INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO EXPERIMENTAL– 2025-26

Resoluções de exercícios práticos de Regressão Linear - Abordagem Descritiva

1. Admite-se que foi criado o objecto `Cereais`, tal como indicado no enunciado. Para ver *o conteúdo* desse objecto `Cereais`, escrevemos o seu nome, como ilustrado de seguida (tendo sido omitidas várias linhas do conteúdo por razões de espaço):

```
> Cereais
    ano    area
1  1986 8789.69
2  1987 8972.11
3  1988 8388.94
4  1989 9075.35
5  1990 7573.48
(...)
24 2009 3398.99
25 2010 3041.18
26 2011 2830.96
```

NOTA: O comando `read.csv` parte do pressuposto que o ficheiro indicado contém colunas de dados - cada coluna correspondente a uma variável. O objecto `Cereais` criado no comando acima é uma *data frame*, que pode ser encarada como uma tabela de dados em que cada coluna corresponde a uma variável. As variáveis (colunas) individuais da *data frame* podem ser acedidas através duma indexação análoga à utilizada para objectos de tipo matriz, refereciando o número da respectiva coluna:

```
> Cereais[,2]
 [1] 8789.69 8972.11 8388.94 9075.35 7573.48 8276.47 7684.20 7217.93 6773.54
[10] 6756.57 6528.18 6902.34 5065.38 5923.45 5779.21 4927.15 5149.21 4507.98
[19] 4636.46 3893.43 3731.92 3120.99 3653.74 3398.99 3041.18 2830.96
```

Alternativamente, as variáveis que compõem uma *data frame* podem ser acedidas através do nome da *data frame*, seguido dum cifrão e do nome da variável:

```
> Cereais$area
 [1] 8789.69 8972.11 8388.94 9075.35 7573.48 8276.47 7684.20 7217.93 6773.54
[10] 6756.57 6528.18 6902.34 5065.38 5923.45 5779.21 4927.15 5149.21 4507.98
[19] 4636.46 3893.43 3731.92 3120.99 3653.74 3398.99 3041.18 2830.96
```

(a) `> plot(Cereais)`

O gráfico obtido revela uma forte relação linear (decrescente) entre anos e superfície agrícola dedicada à produção de cereais.

Repare-se que o comando funciona correctamente nesta forma muito simples porque: (i) a *data frame* `Cereais` apenas tem duas variáveis; e (ii) a ordem dessas variáveis coincide com a ordem desejada no gráfico: a primeira variável no eixo horizontal e a segunda no eixo vertical. Existe uma forma mais geral do comando que também poderia ser usada neste caso: `plot(x,y)`, onde `x` e `y` indicam os nomes das variáveis que desejamos ocupem, respectivamente o eixo horizontal e o eixo vertical. No nosso exemplo, poderíamos escrever:

```
> plot(Cereais$ano, Cereais$area)
```

ou, alternativamente,

```
plot (area~ano , data=Cereais)
```

Na fórmula $y \sim x$, a variável do lado esquerdo do til é a variável resposta, e a do lado direito é a variável preditora.

(b) O gráfico obtido na alínea anterior apresenta uma tendência linear descrescente, pelo que o coeficiente de correlação será negativo. A tendência linear é bastante acentuada, pelo que é de supor que o coeficiente de correlação seja próximo de -1.
O comando `cor` do `R` calcula coeficientes de correlação. Se os seus argumentos forem dois vectores (necessariamente de igual dimensão), é devolvido o coeficiente de correlação. Se o seu argumento fôr uma *data frame*, é devolvida uma matriz de correlações entre todos os pares de variáveis da *data frame*. No nosso caso, esta segunda alternativa produz:

```
> cor(Cereais)
            ano       area
ano   1.0000000 -0.9826927
area -0.9826927  1.0000000
```

O coeficiente de correlação entre `ano` e `area` é, como previsto, muito próximo de -1, confirmando a existência duma forte relação linear decrescente entre anos e superfície agrícola para a produção de cereais em Portugal, nos anos indicados.

(c) Os parâmetros da recta podem ser calculados, quer a partir da sua definição, quer utilizando o comando do R que ajusta uma regressão linear: o comando `lm` (as iniciais, pela ordem em inglês, de *modelo linear*). Sabemos que:

$$b_1 = \frac{cov_{xy}}{s_x^2} \qquad \text{e} \qquad b_0 = \overline{y} - b_1\,\overline{x}\ .$$

Utilizando o R, é possível calcular os indicadores estatísticos nas definições:

```
> cov(Cereais$ano, Cereais$area)
[1] -15137.48
> var(Cereais$ano)
[1] 58.5
> -15137.48/58.5
[1] -258.7603
> mean(Cereais$area)
[1] 5869.187
> mean(Cereais$ano)
[1] 1998.5
> 5869.187-(-258.7603)*1998.5
[1] 523001.6
```

