INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO EXPERIMENTAL– 2024-25

Resoluções de exercícios práticos de Regressão Linear - Abordagem Inferencial

1. (a) i. A informação essencial sobre a regressão pedida pode ser obtida através do comando `summary`:

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

Assim a recta de regressão ajustada tem equação $y = -0.363076 + 0.415755\,x$. O coeficiente de determinação $R^2 = 0.9271$ indica que esta regressão explica quase 93% da variabilidade observada nas larguras das pétalas.

ii. As estimativas dos desvios padrão associados à estimação de cada um dos parâmetros são indicadas na tabela, na coluna de nome `Std.Error` (ou seja, erro padrão). Assim, o desvio padrão associado à estimação da ordenada na origem é $\hat{\sigma}_{\hat{\beta}_0} = 0.039762$. A variância correspondente é o quadrado deste valor, $\hat{\sigma}^2_{\hat{\beta}_0} = 0.001581$. Seria igualmente possível calcular esta variância estimada a partir da sua fórmula: $\hat{\sigma}^2_{\hat{\beta}_0} = QMRE\left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)\,s_x^2}\right]$. O valor de $QMRE$ pode ser obtido a partir da listagem acima, uma vez que, sob a designação `Residual standard error`, a listagem indica o valor $\sqrt{QMRE} = 0.2065$. Os outros valores constantes da expressão podem ser calculados como em exercícios anteriores. De forma análoga, o desvio padrão associado à estimação do declive da recta é $\hat{\sigma}_{\hat{\beta}_1} = 0.009582$, e o seu quadrado é a variância estimada de $\hat{\beta}_1$: $\hat{\sigma}^2_{\hat{\beta}_1} = 9.181472 \times 10^{-5}$. Também aqui, este valor pode ser obtido a partir da expressão $\hat{\sigma}^2_{\hat{\beta}_1} = \frac{QMRE}{(n-1)\,s_x^2}$.
Os valores das estimativas das variâncias dos estimadores dos parâmetros surgem na diagonal principal da matriz de (co)variâncias do estimador $\vec{\hat{\boldsymbol{\beta}}}$, que é estimada por $\hat{V}[\vec{\hat{\boldsymbol{\beta}}}] = QMRE\,(\mathbf{X}^t\mathbf{X})^{-1}$. Esta matriz pode ser calculada no `R` da seguinte forma:

```
> vcov(iris.lm)
              (Intercept)  Petal.Length
(Intercept)   0.0015810158 -3.450711e-04
Petal.Length -0.0003450711  9.182308e-05
```

iii. A afirmação do enunciado corresponde à hipótese $\beta_1 = 0$. De facto, se $\beta_1 = 0$, a equação do modelo que relaciona x e Y reduz-se a $Y_i = \beta_0 + \epsilon_i$, não existindo relação linear entre x e Y. Os cinco passos do teste são:

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

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

Nível de significância: $\alpha = 0.05$

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

Conclusões: O valor calculado da estatística do teste é: $T_{calc} = \frac{0.4157550}{0.009582} = 43.387$. Logo, tem-se uma rejeição clara da hipótese nula $\beta_1 = 0$: o valor estimado $b_1 = 0.415755$ é *significativamente diferente* de zero (ao nível $\alpha = 0.05$), pelo que a recta tem alguma utilidade para prever valores de y (largura da pétala) a partir dos valores de x (comprimento da pétala). Esta conclusão também se pode justificar a partir do valor de prova ($p - value$) do valor calculado da estatística, que é muito pequeno, sendo mesmo inferior à precisão de máquina, $p < 2 \times 10^{-16}$. Mesmo para níveis de significância como $\alpha = 0.01$ ou $\alpha = 0.005$, a conclusão seria a de rejeição de H_0.

Para o caso particular do valor do parâmetro $\beta_1 = 0$ a informação relativa ao teste já é indicada na listagem produzida pelo comando `summary`, nas terceira e quarta colunas da tabela `Coefficients`.

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

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

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

Nível de significância: $\alpha = 0.05$.

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

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

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

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

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

Nível de significância: $\alpha = 0.05$.

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

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

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

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

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

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

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

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

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

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

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

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

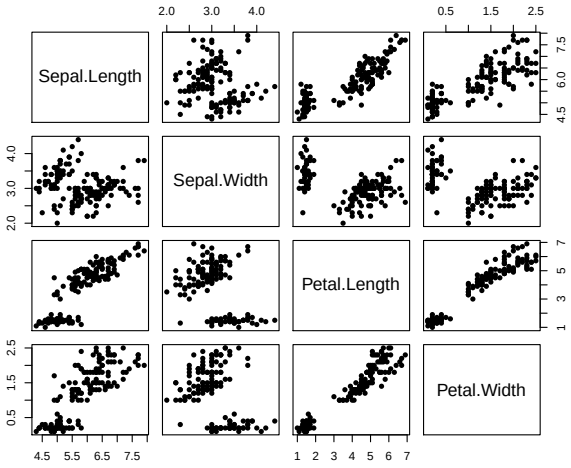
```
> predict(iris.lm, new=data.frame(Petal.Length=4.5), int="pred")
       fit      lwr      upr
1 1.507824 1.098187 1.917461
```

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

(b) 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.

i. O comando a usar no `R` para produzir as nuvens de pontos pedidas pode ser:

```
> 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 na alínea anterior) 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.

ii. 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 $y = -0.2403 + 0.5241\,PL - 0.2073\,SL + 0.2228\,SW$, onde y indica a largura da pétala, x_1 indica o respectivo comprimento, x_2 indica o comprimento da sépala e x_3 a respectiva largura.
Já vimos na alínea anterior que o coeficiente de determinação da regressão linear simples da largura das pétalas sobre o seu comprimento era $R^2 = 0.9271$. O novo valor $R^2 = 0.9379$ é superior, como teria de obrigatoriamente ser num modelo em que se acrescentaram preditores, mas não muito superior. Trata-se, de qualquer forma, dum valor muito elevado, sugerindo que se trata dum bom modelo linear.

iii. 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 pergunta 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.

iv. 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: $PW = -3.2002 + 0.7529SL$.

v. Sabemos que a expressão genérica para os IC a $(1-\alpha)\times 100\%$ para qualquer parâmetro β_j $(j = 0, 1, 2, ..., p)$ é:

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

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

