.0444 *
x5            0.4725     0.4910   0.962   0.3444
x6            2.4893     1.1445   2.175   0.0386 *
x9           -0.8320     0.5515  -1.509   0.1430
---
Residual standard error: 7.461 on 27 degrees of freedom
Multiple R-squared:  0.7299,Adjusted R-squared:  0.6799
F-statistic: 14.59 on 5 and 27 DF,  p-value: 5.835e-07
```

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

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

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

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

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

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

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

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

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

Como foi indicado nas aulas 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. Em relação ao algoritmo baseado nos testes t aos parâmetros β_j, acima ilustrado, apenas pode diferir no momento da paragem do algoritmo: enquanto houver exclusão de variáveis, as variáveis excluídas coincidem nas duas abordagens. Neste exemplo, as duas variantes do algoritmo de exclusão sequencial produzem o mesmo submodelo final, como se pode constatar na parte final desta listagem:

```
> step(milho.lm)                    <--- Comando do R
Start:  AIC=143.79                  <--- AIC do modelo completo
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9

[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
- x7    1      0.42 1405.1 141.79    <--- exclusão de x7 produz o menor (melhor) AIC
- x8    1      9.43 1414.1 142.01    <--- exclusão de x8 (sem excluir x7) é a 2a. melhor opção
- x4    1     33.04 1437.7 142.55
- x5    1     43.28 1448.0 142.79
- x9    1     59.12 1463.8 143.15
- x3    1     70.25 1475.0 143.40
<none>              1404.7 143.78    <--- o modelo inicial
- x2    1    204.13 1608.8 146.26    <--- excluir x2 produz um submodelo com pior (maior) AIC
- x6    1    210.73 1615.4 146.40
- x1    1   1336.47 2741.2 163.85    <--- exclusão de x1: o pior AIC

[Excluída a variável x7, inicia-se novo passo, onde se ensaia a
exclusão de cada uma das variáveis preditoras ainda presentes:]

Step:  AIC=141.8                          <--- AIC do modelo escolhido no passo acima
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x8 + x9  <--- modelo do passo anterior (sem x7)

       Df Sum of Sq    RSS    AIC
- x8    1      9.88 1415.0 140.03    <--- excluir x8 melhora o AIC
- x4    1     37.34 1442.5 140.66    <--- excluir x4 também, mas menos
- x5    1     43.07 1448.2 140.79
- x9    1     61.84 1467.0 141.22
- x3    1     69.96 1475.1 141.40
<none>              1405.1 141.79    <--- o submodelo inicial deste passo
- x2    1    221.75 1626.9 144.63    <--- excluir x2 sobe (piora) AIC
- x6    1    231.80 1636.9 144.83
```

```
- x1    1   1723.38 3128.5 166.21

[Ajusta-se o novo modelo resultante da excluir (também) a variável x8; inicia-se
novo passo para estudar o efeito de excluir um dos dois preditores sobrantes:]

Step:  AIC=140.03
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x9

       Df Sum of Sq    RSS    AIC
- x4    1     41.23 1456.2 138.97   <--- exclusão de x4 melhora AIC
- x5    1     54.23 1469.2 139.27   <--- excluir x5 também, mas menos
- x3    1     70.04 1485.0 139.62
- x9    1     87.98 1503.0 140.02
<none>              1415.0 140.03   <--- submodelo inicial deste passo
- x6    1    222.36 1637.4 142.84   <--- excluir x6 piora AIC
- x2    1    286.50 1701.5 144.11
- x1    1   1800.93 3215.9 165.12

Step:  AIC=138.97                   <--- AIC do submodelo escolhido
y ~ x1 + x2 + x3 + x5 + x6 + x9     <--- submodelo excluindo (também) x4

       Df Sum of Sq    RSS    AIC
- x3    1     46.81 1503.0 138.02   <--- excluir x3 melhor AIC
- x9    1     78.53 1534.8 138.71
- x5    1     82.22 1538.5 138.79
<none>              1456.2 138.97   <--- submodelo inicial do passo
- x2    1    255.37 1711.6 142.31
- x6    1    300.66 1756.9 143.17
- x1    1   1831.49 3287.7 163.85

Step:  AIC=138.02                   <--- AIC do submodelo escolhido acima
y ~ x1 + x2 + x5 + x6 + x9          <--- submodelo sem x4