Mas o comando `lm` devolve directamente os parâmetros da recta de regressão:

```
> lm(area ~ ano, data=Cereais)
Call:
lm(formula = area ~ ano, data = Cereais)
Coefficients:
(Intercept)          ano
   523001.7       -258.8
```

NOTA: Na fórmula $y \sim x$, a variável do lado esquerdo do til é a variável resposta, e a do lado direito é a variável preditora. O argumento `data` permite indicar o objecto onde se encontram as variáveis cujos nomes são referidos na fórmula.
O resultado deste ajustamento pode ser guardado como um novo objecto, que poderá ser invocado sempre que se deseje trabalhar com a regressão agora ajustada:

```
> Cereais.lm <- lm(area ~ ano, data=Cereais)
```

Interpretação dos coeficientes:

- Declive: $b_1 = -258.8\,km^2/ano$ indica que, em cada ano que passa, a superficie agrícola dedicada à produção de cereais diminui, em média, $258,8\,km^2$. Em geral (e como se pode comprovar analisando a fórmula para o declive da recta de regressão), as unidades de b_1 são as unidades da variável resposta y a dividir pelas unidades da variável preditora x. Fala-se em "variação média" porque a recta apenas descreve a tendência de fundo, na relação entre x e y.
- Ordenada na origem: $b_0 = 523001.7\,km^2$. Em geral, as unidades de b_0 são as unidades da variável resposta y. A interpretação deste valor é, neste caso, estranha: a superfície agrícola utilizada na produção de cereais no ano $x = 0$, seria cerca de 5 vezes superior à área total do país, uma situação claramente impossível. A impossibilidade ilustra a ideia geral de que, *na ausência de mais informação, a validade duma relação linear não poder ser extrapolada para longe da gama de valores de x observada* (neste caso, os anos 1986-2011).

(d) Sabe-se que, numa regressão linear simples entre variáveis x e y, o coeficiente de determinação é o quadrado do coeficiente de correlação entre as variáveis, ou seja: $R^2 = r_{xy}^2$. O valor do coeficiente de correlação entre x e y pode ser obtido através do comando `cor`:

```
> cor(Cereais$ano, Cereais$area)
[1] -0.9826927
> cor(Cereais$ano, Cereais$area)^2
[1] 0.9656849
```

No nosso caso $R^2 = 0.9656849$, ou seja, cerca de $96,6\%$ da variabilidade total observada para a variável resposta y é explicada pela regressão.

O comando `summary`, aplicando ao resultado da regressão ajustada, produz vários resultados de interesse relativos à regressão. O coeficiente de determinação pedido nesta alínea é indicado na penúltima linha da listagem produzida:

```
> summary(Cereais.lm)
(...)
Multiple R-squared: 0.9657
(...)
```

(e) O comando `abline(Cereais.lm)` traça a recta pedida em cima do gráfico anteriormente criado pelo comando `plot`. Confirma-se o bom ajustamento da recta à nuvem de pontos, já indiciado pelo valor muito elevado do R^2.

Nota: Em geral, o comando `abline(a,b)` traça, num gráfico já criado, a recta de equação $y = a + bx$. No caso do *input* ser o ajustamento duma regressão linear simples (obtido através do comando `lm` e que devolve o par de coeficientes b_0 e b_1), o resultado é o gráfico da recta $y = b_0 + b_1\,x$.

(f) Sabemos que $SQT = (n-1)\,s_y^2$, pelo que podemos calcular este valor através do comando:

```
> (length(Cereais$area)-1)*var(Cereais$area)
[1] 101404176
```

(g) Sabemos que $R^2 = \frac{SQR}{SQT}$, pelo que $SQR = R^2 \times SQT$:

```
>  0.9656849*101404176
[1] 97924482
```

Alternativamente, e uma vez que $SQR = (n-1)\, s_{\hat{y}}^2$, pode-se usar o comando `fitted` para obter os valores ajustados de y ($\hat{y}_i$) e seguidamente obter o valor de SQR:

```
> fitted(Cereais.lm)
       1        2        3        4        5        6        7        8
9103.691 8844.930 8586.170 8327.410 8068.649 7809.889 7551.129 7292.368
       9       10       11       12       13       14       15       16
7033.608 6774.848 6516.087 6257.327 5998.567 5739.806 5481.046 5222.286
(...)
> (length(Cereais$area)-1)*var(fitted(Cereais.lm))
[1] 97924480
```

NOTA: A pequena discrepância nos dois valores obtidos para SQR deve-se a erros de arredondamento.