```
> summary(iris2.lm)
(...)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.24031    0.17837  -1.347     0.18
Petal.Length  0.52408    0.02449  21.399  < 2e-16 ***
Sepal.Length -0.20727    0.04751  -4.363 2.41e-05 ***
Sepal.Width   0.22283    0.04894   4.553 1.10e-05 ***
---
Residual standard error: 0.192 on 146 degrees of freedom
Multiple R-squared: 0.9379, Adjusted R-squared: 0.9366
F-statistic: 734.4 on 3 and 146 DF,  p-value: < 2.2e-16
```

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

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

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

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

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

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

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

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

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

vi. 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:

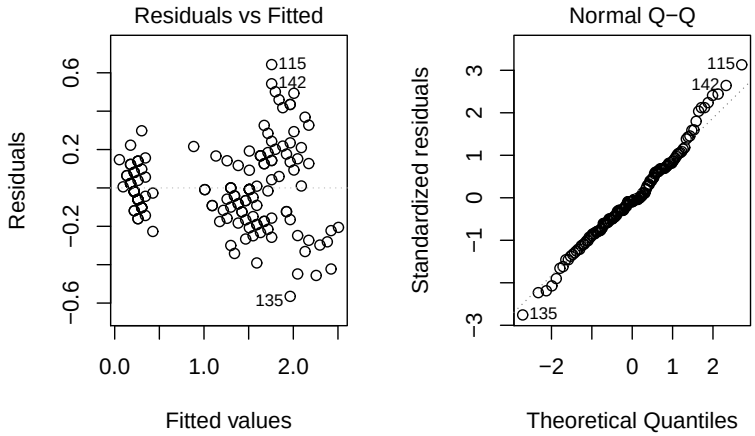
A. Como o valor estimado de β_2 é negativo ($b_2 = -0.20727$) caso se tivesse dado o benefício da dúvida à hipótese $\beta_2 < 0$, nunca se poderia rejeitar essa hipótese;

B. 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.

(c) Dos gráficos de resíduos produzidos pelo comando

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

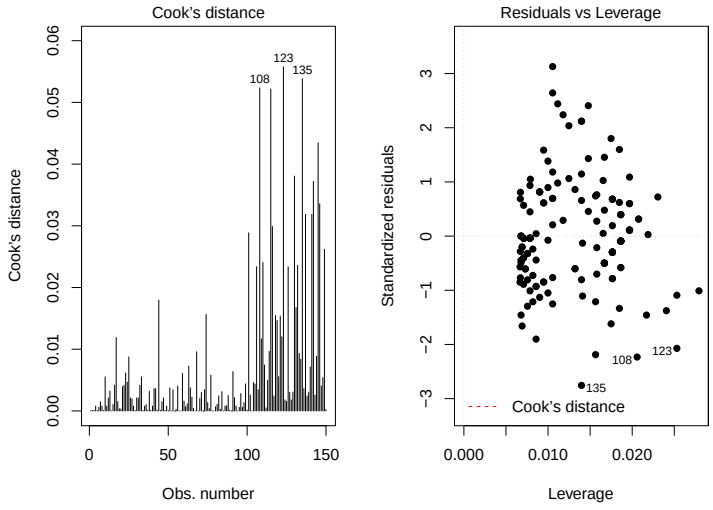
verifica-se que pode existir um problema em relação à hipótese de homogeneidade de variâncias. O gráfico da esquerda sugere que os lírios com comprimento de pétala mais pequeno (do lado esquerdo do gráfico) parecem ter menor variabilidade dos resíduos do que os restantes. Já a linearidade aproximada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



Quanto aos gráficos de diagnóstico produzidos pelo comando

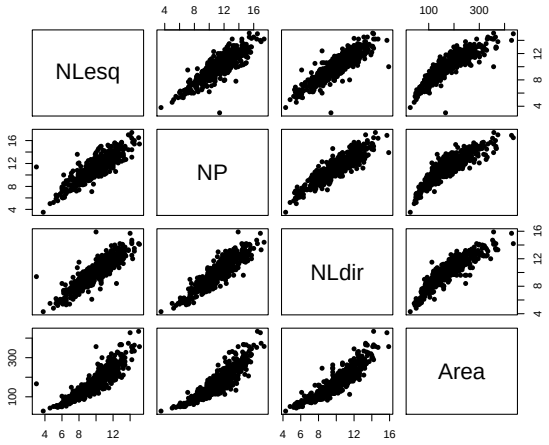
```
> plot(lm(Petal.Width ~ Petal.Length, data=iris),which=c(4,5))
```

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



2. (a) O comando para construir as nuvens de pontos pedidas pode ser:

```
> 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.

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

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

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.

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

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

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

Nível de significância: $\alpha = 0.01$.

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

Conclusões: Tem-se $T_{\text{calc}} = \frac{b_1 - 7}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{9.987 - 7}{1.192} = 2.506 < 2.584$. 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.

(e) 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: $H_0 : \beta_2 - \beta_3 = 0 \quad$ vs. $\quad H_1 : \beta_2 - \beta_3 \neq 0$

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

Nível de significância: $\alpha = 0.05$

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

Conclusões: Conhecem-se as estimativas $b_2 = 11.078$ e $b_3 = 11.895$, mas precisamos ainda de conhecer o valor do erro padrão associado à estimação de $\beta_2 - \beta_3$ que, como foi visto nas aulas 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 $\vec{\hat{\boldsymbol{\beta}}}$, que é estimada por $\hat{V}[\vec{\hat{\boldsymbol{\beta}}}] = QMRE\,(\mathbf{X}^t\mathbf{X})^{-1}$. Esta matriz pode ser calculada no `R` da seguinte forma:

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

Assim,

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

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

(f) i. Substituindo na equação do hiperplano ajustado, obtido na alínea 2c, obtém-se os seguintes valores estimados:

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

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

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

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

ii. Estes intervalos de confiança para $\mu_{Y|X} = E[Y|X_1 = x_1, X_2 = x_2, X_3 = x_3]$ (com os valores de x_1, x_2 e x_3 indicados no enunciado, para cada uma das três folhas) obtêm-se subtraindo e somando aos valores ajustados obtidos na subalínea anterior a semi-amplitude do IC, dada por $t_{\alpha/2(n-(p+1))}\cdot\hat{\sigma}_{\hat{\mu}_{Y|X}}$, sendo $\hat{\sigma}_{\hat{\mu}_{Y|X}} = \sqrt{QMRE\cdot\vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}$ onde os vectores $\vec{\mathbf{a}}$ são os vectores da forma $\vec{\mathbf{a}} = (1, x_1, x_2, x_3)$. Estas contas, algo trabalhosas, resultam fáceis recorrendo de novo ao comando `predict` do `R`, mas desta vez com o argumento `int="conf"`, como indicado de seguida:

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

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

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

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

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

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

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

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

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

(g) 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: $\alpha = 0.05$.