       Df Sum of Sq    RSS    AIC
- x5    1     51.56 1554.6 137.13   <--- ainda há exclusões a fazer: x5
<none>              1503.0 138.02   <--- modelo inicial do passo
- x9    1    126.71 1629.8 138.69
- x2    1    247.57 1750.6 141.05
- x6    1    263.35 1766.4 141.35
- x1    1   1838.51 3341.6 162.38

Step:  AIC=137.13                   <--- AIC do modelo sem x5
y ~ x1 + x2 + x6 + x9               <--- submodelo excluindo x5

       Df Sum of Sq    RSS    AIC
<none>              1554.6 137.13  <--- <none> na 1a. linha indica que não há
- x9    1    187.95 1742.6 138.90       melhorias de AIC com mais exclusões
- 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)  <--- submodelo final

Coefficients:
(Intercept)           x1           x2           x6           x9
    87.1589       0.8519       0.5989       2.3613      -0.9755  <--- coef ajustados
```

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
[...]
$size
[1]  2  3  4  5  6  7  8  9 10
$r2
[1] 0.5633857 0.6566246 0.6868757 0.7206491 0.7299145 0.7383258 0.7457353
[8] 0.7475100 0.7475856
```

Também é possível utilizar outros indicadores como critério de escolha de submodelos. A utilização do valor do R^2 modificado, R^2_{mod}, como critério de selecção tem de produzir a mesma escolha de subconjuntos óptimos para cada cardinalidade (uma vez que existe uma relação monótona entre as duas variantes de R^2, para iguais conjuntos de dados e subconjuntos de preditores). Essa situação é ilustrada neste exemplo, e de novo recorrendo ao comando `leaps`, com o argumento `method="adjr2"`. Como se pode verificar em baixo, os valores de R^2_{mod} (ao contrário dos valores do coeficiente de determinação usual) não são sempre monótonos para conjuntos encaixados de preditores: assim, na passagem de $k=5$ para $k=6$ preditores, as soluções óptimas apenas diferem na nova variável $x4$, mas o conjunto de preditores maior tem um valor menor de R^2_{mod}. Este indicador pode assim ser um auxiliar na selecção de quantos preditores usar num submodelo RLM.

```
> leaps(y=milho$y, x=milho[,-10], method="adjr2", 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
```

```
[...]
$size
[1]  2  3  4  5  6  7  8  9 10
$adjr2
[1] 0.5493014 0.6337329 0.6544835 0.6807418 0.6798986 0.6779395 0.6745412
[8] 0.6633467 0.6488148
```

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

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

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

Comparando esta listagem com os resultados do modelo final produzido pelo algoritmo de exclusão sequencial, nas unidades de medida originais (ver alínea 5h), constata-se que as quantidades associadas à qualidade do ajustamento global (R^2, valor da estatística F no teste de ajustamento global) mantêm-se inalteradas. Trata-se duma consequência do facto de que as transformações de variáveis foram todas transformações lineares (afins). No entanto, e tal como sucedia na RLS, os valores das estimativas b_j são diferentes. O facto de que a informação relativa aos testes a $\beta_j = 0$ se manter igual, para os coeficientes β_j que multiplicam as variáveis preditoras (isto é, quando $j > 0$), sugere que se trata de alterações que apenas visam adaptar as estimativas às novas unidades de medida, não alterando globalmente o ajustamento.

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

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

(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 $R^2_b = (0.9555627)^2 = 0.9131001$.

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 duma 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 4 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 `cl`, 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 `dl`, `b` e `r`.

ii. Sabemos que a estatística do teste F parcial, quando se compara um modelo e submodelo que diferem numa única variável $\vec{\mathbf{x}}_j$, 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:

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

Assim, a quatro casas decimais não há alteração no valor de R^2, resultante da exclusão de `cl`. 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,

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

iii. Prosseguimos no algoritmo de exclusão sequencial, a partir do submodelo com os três preditores `dl`, `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 `cl` 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^2_s 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 `dl` e `b`, cujo coeficiente de determinação associado é $R^2_s = 0.9229$. 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 $R^2_c = 0.9256$ 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 `r`, cujo coeficiente é β_3:

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

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

Nível de significância: $\alpha = 0.05$.