(h) O comando `residuals` devolve os resíduos dum modelo ajustado. Logo,

```
> residuals(Cereais.lm)
         1          2          3          4          5          6          7
-314.00068  127.17965 -197.23002  747.94031 -495.16936  466.58097  133.07131
         8          9         10         11         12         13         14
 -74.43836 -260.06803  -18.27770   12.09263  645.01296 -933.18670  183.64363
(...)
> sum(residuals(Cereais.lm)^2)
[1] 3479697
```

É fácil de verificar que se tem $SQR + SQRE = SQT$:

```
>  97924480+3479697
[1] 101404177
```

(i) Com o auxílio do `R`, podemos efectuar o novo ajustamento. No caso de se efectuar uma transformação duma variável, esta deve ser efectuada, na fórmula do comando `lm`, com a protecção `I()`, como indicado no comando seguinte:

```
> lm(I(area*100) ~ ano, data=Cereais)
Call:
lm(formula = I(area * 100) ~ ano, data = Cereais)
Coefficients:
(Intercept)          ano
   52300171       -25876
```

Comparando estes valores dos parâmetros ajustados com os que haviam sido obtidos incialmente, pode verificar-se que ambos os parâmetros ajustados aparecem multiplicados por 100. Não se trata duma coincidência, o que se pode verificar inspeccionando o efeito da transformação $y \to y^* = c\, y$ (para qualquer constante c) nas fórmulas dos parâmetros da recta ajustada. Indicando por b_1 e b_0 os parâmetros na recta original e por b_1^* e b_0^* os novos parâmetros, obtidos com a transformação indicada, temos (recordando que $cov(x, cy) = c\, cov(x, y)$):

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

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

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

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

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

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

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

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

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

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

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

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

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

2. (a) Seguindo as instruções do enunciado, cria-se o ficheiro de texto `Azeite.txt` na directoria da sessão de trabalho do R, que se recomenda ser uma pasta de nome `AulasED`, num dispositivo de armazenamento de fácil acesso (por exemplo, uma *pen*). Para se saber qual a directoria de trabalho duma sessão do `R`, pode ser dado o seguinte comando:

```
> getwd()
```

(b) O comando de leitura, a partir da sessão do R, pode ser:

```
> azeite <- read.table("Azeite.txt", header=TRUE)
```

Caso o ficheiro `Azeite.txt` esteja numa directoria diferente da directoria de trabalho do R, o nome do ficheiro deverá incluir a sequência de pastas e subpastas que devem ser percorridas para chegar até ao ficheiro.

NOTA: O argumento `header` tem valor lógico que indica se a primeira linha do ficheiro a ser lido contém, ou não, os nomes das variáveis. Por omissão, neste comando, o argumento tem o valor lógico `FALSE`, que considera que na primeira linha do ficheiro já há valores numéricos. Como no ficheiro `Azeite.txt` a primeira linha contém os nomes das variáveis, foi necessário indicar explicitamente o valor lógico `TRUE`.
O resultado do comando pode ser visto escrevendo o nome do objecto agora lido:

```
> azeite
    Ano Azeitona Azeite
1  1995   311257 477728
2  1996   275143 452038
3  1997   309090 423584
4  1998   225616 360948
5  1999   320865 512264
6  2000   167161 249433
7  2001   218522 349502
8  2002   211574 310474
9  2003   232947 364976
10 2004   300699 500658
11 2005   203909 318174
12 2006   362301 518466
13 2007   203968 352574
14 2008   336479 587422
15 2009   414687 681850
16 2010   435009 686832
```

(c) Quando aplicado a uma *data frame*, o comando `plot` produz uma "matriz de gráficos" de cada possível par de variáveis (confirme!). Neste caso, não é pedido qualquer gráfico envolvendo a primeira variável da *data frame*. Existem várias maneiras alternativas de pedir apenas o gráfico das segunda e terceira variáveis, uma das quais envolve o conceito de *indexação negativa*, que tanto pode ser utilizado em *data frames* como em matrizes: índices negativos representam linhas ou colunas a serem *omitidas*. Assim, qualquer dos seguintes comandos (alternativos) produz o gráfico pedido no enunciado:

```
> plot(azeite[,-1])
> plot(azeite[,c(2,3)])
> plot(azeite$Azeitona, azeite$Azeite)
ou, alternativamente,
> plot(Azeite ~Azeitona, data=azeite)
```

(d) O comando `cor` do R calcula a matriz dos coeficientes de correlação entre cada par de variáveis da *data frame*.

```
> cor(azeite)
               Ano  Azeitona    Azeite
Ano      1.0000000 0.3999257 0.4715217
Azeitona 0.3999257 1.0000000 0.9722528
Azeite   0.4715217 0.9722528 1.0000000
```

O valor da correlação pedido é $r_{xy} = 0.9722528$, um valor positivo muito elevado, que indica uma relação linear crescente muito forte, entre produção de azeitona e produção de azeite.

(e) Utilizando o comando `lm` do R, tem-se:

```
> lm(Azeite ~ Azeitona, data=azeite)
Call: lm(formula = Azeite ~ Azeitona, data = azeite)
Coefficients:
(Intercept)     Azeitona
  -5151.793        1.596
```

Por cada tonelada adicional de produção de azeitona oleificada, há um aumento médio de $1.596hl$ de produção de azeite. De novo, o valor da ordenada na origem não faz sentido no contexto do problema: indica que, na ausência de produção de azeitona, a produção média de azeite seria negativa ($b_0 = -5151.793hl$). O modelo não deve ser utilizado (nem tal faria sentido) para produções de azeitona próximas de zero. Em geral, deve ser usado com muito cuidado fora da gama de valores observados de x.