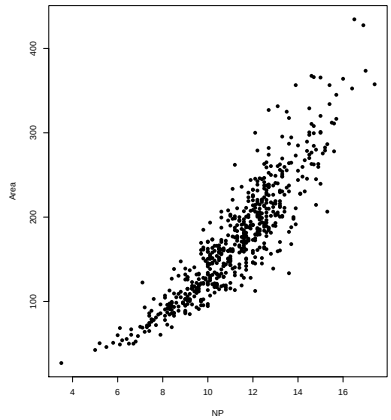
Região Crítica (Unilateral direita): Rej. H_0 se $F_{calc} > f_{\alpha\,(p,n-(p+1))} = f_{0.05(3,596)} \approx 2.62$.

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.

(h) i. A nuvem de pontos pedida, e o correspondente coeficiente de correlação linear, obtêm-se através dos comandos:

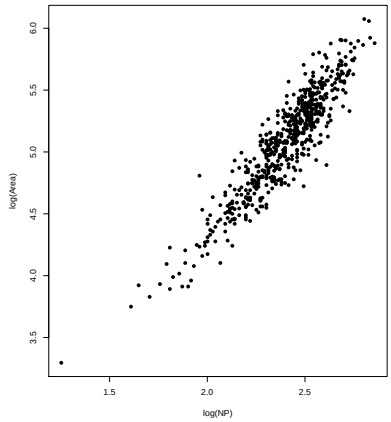
```
> plot(Area ~ NP, data=videiras, pch=16)
> cor(videiras$Area, videiras$NP)
```

O coeficiente de correlação $r_{xy} = 0.89457$ é bastante elevado. O coeficiente de determinação correspondente é $R^2 = 0.89457^2 = 0.8002556$, o que implica que uma recta de regressão explicaria cerca de 80% da variabilidade observada nas áreas foliares. No entanto, uma inspecção da nuvem de pontos revela uma curvatura, sugerindo que a relação linear ajustada não será o modelo mais adequado:



ii. Logaritmizando as duas variáveis obtém-se uma maior linearidade na nuvem de pontos:

```
> plot(log(Area) ~ log(NP), data=videiras, pch=16)
```



Uma regressão linear de log-áreas sobre log-comprimentos da nervura principal tem associado um coeficiente de correlação, e por conseguinte, um coeficiente de determinação, mais elevados. No entanto, os valores não devem ser directamente comparados entre si, uma vez que as escalas onde se medem as Somas de Quadrados (logo os valores de r_{xy} e R^2) são diferentes. Por exemplo, a regressão linear associada a esta alínea explicará cerca de 85% da variabilidade *das log-áreas* (e não *das áreas*) observadas:

```
> cor(log(videiras$Area), log(videiras$NP))
[1] 0.9228187
> cor(log(videiras$Area), log(videiras$NP))^2
[1] 0.8515943
```

iii. A recta de regressão envolvendo as variáveis logaritmizadas obtém-se assim:

```
> vidRLSlog.lm <- lm(log(Area) ~ log(NP), data=videiras)
> vidRLSlog.lm

Call:
lm(formula = log(Area) ~ log(NP), data = videiras)

Coefficients:
(Intercept)      log(NP)
     0.5787       1.8764
```

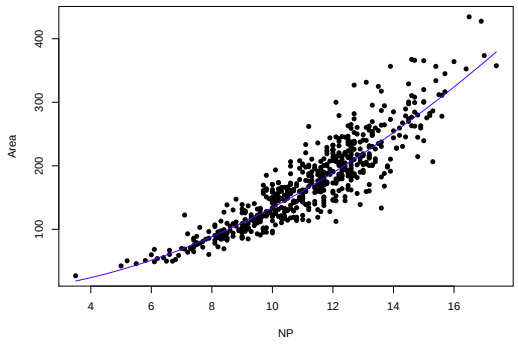
Assim, a recta ajustada é $y^* = 0.5787 + 1.876\,x^*$, onde y^* indica log-áreas e x^* log-comprimentos da nervura principal. Para traçar esta recta no gráfico obtido na alínea 2(h)ii, usa-se o comando:

```
> abline(vidRLSlog.lm)
```

iv. Para obter a relação directa entre áreas e comprimentos (sem a logaritmização), será necessário exponenciar os dois lados da equação:

$$\ln(y) = 0.5787 + 1.876\,\ln(x) \quad \Leftrightarrow \quad y = \mathrm{e}^{0.5787+1.876\,\ln(x)} \quad \Leftrightarrow \quad y = \mathrm{e}^{0.5787}\cdot \mathrm{e}^{1.876\,\ln(x)}$$
$$\Leftrightarrow \quad y = 1.783718\cdot \mathrm{e}^{\ln(x^{1.876})} \quad \Leftrightarrow \quad y = 1.783718\cdot x^{1.876}$$

Assim, a regressão linear ajustada corresponde a uma relação de tipo potência entre a área foliar e o comprimento da nervura principal. Mais concretamente, a área foliar é proporcional à potência 1.876 do comprimento da nervura principal. Eis a curva ajustada:



Esta figura foi traçada com os seguintes comandos do R que permitem, de forma relativamente automatica, sobrepôr a curva potência agora identificada à nuvem de pontos na escala original (sem logaritmização):

```
> plot(Area ~ NP , data=videiras, pch=16)
> coef(videiras.loglm)
(Intercept)      log(NP)
  0.5786896   1.8764239
> c <- exp(coef(videiras.loglm)[1])
> d <- coef(videiras.loglm)[2]
> curve(c*x^d, add=TRUE, col="blue")
```

v. O intervalo a $(1-\alpha)\times 100\%$ de confiança para o declive populacional β_1 duma recta de regressão é:

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

Para obter o erro padrão correspondente à estimação de β_1, $\hat{\sigma}_{\hat{\beta}_1}$, podemos usar a listagem produzida pelo comando `summary` do R:

```
> summary(vidRLSlog.lm)
Call: lm(formula = log(Area) ~ log(NP), data = videiras)
[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.57869    0.07703   7.513 2.12e-13
```

```
log(NP)      1.87642    0.03203  58.579  < 2e-16
---
Residual standard error: 0.1597 on 598 degrees of freedom
Multiple R-squared: 0.8516,  Adjusted R-squared: 0.8513
F-statistic:  3431 on 1 and 598 DF,  p-value: < 2.2e-16
```

Assim, temos $b_1 = 1.87642$ e $\hat{\sigma}_{\hat{\beta}_1} = 0.03203$. Precisa-se ainda do quantil 0.975 da distribuição *t-Student* com $n-2 = 598$ graus de liberdade, ou seja, o valor $t_{\frac{\alpha}{2}(598)} = t_{0.025(598)}$. Usando o valor indicado nas tabelas como sendo o dos infinitos graus de liberdade (e que é, na realidade, o quantil correspondente numa Normal reduzida, $\mathcal{N}(0,1)$), tem-se $t_{0.025,598} \approx 1.96234$. Substituindo, obtém-se o IC pedido:] 1.813566 , 1.939274 [. Este intervalo de confiança pode ser interpretado duma de duas formas: directamente em termos da relação entre as variáveis log-transformadas (em que β_1 é o declive da recta populacional); ou em termos da relação potência entre as variáveis originais (sendo o mesmo β_1 a potência da variável preditora). Assim, pode dizer-se com 95% de confiança que o declive da recta populacional relacionando log-áreas e log-comprimentos está entre 1.813566 e 1.939274, ou seja, que por cada unidade a mais no log-comprimento da nervura principal, a log-área foliar aumentará entre 1.813566 e 1.939274 unidades. Alternativamente, pode afirmar-se que a área foliar é proporcional a uma potência do comprimento da nervura principal, potência essa que, com 95% de confiança, está contida entre 1.813566 e 1.939274.

vi. O enunciado postula a hipótese que $y = c\,x^2$, sendo y a área foliar e x o comprimento da nervura principal. Com base no intervalo a 95% de confiança obtido na alínea anterior, já poderíamos rejeitar a hipótese agora formulada, uma vez que o valor $\beta_1 = 2$ não pertence ao intervalo de confiança para β_1. Mas o enunciado pede que a resposta seja dada com base num Teste de Hipóteses. Os cinco passos do teste são:

Hipóteses: $H_0 : \beta_1 = 2$ vs. $H_1 : \beta_1 \neq 2$.

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

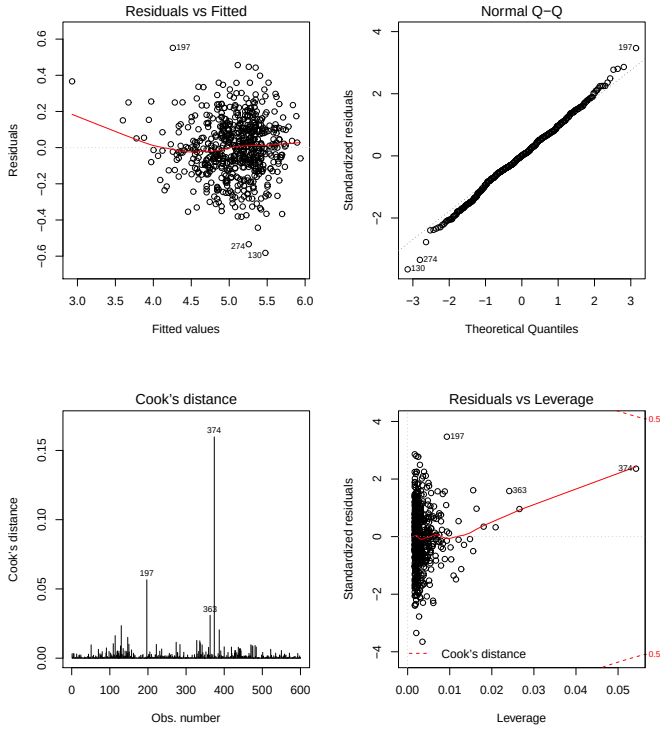
Nível de significância: $\alpha = 0.05$.

Região Crítica (Bilateral): Rejeitar H_0 se $|T_{calc}| > t_{\frac{\alpha}{2}(n-2)} = t_{0.025(598)} = 1.96$.

Conclusões: O valor calculado da estatística do teste é: $T_{calc} = \frac{1.87642-2}{0.03203} = -3.858258$. Logo, rejeita-se claramente a hipótese nula. Ou seja, não se pode afirmar que a área foliar seja proporcional ao quadrado do comprimento da nervura principal, ao nível de significância $\alpha = 0.05$.

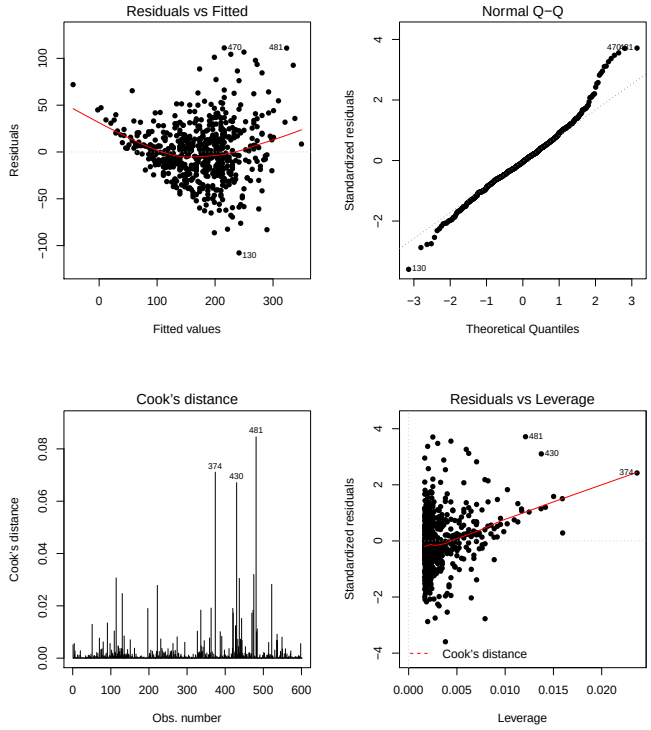
vii. Eis os quatro gráficos de resíduos e diagnósticos considerados nas aulas:

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



Nenhum dos gráficos revela a existência de problemas. No primeiro gráfico, de resíduos e_i vs. valores ajustados $\hat{y}_i$, constata-se uma dispersão adequada, numa banda horizontal em torno do valor zero, que não aponta para qualquer violação dos pressupostos de linearidade ou de variâncias homogéneas. No segundo gráfico, um *qq-plot* à Normalidade dos resíduos, constata-se uma fortíssima linearidade, coerente com o pressuposto de Normalidade dos erros aleatórios. Nos restantes gráficos, e apesar de uma das observações (a 374) ter uma distância de Cook e um efeito alavanca bastante superior às restantes, constata-se que esses valores ($D_i \approx 0.15$ e $h_{374,374} \approx 0.05$) estão perfeitamente dentro dos valores expectáveis. Assim, tudo aponta para a validade do Modelo Linear subjacente à inferência estatística acima indicada.
É de assinalar que um estudo análogo sobre a regressão linear de `Area` sobre `NP` (sem logaritmização) assinala a existência de alguns problemas. Como se constata em baixo, no primeiro gráfico são visíveis uma curvilinearidade já detectada directamente na nuvem de pontos de `Area` sobre `NP` e um efeito funil indicativo de heterogeneidade das variâncias dos erros aleatórios (ou seja, a variância da `Area` tende a crescer com o aumento do comprimento da nervura principal). No segundo gráfico, o *qq-plot*, é visível algum desvio à linearidade para as observações mais à direita (os quantis mais elevados), que indicia algum desvio à Normalidade dos erros aleatórios.

```
> plot(lm(Area ~ NP , data=videiras), which=c(1,2,4,5), pch=16)
```



3. (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. 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 $F_{calc} > f_{\alpha\,(p,n-(p+1))} = f_{0.05(9,23)} \approx 2.3$.

Conclusões: O valor calculado da estatística é dado na listagem produzida pelo R ($F_{calc} = 7.569$). Logo, rejeita-se a hipótese nula, que corresponde à hipótese dum modelo inútil. Esta conclusão também resulta directamente da análise do valor de prova (*p-value*) associado à estatística de teste calculada: $p = 0.00004349$ corresponde a uma rejeição para o nível de significância 0.05.

(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 ($p+1 = 10$). 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.

(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 2d). 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) i. Com base na listagem de resultados obtidos na alínea 3a), 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: $T_{calc} = 0.083$.

ii. Como se viu no ponto 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: $H_0 : \beta_j = 0\ ,\ \forall\, j = 6,7,8,9$ [modelos equivalentes] vs. $H_1 : \exists\, j = 6,7,8,9$ tal que $\beta_j \neq 0$ [modelos diferentes]

ou alternativamente,

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

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

Nível de significância: $\alpha = 0.05$

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

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

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

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

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

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

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

(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 3a), 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:

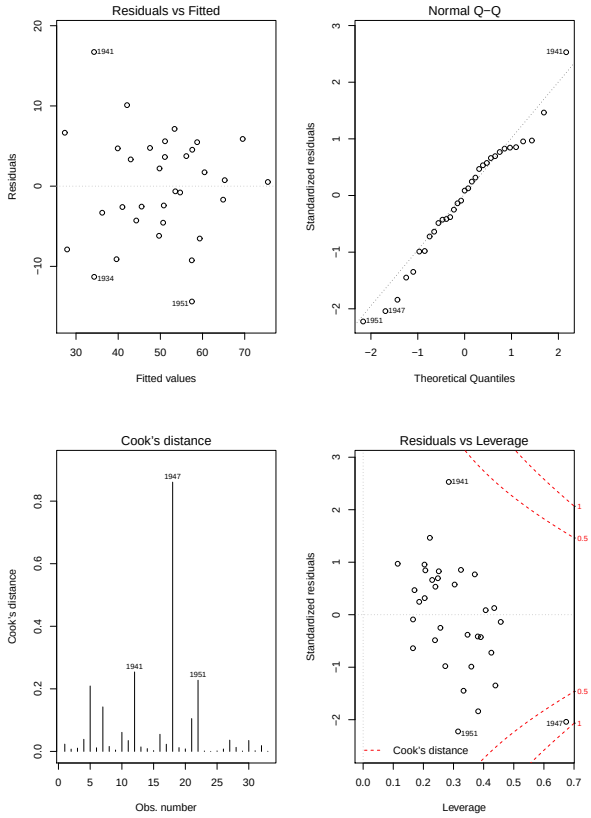
```
> step(milho.lm)                    <--- Comando do R
Start:  AIC=143.79                  <--- AIC do modelo completo
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9

[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
correspodente 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, agora sem x3

       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).

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

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

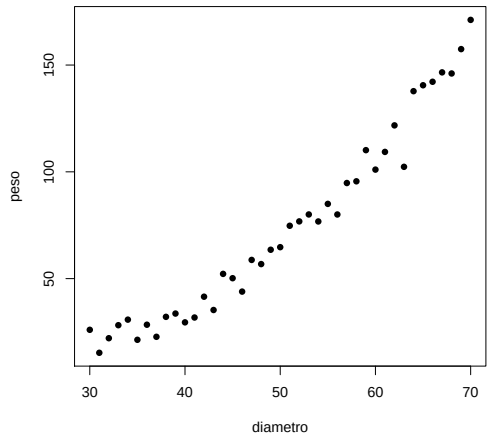


O gráfico de resíduos usuais *vs.* valores ajustados $\hat{y}_i$ (no canto superior esquerdo) não apresenta qualquer padrão digno de registo, dispersando-se os resíduos numa banda horizontal. Assim, nada sugere que não se verifiquem os pressupostos de linearidade e de homgeneidade de variâncias, admitidos 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 ($p+1=10$ 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.

4. (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 $R^2 = 0.9406$.

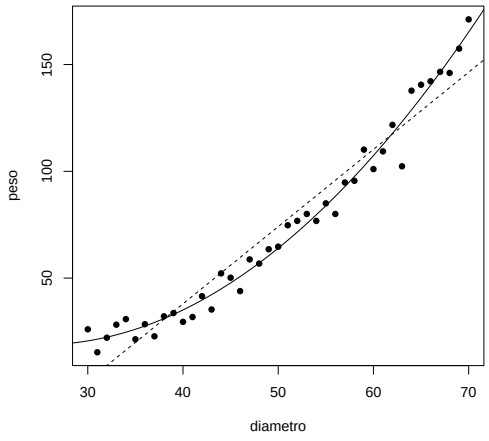
Os parâmetros do modelo (β_0, β_1 e β_2) são estimados, respectivamente, por: $b_0 = 63.763698$, $b_1 = -3.604849$ e $b_2 = 0.072196$. Logo, a parábola ajustada tem a seguinte equação:

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

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.

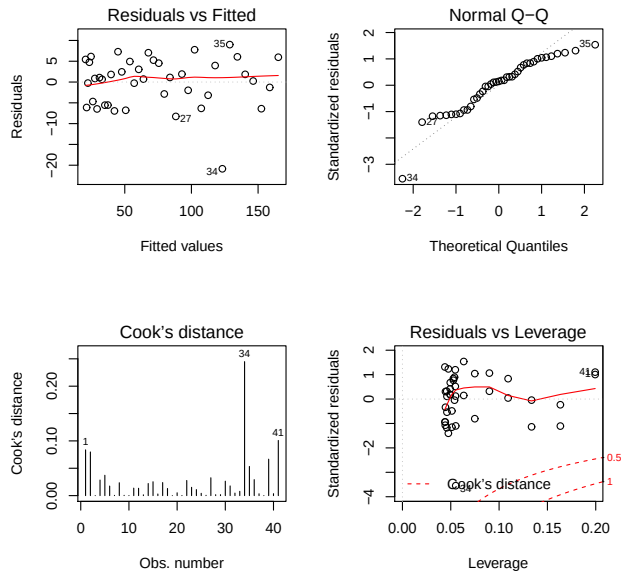
Alternativamente, seria possível (e equivalente) usar um teste F parcial para comparar o modelo quadrático com o submodelo linear. Vamos utilizar o comando `anova` do F para efectuar esse teste:

```
> 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 4b 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.

5. (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:

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

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

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

(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.

6. (a) i. O coeficiente de determinação nesta regressão linear simples é dado pelo quadrado do coeficiente de correlação entre o preditor e a variável resposta, ou seja, $R^2 = (r_{x,y})^2 = (0.9397929)^2 = 0.8832107$. Assim, a recta de regressão ajustada explica cerca de 88.3% da variabilidade observada dos pesos das pêras.

ii. 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 $F_{calc} > f_{\alpha\,(p,n-(p+1))} = f_{0.05(1,118)} \approx 3.9$.

Conclusões: O valor calculado da estatística é $F_{calc} = 892.3665$. Logo, para o nível de significância 0.05. rejeita-se a hipótese nula, isto é, rejeita-se a hipótese de um modelo inútil.

iii. A recta de regressão ajustada de `peso` (y) sobre `diametro` (x) tem equação $y = b_0 + b_1\, x$. O declive é dado por $b_1 = \frac{cov_{xy}}{s_x^2} = r_{xy} \frac{s_y}{s_x} = 0.9397929 \times \sqrt{\frac{1131.675}{27.4688375}} = 6.03216$. A ordenada na origem é dada por $b_0 = \overline{y} - b_1\, \overline{x} = 139.8 - (6.03216)(59.71) = -220.3803$. Assim, a equação da recta ajustada é $y = -220.3803 + 6.03216\, x$. Uma vez que se trata duma recta crescente (declive positivo), para identificar a partir de que diâmetro é que a recta prevê pesos positivos, basta identificar para que diâmetro é que a recta corta o eixo dos xx. Ora,

$$0 = -220.3803 + 6.03216\, x \qquad \Leftrightarrow \qquad x = \frac{220.3803}{6.03216} = 36.53423 \ .$$

Assim, para diâmetros superiores a $36.534\, mm$, os pesos previstos pela recta serão positivos. No entanto, importa assinalar que os valores de diâmetros usados para ajustar a recta são todos razoavelmente superiores a este valor (o menor diâmetro usado no ajustamento foi $48\, mm$, pelo que usar a recta para prever pesos de pêras com diâmetros muito inferiores a $48\, mm$ não é aconselhável.

iv. Dada a fórmula do efeito alavanca numa regressão linear simples (disponível no formulário), é evidente que o efeito alavanca é uma função crescente da distância a que os valores observados da variável preditora, x_i, se encontram da média $\overline{x}$ dos valores do preditor usados no ajustamento. Assim, a observação com o maior valor do efeito alavanca só pode ser a observação com o menor, ou o maior, diâmetro. A distância do menor diâmetro ao diâmetro médio ($|48 - 59.71| = 11.71$) é menor que a distância do maior diâmetro ao diâmetro médio ($|73 - 59.71| = 13.29$). Logo, a observação com o maior efeito alavanca é a observação a que corresponde o diâmetro máximo $x_i = 73.00$. Na nuvem de pontos, essa observação encontra-se no canto superior direito.

A. O valor do efeito alavanca pode ser calculado usando a fórmula disponível no formulário:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{(n-1)s_x^2} = \frac{1}{120} + \frac{(73 - 59.71)^2}{119 \times 27.4688375} = 0.06237 \ .$$

Uma vez que o valor médio dos efeitos alavanca numa regressão linear simples é $\overline{h} = \frac{2}{n}$, que no nosso caso corresponde a $\overline{h} = 0.01666$, o maior efeito alavanca é cerca de quatro vezes maior que o valor médio, mas está muito longe do maior valor possível (1).