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

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 `dl` e `b` a que chegámos, com os modelos de regressão linear simples de `v` com cada uma daquelas variáveis preditoras. Como se viu na alínea (a), a melhor destas duas variáveis preditoras, considerada isoladamente, é o preditor `b`, cujo coeficiente de determinação associado é $R^2_s = 0.9131$. 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 `b` e `dl`. Este último será o submodelo final.

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

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 $b_0 = \overline{\mathrm{v}} - b_1\,\overline{\mathrm{b}}$. As médias e variâncias de cada variável são dadas no enunciado, logo sabemos que $\overline{\mathrm{v}} = 16.43750$, $\overline{\mathrm{b}} = 17.53125$ 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 $\mathrm{v} = 0.8859 + 0.8871\,\mathrm{b}$. 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 $\mathrm{v} \leq \mathrm{b}$. No gráfico de `v` vs. `b`, disponibilizado no enunciado, essa relação reflecte-se no facto de todos os pontos estarem abaixo da bissectriz $\mathrm{v} = \mathrm{b}$. 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 $\mathrm{R}_{13} \approx -4$ (recorde-se que as distâncias de Cook podem ser escritas como $D_i = \mathrm{R}_i^2 \cdot \left(\frac{h_{ii}}{1-h_{ii}}\right) \cdot \frac{1}{p+1}$ pelo que distâncias de Cook elevadas correspondem a grandes resíduos estandardizados $|\mathrm{R}_i|$ 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 `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.* `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: $\mathtt{v} \leq \mathtt{b}$. 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 ($\mathtt{v}=\beta_0+\beta_1\mathtt{b}$), 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 $\mathtt{v} \leq \mathtt{b}$) 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 $\mathtt{v}=\mathtt{b}$, já discutidas a propósito do gráfico inicial relacionando estas duas variáveis. De facto, para as observações i em que $\mathrm{v}_i = \mathrm{b}_i$, temos que a recta de regressão gera valores ajustados $\hat{\mathrm{v}}_i = b_0 + b_1\,\mathrm{b}_i = b_0 + b_1\,\mathrm{v}_i$, o que equivale a dizer $\mathrm{v}_i = \frac{\hat{\mathrm{v}}_i - b_0}{b_1}$. Logo, os resíduos correspondentes serão: $e_i = \mathrm{v}_i - \hat{\mathrm{v}}_i = \left(\frac{\hat{\mathrm{v}}_i}{b_1} - \frac{b_0}{b_1}\right) - \hat{\mathrm{v}}_i = -\frac{b_0}{b_1} + \left(\frac{1}{b_1} - 1\right)\hat{\mathrm{v}}_i$. Ou seja, há uma relação linear exacta entre resíduos e valores esperados da variável resposta, nas observações para as quais $\mathrm{v}_i = \mathrm{b}_i$. O grupo de pontos em linha recta no primeiro gráfico de resíduos será, assim, o grupo de pontos em cima da bissectriz no gráfico original de `v` *vs.* `b`.

7. O conjunto de dados subjacente a este Exercício encontra-se num objecto de nome `trees`, no `R`. Nesse objecto, a variável `Diametro` é designada `Girth`, a variável `Altura` é designada `Height`, e a variável `Volume` tem o mesmo nome. No entanto, toda a informação necessária para a resolução encontra-se no enunciado do Exercício.

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

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

que $\beta_1 \neq 0$, já o valor de prova no teste a $H_0 : \beta_2 = 0$ é $p = 0.0145$ e portanto superior a $\alpha = 0.01$. Assim, e embora por pouco, não se rejeita a hipótese $H_0 : \beta_2 = 0$ ao nível de significância $\alpha = 0.01$. Como tal, uma regressão linear simples de `Volume` sobre `Diametro` não difere significativamente (para $\alpha = 0.01$) da regressão com dois preditores ajustada no enunciado.