(f) A precisão da recta é uma designação alternativa para o coeficiente de determinação R^2. Sabe-se que, numa regressão linear simples, $R^2 = r_{xy}^2$. Logo, e tendo em conta os resultados já obtidos, a forma mais fácil de calcular R^2 é $R^2 = 0.9722528^2 = 0.9452755$. Assim, cerca de 94.5% da variabilidade na produção de azeite é explicável pela regressão linear simples sobre a produção de azeitona.

3. Os dados referidos no enunciado são obtidos como se indica a seguir:

```
> library(MASS)
> Animals
                     body  brain
Mountain beaver      1.350    8.1
Cow                465.000  423.0
Grey wolf           36.330  119.5
Goat                27.660  115.0
Guinea pig           1.040    5.5
Dipliodocus      11700.000   50.0
Asian elephant    2547.000 4603.0
Donkey             187.100  419.0
Horse              521.000  655.0
Potar monkey        10.000  115.0
Cat                  3.300   25.6
Giraffe            529.000  680.0
Gorilla            207.000  406.0
Human               62.000 1320.0
African elephant  6654.000 5712.0
Triceratops       9400.000   70.0
Rhesus monkey        6.800  179.0
Kangaroo            35.000   56.0
Golden hamster       0.120    1.0
Mouse                0.023    0.4
Rabbit               2.500   12.1
Sheep               55.500  175.0
Jaguar             100.000  157.0
Chimpanzee          52.160  440.0
Rat                  0.280    1.9
Brachiosaurus    87000.000  154.5
Mole                 0.122    3.0
Pig                192.000  180.0
```

(a) A nuvem de pontos pedida pode ser obtida através do comando `plot(Animals)`. Quanto ao coeficiente de correlação, tem-se:

```
> cor(Animals)
              body        brain
body   1.000000000 -0.005341163
brain -0.005341163  1.000000000
```

O valor quase nulo do coeficiente de correlação indica ausência de relacionamento linear entre os pesos do corpo e do cérebro, facto que se confirma visualmente no gráfico.

(b) Pedem-se vários gráficos com transformações de uma ou ambas as variáveis. Aproveita-se este exercício para introduzir uma forma alternativa de pedir uma nuvem de pontos, que utiliza uma sintaxe parecida com as usadas para escrever as fórmulas no comando `lm`:

i. O gráfico de log-pesos do cérebro (no eixo vertical) vs. pesos do corpo (eixo horizontal) pode ser obtido através da tradicional forma `plot(x,y)`, que no nosso caso seria
```
> plot(Animals$body, log(Animals$brain))
```
Alternativamente, pode dar-se o seguinte comando equivalente:
```
> plot(log(brain) ~ body, data=Animals)
```

ii. Usando a forma do comando agora introduzida, a nuvem de pontos pedida é dada por:
```
> plot(brain ~ log(body), data=Animals)
```

iii. Neste caso, e uma vez que a transformação logarítimica se aplica às duas variáveis da *data frame* `Animals`, basta dar o comando
```
> plot(log(Animals))
```
ou, alternativamente,
```
> plot(log(brain) ~ log(body), data=Animals)
```
NOTA: Os logaritmos aqui referidos são os logaritmos naturais, `ln`. Por omissão, o comando `log` do R calcula logaritmos naturais.

(c) Como se viu nas aulas teóricas, uma relação linear entre $ln(y)$ e $ln(x)$ corresponde a uma relação potência (alométrica) entre as variáveis originais: $y = c\,x^d$. Neste caso, tem-se uma relação de tipo alométrico entre pesos duma parte do organismo (cérebro) e do todo (corpo). O último gráfico da alínea anterior indica que é aceitável admitir uma relação potência entre o peso do cérebro e o peso do corpo, nas espécies animais consideradas.

(d) Os coeficientes de correlação e de determinação entre log-pesos do corpo e log-pesos do cérebro podem ser calculados, com o auxílio do `R`, da seguinte forma:

```
> cor(log(Animals$body), log(Animals$brain))    # coeficiente de correlação
[1] 0.7794935
> cor(log(Animals$body), log(Animals$brain))^2  # coeficiente de determinação
[1] 0.6076101
```

Dado o valor $R^2 = 0.6076$, a regressão linear entre log-peso do cérebro e log-peso do corpo explica menos de 61% da variabilidade total dos log-pesos do cérebro observados. Este valor, aparentemente contraditório com a relativamente forte relação linear para a maioria das espécies, é reflexo da presença nos dados das três espécies (pontos) que são claramente atípicas face às restantes.

(e) Os comandos pedidos são:

```
> Animals.loglm <- lm(log(brain) ~ log(body), data=Animals)
> Animals.loglm
Call: lm(formula = log(brain) ~ log(body), data = Animals)
Coefficients:
(Intercept)    log(body)
      2.555        0.496
> abline(Animals.loglm)
```

(admitindo que o último comando `plot` dado antes deste comando `abline` fosse o do gráfico correspondente à dupla logaritmização).