B. O resíduo duma observação numa regressão linear simples é dado por $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1\, x_i)$. No nosso caso, queremos o resíduo correspondente à observação com $x_i = 73$ (o maior diâmetro). São conhecidos os dois parâmetros b_0 e b_1 da recta ajustada. Falta conhecer o valor observado da variável resposta correspondente à

observação com o maior valor do preditor x. Logo, tendo em conta que o maior valor observado de peso é $y_i = 255$ (ver enunciado), o valor do resíduo pedido é: $e_i = 255.0 - (-220.3803 + 6.03216 \times 73.00) = 35.0326$.

(b) Foi ajustado um modelo quadrático, cuja equação é $peso = \beta_0 + \beta_1\, diametro + \beta_2\, diametro^2$. Os modelos quadrático e linear são equivalentes se $\beta_2 = 0$. Vamos testar as hipóteses $H_0 : \beta_2 = 0$ *vs.* $H_1 : \beta_2 \neq 0$. A Região Crítica deste teste é bilateral, e para um nível de significância $\alpha = 0.05$ tem-se que devemos rejeitar H_0 se $|T_{calc}| > t_{0.025(117)} \approx 1.980$. Ora, a partir do enunciado verifica-se que o valor calculado da estatística (com $\beta_{2|H_0} = 0$) é $t_{calc} = 3.021$, pelo que se rejeita H_0. Esta conclusão pode também ser obtida a partir do valor de prova (*p-value*) associado ao valor calculado da estatística neste teste bilateral, que é igualmente dado no enunciado: $p = 0.0031$, um valor menor que qualquer dos níveis de significância α usuais. A rejeição de $H_0 : \beta_2 = 0$ corresponde a dizer que o ajustamento do modelo linear e do modelo quadrático são significativamente diferentes, devendo-se nesse caso, optar pelo modelo quadrático, uma vez que garante um melhor ajustamento do que o (sub)modelo linear.

7. (a) i. A melhor variável preditora para uma regressão linear simples (RLS) é a variável mais fortemente correlacionada (em valor absoluto) com a variável reposta `BIO`. De acordo com a matriz de correlações disponível no enunciado, trata-se da variável `pH`, tendo-se $r_{pH,BIO} = 0.7742$. Assim, o coeficiente de determinação dessa RLS é $R^2 = 0.7742^2 = 0.5993856$, podendo afirmar-se que quase 60% da variabilidade da biomassa da *Spartina alterniflora* é explicada pela acidez (`pH`) do solo.

ii. A. A recta de regressão de `BIO` (y) sobre `pH` (x) tem equação $y = b_0 + b_1\, x$. O declive é dado por $b_1 = \frac{cov_{xy}}{s_x^2} = r_{xy}\frac{s_y}{s_x} = 0.7742 \times \frac{660}{1.247} = 409.761$. A ordenada na origem é dada por $b_0 = \overline{y} - b_1\,\overline{x} = 1001 - (409.761)(4.602) = -884.7202$. Assim, a equação da recta de regressão ajustada é $y = -884.7202 + 409.761\, x$.

B. A variância comum aos erros aleatórios ϵ_is é, de acordo com o modelo de regressão linear simples, $V[\epsilon_i] = \sigma^2$ (para qualquer $i = 1, 2, ..., n$). Sabemos que em qualquer modelo linear esta quantidade é estimada pelo Quadrado Médio Residual. Assim, o que é pedido no enunciado é o valor de $QMRE$. Sabemos que $SQT = (n-1)\, s_y^2 = 44 \times 660^2 = 19\,166\,400$. Tem-se $SQR = R^2 \times SQT = 0.5993856 \times 19\,166\,400 = 11\,488\,064$. Logo, $SQRE = SQT - SQR = 7\,678\,336$. Finalmente, $QMRE = \frac{SQRE}{n-2} = \frac{7\,678\,336}{43} = 178\,565.9$.

C. Pede-se um intervalo de predição para uma observação de Y associada a $x = 4.5$. Sabe-se (pelo formulário) que este intervalo de predição é da forma:

$$\left]\, (b_0 + b_1\, x) - t_{\frac{\alpha}{2}(n-2)} \sqrt{QMRE\left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}\right]}\ ,\ (b_0 + b_1\, x) + t_{\frac{\alpha}{2}(n-2)} \sqrt{QMRE\left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}\right]}\, \right[.$$

Ora, $b_0 + b_1\, x = -884.7202 + (409.761)(4.5) = 959.2043$. Por outro lado, $QMRE = 178\,565.9$ (como indicado no enunciado da alínea anterior); $n = 45$, $\overline{x} = 4.602$ e $s_x^2 = 1.247^2$, pelo que $\sqrt{QMRE\left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}\right]} = 427.2718$. Finalmente, $t_{0.025(43)} \approx 2.02$. Logo, o intervalo de predição pedido vem: $]\, 96.115\ ,\ 1822.293\, [$.

(b) i. Não é possível fazer esta afirmação. De facto, os valores de prova (*p-values*) indicados referem-se a testes às hipóteses nulas $\beta_j = 0$, *admitindo que todas as variáveis preditoras*

estão presentes no modelo. A exclusão de um qualquer (ou vários) preditor(es) altera todo o ajustamento, incluíndo a significância dos coeficientes das restantes variáveis. Não há, assim, justificação estatística para proceder à exclusão simultânea de várias variáveis preditoras.

ii. Pede-se um teste F parcial para comparar o modelo completo de equação

$$BIO \;=\; \beta_0 \;+\; \beta_1\, SAL \;+\; \beta_2\, pH \;+\; \beta_3\, K \;+\; \beta_4\, Na \;+\; \beta_5\, Zn + \epsilon$$

e o submodelo de regressão linear simples obtido na primeira pergunta deste exercício, e cuja equação é:

$$BIO \;=\; \beta_0 \;+\; \beta_2\, pH + \epsilon$$

Como de costume, a hipótese nula desse teste corresponde à equivalência do modelo e submodelo, enquanto que a hipótese alternativa corresponde a afirmar que se trata de dois modelos diferentes:

Hipóteses: $H_0 : \beta_1 = \beta_3 = \beta_4 = \beta_5 = 0 \quad vs. \quad H_1 : \exists j \in \{1,3,4,5\}$ tal que $\beta_j \neq 0$.