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

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

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

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

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

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

(c) A troca de variável resposta piorou claramente o valor de R^2 do ajustamento. Este resultado pode parecer surpreendente à primeira vista, uma vez que do ponto de vista algébrico, uma relação da forma $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ é equivalente a $x_2 = \frac{y - \beta_0 - \beta_1 x_1}{\beta_2} = \beta_0^* + \beta_1^* x_1 + \beta_2^* y$ (com $\beta_0^* = \frac{-\beta_0}{\beta_2}$, $\beta_1^* = \frac{-\beta_1}{\beta_2}$ e $\beta_2^* = \frac{1}{\beta_2}$). Além disso, numa regressão linear simples, a troca do preditor e da variável resposta, se bem que muda a equação da recta ajustada, não muda a qualidade do ajustamento (uma vez que $R^2 = r_{xy}^2$, e o coeficiente de correlação é simétrico

nos seus argumentos). Mas numa regressão linear múltipla, permutar a variável resposta com um dos preditores pode, como este exemplo ilustra, gerar um modelo de qualidade bastante diferente. O exemplo sugere a razão de ser deste facto: as variáveis `Volume` e `Diametro` estão fortemente correlacionadas entre si. Qualquer modelo de regressão linear que tenha uma dessas variáveis como variável resposta, e a outra como preditor, terá de ter $R^2 \geq (0.9671194)^2 = 0.9353199$. Mas a variável `Altura`, que foi agora colocada como variável resposta, não está fortemente correlacionada com nenhuma das duas outras. Ao desempenhar o papel de variável resposta, com as outras duas variáveis como preditores, o valor do R^2 resultante poderá ser elevado, mas como este exemplo ilustra, poderá não o ser.

8. (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:

```
> ameixas.lm <- lm(peso ~ diametro, data=ameixas)
> ameixas.lm
[...]
Coefficients:
(Intercept)     diametro
-106.618        3.615
```

O gráfico da recta $y = -106.618 + 3.615\,x$ 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 (`diametro`). 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 **
diametro      -3.604849   0.759323  -4.747 2.91e-05 ***
I(diametro^2)  0.072196   0.007551   9.561 1.17e-11 ***
---
Residual standard error: 6.049 on 38 degrees of freedom
Multiple R-squared: 0.9826,Adjusted R-squared: 0.9816
F-statistic:  1071 on 2 and 38 DF,  p-value: < 2.2e-16
```

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

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

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(ameixas.lm, lty="dashed")
```

(c) Pede-se para testar se vale 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 a esta pergunta, basta fazer um teste T à hipótese de que o coeficiente do termo quadrático β_2 seja nulo. De facto, a equação do modelo quadrático é $Y = \beta_0 + \beta_1 X + \beta_2 X^2$. Se $\beta_2 = 0$, recupera-se a equação do modelo linear, $Y = \beta_0 + \beta_1 X$. Eis os passos deste teste:

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

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

Nível de significância: $\alpha = 0.05$.

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

Conclusões: Como $T_{\text{calc}} = \frac{b_2 - 0}{\hat{\sigma}_{\hat{\beta}_2}} = \frac{0.072196}{0.007551} = 9.561$ (valor disponível na coluna de nome `t value`) é maior que $t_{0.025(28)} = 2.048$, rejeita-se H_0 ao nível de significância de 0.05, isto é, o modelo quadrático tem um ajustamento significativamente diferente (melhor) que o modelo linear. Registe-se que o valor de prova (*p-value*) associado ao valor calculado da estatística está na listagem do ajustamento do modelo, ao lado do valor da estatística correspondente ao teste a $\beta_2 = 0$, sendo 1.17×10^{-11}, pelo que a conclusão é válida para qualquer dos níveis usuais de α.

Alternativamente, seria possível (e equivalente) usar um teste F parcial para comparar o modelo quadrático com o submodelo linear. Vamos utilizar o comando `anova` do F para efectuar esse teste:

```
> anova(ameixas.lm, ameixas2.lm)
Analysis of Variance Table

Model 1: peso ~ diametro
Model 2: peso ~ diametro + I(diametro^2)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     39 4735.3
2     38 1390.5  1    3344.9 91.411 1.171e-11 ***
```

Como se pode constatar, o valor da estatística deste teste F parcial, que compara um modelo completo (quadrático) e um submodelo (linear) que diferem num único preditor ($x_2 = x^2$) é

(a menos de erros de arredondamento) o quadrado do valor da estatística do teste T a que o coeficiente do único preditor que distingue os dois modelos seja nulo: $F_{calc}=91.411=9.561^2$. Os *p-values* são, nos dois casos, iguais. Trata-se dum teste equivalente.

(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

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

No entanto, o acréscimo no valor do valor de R^2 não se faz sentir nas quatro casas decimais mostradas, indicando que o ganho na qualidade de ajustamento com a passagem dum modelo quadrático para um modelo cúbico é quase inexistente. 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 coincidiam) 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 8b) 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.