(f) O declive $b_1^* = 0.496$ da recta ajustada tem duas leituras possíveis. Na relação entre as variáveis logaritmizadas tem a habitual leitura de qualquer declive duma recta de regressão: o log-peso do cérebro aumenta em média 0.496 log-gramas, por cada aumento de 1 log-kg no peso do corpo. Mais compreensível é a interpretação na relação potência entre as variáveis originais. Como se viu nas aulas teóricas, a relação original entre y e x é da forma $y = c\,x^d$ com $d = b_1^* = 0.496$ e $b_0^* = \ln(c) = 2.555 \Leftrightarrow c = \mathrm{e}^{2.555} = 12.871$. No nosso caso, a tendência de fundo na relação entre peso do corpo (x) e peso do cérebro (y) é $y = 12.871\,x^{0.496}$. O valor de d muito próximo de 0.5 permite simplificar a relação dizendo que o ajustamento indica que o peso do cérebro é aproximadamente proporcional à *raíz quadrada* do peso do corpo.

(g) O comando

```
> identify(log(Animals))
```

permite, com o auxílio do rato, identificar pontos seleccionados pelo utilizador. (Para sair do modo interactivo, clicar no botão direito do rato com o cursor em cima da janela gráfica).

NOTA: É necessário explicitar as coordenadas dos pontos no gráfico que se vai aceder com o comando. No nosso caso, isso significa explicitar as coordenadas dos dados logaritmizados: `log(Animals)`.

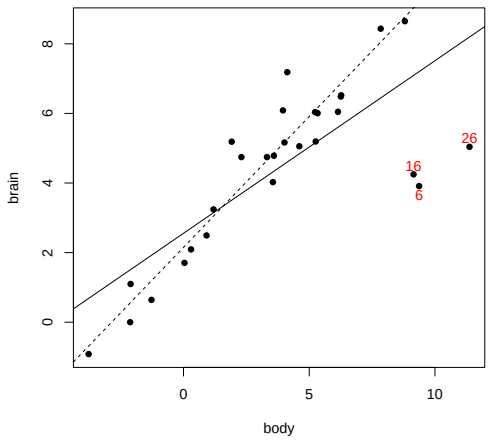
O enunciado pede para identificar os pontos que se destacam da relação linear, e que são os pontos 6, 16 e 26. Selecionando as linhas com esses números podemos identificar as espécies em questão, e verificar que se trata de espécies de dinossáurios, as únicas espécies de animais extintos presentes no conjunto de dados:

```
> Animals[c(6,16,26),]
               body brain
Dipliodocus   11700  50.0
Triceratops    9400  70.0
Brachiosaurus 87000 154.5
```

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

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

O gráfico resultante é reproduzido abaixo. A exclusão das três espécies de dinossáurios (as observações atípicas) permitiu que a recta ajustada acompanhe melhor a relação linear existente entre a generalidade das espécies do conjunto de dados. Este exemplo ilustra que *as rectas de regressão são sensíveis à presença de observações atípicas*. Neste caso, as espécies de dinossáurios "atraem" a recta de regressão, afastando-a da generalidade das restantes espécies.



O ajustamento sem as espécies extintas produz os seguintes parâmetros da recta:

```
> Animals.loglm.sub <- lm(log(brain) ~ log(body),data=Animals[-c(6,16,26),])
> Animals.loglm.sub
Call: lm(formula = log(brain) ~ log(body), data = Animals[-c(6,16,26),])
Coefficients:
(Intercept)    log(body)
     2.1504       0.7523
```

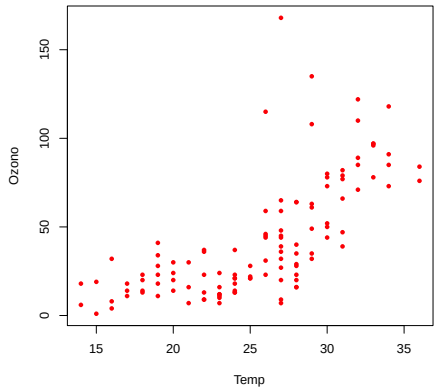
Note-se como os parâmetros da recta se alteram: o declive da recta cresce para mais de 0.75 e a ordenada na origem decresce um pouco. Além disso, podemos analisar o efeito sobre o coeficiente de determinação, através da aplicação do comando `summary` à regressão agora ajustada:

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

Com a exclusão das espécies extintas, a recta de regressão passa a explicar mais de 92% da variabilidade total nos restantes log-pesos do cérebro, a partir dos log-pesos do corpo.