Estatística do teste: $F = \frac{n-(p+1)}{p-k}\,\frac{R_c^2 - R_s^2}{1-R_c^2}$, onde p é o número de preditores do modelo completo, k é o número de preditores do submodelo, R_c^2 e R_s^2 indicam os coeficientes de determinação, respectivamente do modelo completo e do submodelo. Esta estatística de teste tem, caso seja verdade a hipótese nula de igualdade dos dois modelos encaixados, uma distribuição $F_{(p-k,n-(p+1))}$.

Nível de significância: $\alpha = P[Erro\ Tipo\ I] = P[Rej.H_0 | H_0\ verd.] = 0.10$.

Região Crítica unilateral direita: Rejeitar H_0 se $F_{calc} > f_{0.10(4,39)} \approx 2.09$ (entre os valores tabelados 2.09 e 2.14).

Tem-se $F_{calc} = \frac{39}{4} \times \frac{0.6773-0.5994}{1-0.6773} = 2.3536$, pelo que (qualquer que seja o valor exacto da fronteira da região crítica) F_{calc} pertence a essa região de rejeição. Logo, rejeita-se H_0 e conclui-se que o modelo de cinco preditores e a RLS apenas com o preditor `pH` são significativamente diferentes, ao nível de 10%.

iii. No enunciado pergunta-se se é admissível considerar que $\beta_5 = -40$. Ora, um intervalo a $(1-\alpha)\times 100\%$ de confiança para β_5 é dado por:

$$\left]\quad b_5 \;-\; t_{\alpha/2(n-(p+1))} \cdot \hat{\sigma}_{\hat{\beta}_5} \quad , \quad b_5 \;+\; t_{\alpha/2(n-(p+1))} \cdot \hat{\sigma}_{\hat{\beta}_5} \quad \right[\ ,$$

onde $b_5 = -20.68$, $\hat{\sigma}_{\hat{\beta}_5} = 15.05$ e $t_{0.025(39)} \approx 2.02$. Logo, o intervalo de confiança pedido é: $]\ -51.081,\ 9.721\ [$. Não se pode rejeitar (a 95% de confiança) a hipótese indicada, uma vez que o intervalo contém o valor -40.

iv. A. Apenas foi excluída uma variável preditora, pelo que terá de tratar-se da variável `Na`, à qual corresponde, no modelo completo, o maior *p-value* ($p = 0.58926$, maior que qualquer nível de significância usual) nos testes t às hipóteses $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$. Assim, os quatro preditores do submodelo escolhido são a salinidade, a acidez, e os teores de potássio e zinco.

B. Sabemos que a estatística do teste F parcial comparando o modelo completo de cinco preditores e o submodelo com apenas quatro preditores é: $F = \frac{n-(p+1)}{p-k}\,\frac{R_c^2 - R_s^2}{1-R_c^2}$. Uma vez que o submodelo apenas difere do modelo numa única variável ($k = p-1 = 4$), sabemos que se tem ainda $F_{calc} = (t_{calc})^2 = (-0.544)^2$, onde $t_{calc} = -0.544$ é a

estatística do teste t ao coeficiente β_4 da variável excluída (`Na`). Logo, e tendo em conta os restantes valores conhecidos na expressão de F_{calc}, tem-se:

$$(-0.544)^2 = \frac{39}{1}\frac{0.6773 - R_s^2}{1 - 0.6773} \Leftrightarrow 0.6773 - \frac{(-0.544)^2(1 - 0.6773)}{39} = R_s^2$$
$$\Leftrightarrow R_s^2 = 0.6747 .$$

O submodelo escolhido tem um coeficiente de determinação que apenas difere do coeficiente de determinação do modelo completo na terceira casa decimal.

8. Solução/resolução resumida

   (a) Pelo teste F de ajustamento global e utilizando os habituais níveis de significância, rejeita-se a hipótese do modelo ajustado ser inútil. No entanto, tendo em conta os baixos valores de R^2 e R^2_{mod}, a qualidade deste modelo não é muito boa.

   (b) O valor indicado é a estimativa de β_3, o coeficiente que multiplica a variável `Ca` no modelo de regressão linear múltipla. As suas unidades são $\mu S.kg/(cm.cmol)$. A condutividade eléctrica média diminui 57.68 $\mu S/cm$ quando a concentração de cálcio aumenta 1 $cmol/kg$ e as restantes variáveis preditoras permanecem constantes.

   (c) Pretende-se analisar se o coeficiente que multiplica a variável `Na` (β_1) é menor que zero. Dando o ónus da prova a esta hipótese, o teste T a realizar tem como hipóteses:

   $$H_0 : \beta_1 \geq 0 \qquad \text{vs.} \qquad H_1 : \beta_1 < 0$$

   De acordo com os dados e ao nível de significância $\alpha = 0.05$, não rejeitamos H_0, isto é, não há evidência experimental para afirmar que um aumento na concentração de sódio leva a um decréscimo na condutividade eléctrica média.

   (d) i. Esta afirmação não é legítima. Os valores das estatísticas dos testes a $\beta_j = 0$ dizem respeito ao modelo com *todos* os preditores. A exclusão de *uma* qualquer variável preditora resultaria no ajustamento de novos modelos, com novas estimativas e diferentes erros padrões dos coeficientes. Sem efectuar cálculos adicionais não é lícito fazer a referida afirmação.

   ii. É pedido um teste para comparar o

   Modelo Completo (C) $CE = \beta_0 + \beta_1 Na + \beta_2 K + \beta_3 Ca + \beta_4 Mg$

   com o

   Submodelo (S) $CE = \beta_0 + \beta_4 Mg$

   De acordo com os dados fornecidos, os coeficiente de determinação do modelo completo e do sub-modelo são $R^2_c = 0.5147$ e $R^2_s = (r_{\text{CE,Mg}})^2 = 0.55346^2$, respectivamente. Um teste F parcial para comparar este modelo e submodelo leva à conclusão que o submodelo difere significativamente do modelo completo (para os níveis usuais de significância), pelo que é preferível trabalhar com modelo com 4 preditores.