4. (a) O comando `plot(ozono)` produz o gráfico pedido. Um gráfico com alguns embelezamentos adicionais é produzido pelo comando:

```
> plot(ozono, col="red", pch=16, cex=0.8)
```



(b) A linearização duma relação exponencial faz-se logaritmizando:

$$y = a\mathrm{e}^{bx} \quad \Leftrightarrow \quad \ln(y) = \ln(a) + b\,x \ ,$$

que é uma relação linear entre x e $y^* = \ln(y)$.

i. O gráfico de log-`Ozono` contra `Temp` pode ser construído pelo comando:
```
> plot(ozono$Temp, log(ozono$Ozono))
```
Uma tendência linear mais ou menos forte neste gráfico indica que a relação exponencial entre as variáveis originais é adequada. Neste caso, o gráfico corresponde a um coeficiente de correlação entre `Temp` e log-`Ozono` de 0.73.

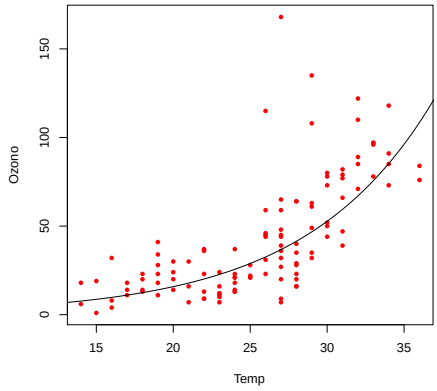
ii. O ajustamento pedido faz-se da seguinte forma:
```
> lm(log(Ozono) ~ Temp, data=ozono)
Call: lm(formula = log(Ozono) ~ Temp, data = ozono)
Coefficients:
(Intercept)         Temp
     0.3558       0.1203
```
O coeficiente de determinação é de cerca de $R^2 = 0.73^2 = 0.53$ (aplicando o comando `summary` ao modelo agora ajustado verifica-se ser $R^2 = 0.5372$), o que significa que a regressão explica pouco mais de 53% da variabilidade dos log-teores de ozono.

iii. O declive estimado da recta $b_1 = 0.1203$ é o coeficiente do expoente, na relação exponencial original, uma vez que estima o parâmetro b que tem esse significado. Já a ordenada na origem da recta ajustada, $b_0 = 0.3558$ corresponde à estimativa de $\ln(a)$, pelo que a constante multiplicativa a da relação exponencial original é: $a = \mathrm{e}^{0.3558} = 1.4273$.

iv. Para prever o *teor de ozono* y utiliza-se a equação exponencial ajustada nas alíneas anteriores, ou seja, a equação $y = 1.4273\,\mathrm{e}^{0.123\,x}$, onde x indica a temperatura máxima. Assim, o valor previsto do teor médio de ozono, num dia de temperatura máxima $x = 25$ é dado por $\hat{y} = 1.4273\,\mathrm{e}^{0.123\times 25} = 28.8839$. É igualmente possível chegar a este valor utilizando directamente a recta de regressão, desde que se tenha em atenção que os valores ajustados por essa recta são de log-teor de ozono, e que se torna necessário desfazer a transformação logarítmica. Assim, o valor *de log-ozono* previsto pela recta, para um dia com temperatura máxima de 25º é dado por: $\widehat{y^*} = \widehat{\ln(y)} = 0.3558 + 0.1203 \times 25 = 3.3633$. E o teor previsto *de ozono* (em ppm) é: $\mathrm{e}^{3.3633} = 28.8843$. A pequena diferenças nos valores obtidos por cada uma das vias acima resulta de erros de arredondamento.

(c) Para sobrepôr a curva exponencial à nuvem de pontos de ozono vs. temperaturas, podem usar-se os seguintes comandos:

```
> plot(ozono, col="red", pch=16, cex=0.8)
> curve(1.4273*exp(0.1203*x), from=10, to=40, add=TRUE)
```



5. (a) Os comandos são:

```
Puromycinexerc<-Puromycin[1:12,]
plot(rate~conc, data=Puromycinexerc)
```

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

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

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

ii. O modelo linearizado ajusta-se através do comando:

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

sendo os resultados obtidos os seguintes:

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

iii. Tendo em conta as relações vistas na alínea anterior, $b_0^* = \frac{1}{a} = 0.0051072$, tem-se $a = 195.802$. Por outro lado, $b_1^* = \frac{b}{a} = 0.0002472$, logo $b = 0.0002472 \times 195.802 = 0.04840225$. Assim, o modelo de Michaelis-Menten ajustado é: $y = \frac{195.802\,x}{0.04840225+x}$. Repare-se que o limite de y quando x tende para $+\infty$ é 195.802, que é assim a estimativa da assintota superior da relação de Michaelis-Menten. O gráfico da relação original sugere que se pode tratar duma subestimação do verdadeiro valor desta assintota horizontal. Traçando a curva correspondente ao ajustamento:

```
curve(195.802*x/(0.048402+x),from=0,to=1.5,add=TRUE)
```

Este exemplo ilustra que pode haver inconvenientes associados à utilização de transformações linearizantes, como indicado nos slides das aulas teóricas.

6. (a) Sabemos que o declive é dado por $b_1 = \frac{cov_{xy}}{s_x^2} = r_{xy} \cdot \frac{s_y}{s_x} = 0.9326 \cdot \sqrt{\frac{0.1077003}{84859.51}} = 0.001050652$. A ordenada na origem é dada por $b_0 = \bar{y} - b_1\,\bar{x} = 0.4010485 - (0.00105652)(2966.882) = -2.716112$. Logo, a recta ajustada é $y = -2.716112 + 0.0010506520\,x$. A proporção de variabilidade total explicada por esta recta de regressão é $R^2 = (r_{xy})^2 = (0.9326)^2 = 0.8697428$, ou seja, cerca de 87% da variabilidade na reflectância é explicada por esta regressão sobre a longitude.

   (b) Vamos construir a transformação sugerida no enunciado. Como $y = \frac{1}{1+\mathrm{e}^{-(c+d\,x)}}$, tem-se $1 - y = 1 - \frac{1}{1+\mathrm{e}^{-(c+d\,x)}} = \frac{\cancel{1}+\mathrm{e}^{-(c+d\,x)}-\cancel{1}}{1+\mathrm{e}^{-(c+d\,x)}}$. Logo, $\frac{y}{1-y} = \frac{\frac{1}{\cancel{1+\mathrm{e}^{-(c+d\,x)}}}}{\frac{\mathrm{e}^{-(c+d\,x)}}{\cancel{1+\mathrm{e}^{-(c+d\,x)}}}} = \frac{1}{\mathrm{e}^{-(c+d\,x)}} = \mathrm{e}^{c+d\,x}$. Assim, $y^* = \ln\left(\frac{y}{1-y}\right) = c + d\,x$. Os parâmetros da relação linearizada são os mesmos que surgem na equação da logística: $b_0^* = c$ e $b_1^* = d$.

   (c) Pede-se o valor de R_*^2 para o modelo linearizado (Nota: não é possível utilizar nem o valor de r_{xy}, nem o valor de s_y^2, dados no enunciado, uma vez que a variável resposta neste modelo foi transformada; seriam precisos os valores de r_{xy^*} ou $s_{y^*}^2$). Por definição $R_*^2 = \frac{SQR}{SQT}$. Sabemos (Exercício teórico 2*d*) que $SQR = (b_1^*)^2\,(n-1)\,s_x^2 = (0.006629)^2 \times 84 \times 84859.51 = 313.239$ (Nota: é possível usar s_x^2, pois a variável X não foi transformada). Por outro lado, $SQT = SQR + SQRE = SQR + QMRE\,(n-2) = 313.239 + 0.6081371 \times 83 = 363.7144$. Logo, $R_*^2 = \frac{313.239}{363.7144} = 0.8612$.

   Alternativamente, e sabendo que $(b_1^*)^2 = (r_{xy^*})^2\,\frac{s_{y^*}^2}{s_x^2} = R_*^2\,\frac{s_{y^*}^2}{s_x^2} = \left(1 - \frac{SQRE}{SQT}\right)\frac{s_{y^*}^2}{s_x^2} = \frac{s_{y^*}^2}{s_x^2} - \frac{QMRE\,(n-2)}{(n-1)\,s_x^2}$. São conhecidos todos os valores desta expressão final, menos $s_{y^*}^2$, pelo que este pode ser obtido e, com o seu auxílio, chegar-se ao valor de R_*^2 para o modelo linearizado, usando a igualdade $(b_1^*)^2 = R_*^2\,\frac{s_{y^*}^2}{s_x^2}$.

7. (a) Proceda como indicado no enunciado para ter disponível a *data frame* `vinhos`.

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

   (c) O ajustamento pedido é:

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

   Trata-se dum péssimo ajustamento, o que não surpreende, tendo em conta a nuvem de pontos deste par de variáveis, obtida na alínea anterior. O coeficiente de determinação é

quase nulo: $R^2 = 0.05608$ e menos de 6% da variabilidade no teor de flavonóides é explicado pela regressão sobre o teor alcoólico.

Como sempre, a Soma de Quadrados Total é o numerador da variância amostral dos valores observados da variável resposta. Ora,

```
> var(vinhos$V8)
[1] 0.9977187
> dim(vinhos)
[1] 178  14
> 177*var(vinhos$V8)
[1] 176.5962
> 177*var(fitted(lm(V8 ~ V2 , data=vinhos)))
[1] 9.903747
> 177*var(residuals(lm(V8 ~ V2 , data=vinhos)))
[1] 166.6925
```

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

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

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

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

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

(e) O modelo pedido no enunciado é:

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

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

Os cinco preditores referidos permitem obter um coeficiente de determinação quase tão bom, embora ligeiramente inferior, ao obtido utilizando apenas o preditor `V7`. O facto de o valor de R^2 ser agora inferior ao valor de R^2 na regressão linear simples de `V8` sobre `V7` não contradiz o facto de submodelos não poderem ter valores do coeficiente de determinação superiores, uma vez que o preditor `V7` não faz parte do grupo de cinco preditores agora considerado (ou seja, o modelo da alínea anterior *não é um submodelo* do que foi considerado nesta alínea).

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

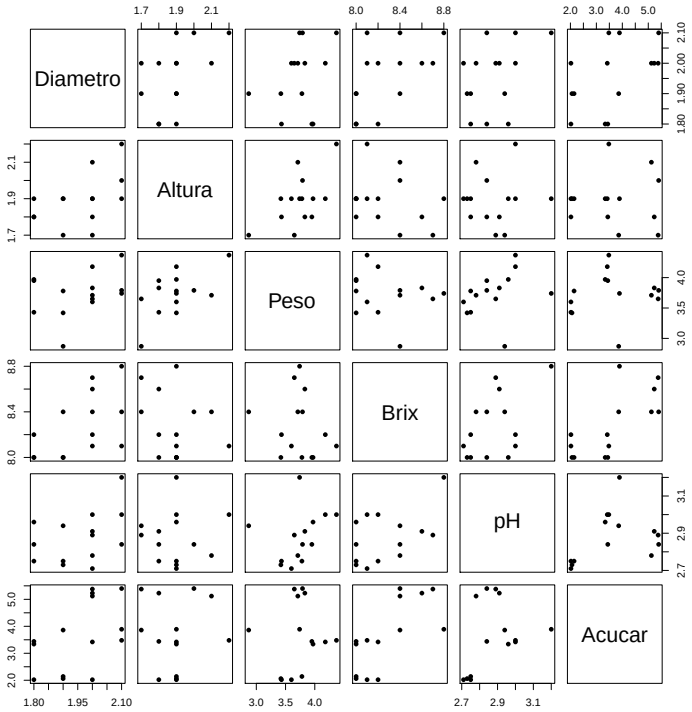
```
> summary(lm(V8~., data=vinhos))

Call:
lm(formula = V8 ~ ., data = vinhos)
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.333e+00  7.558e-01  -1.764  0.07956 .
V2           4.835e-03  5.667e-02   0.085  0.93211
V3          -4.215e-02  3.363e-02  -1.253  0.21191
V4           4.931e-01  1.533e-01   3.216  0.00156 **
V5          -2.325e-02  1.302e-02  -1.786  0.07591 .
V6          -3.559e-03  2.429e-03  -1.465  0.14487
V7           7.058e-01  8.062e-02   8.755 2.33e-15 ***
V9          -1.000e+00  3.061e-01  -3.267  0.00132 **
V10          2.840e-01  6.855e-02   4.143 5.47e-05 ***
V11          1.068e-04  2.201e-02   0.005  0.99614
V12          4.387e-01  2.021e-01   2.171  0.03137 *
V13          3.208e-01  7.639e-02   4.199 4.37e-05 ***
V14          9.557e-05  1.563e-04   0.611  0.54182
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 0.3902 on 165 degrees of freedom
Multiple R-squared:  0.8577,Adjusted R-squared:  0.8474
F-statistic:  82.9 on 12 and 165 DF,  p-value: < 2.2e-16

Também pode calcular SQR e SQRE como:
> 177*var(fitted(lm(V8 ~ . , data=vinhos)))
[1] 151.4735
> 177*var(residuals(lm(V8 ~ . , data=vinhos)))
[1] 25.12269
```

i. O coeficiente de determinação obtido é $R^2 = 0.8577$. De novo, o valor da Soma de Quadrados Total já é conhecido das alíneas anteriores: não depende do modelo ajustado, mas apenas da variância dos valores observados de Y (`V8`, neste exercício), que não se alteraram. Logo, $SQT = 176.5962$, pelo que $SQR = SQT.R^2 = 151.4666$ e $SQRE = SQT - SQR = 25.1296$. Alternativamente, como se pode deduzir da listagem acima, $SQR = (n-1) \cdot s^2_{\hat{y}} = 151.4666$ e $SQRE = (n-1) \cdot s^2_e = 25.12269$. Tem-se, então, $R^2 = \frac{151.4735}{176.5962} = 0.8577$. Refira-se que este valor do coeficiente de determinação *nunca poderia ser inferior ao obtido nas alíneas anteriores*, uma vez que os preditores das

alíneas anteriores formam um subconjunto dos preditores utilizados aqui. Repare como a diferentes modelos para a variável resposta `V8`, correspondem diferentes formas de decompôr a Soma de Quadrados Total comum, $SQT = 176.5962$. Quanto maior a parcela explicada pelo modelo (SQR), menor a parcela associada aos resíduos ($SQRE$), isto é, menor a parcela do que não é explicado pelo modelo.

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

8. A nuvem de pontos e a matriz de correlações dos dados em estudo são apresentados de seguida.




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

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

(a) A equação do hiperlano ajustado (usando os nomes das variáveis como constam da *data frame*) é:

$$Brix \;=\; b_0 + b_1\, Diametro + b_2\, Altura + b_3\, Peso + b_4\, pH + b_5\, Acucar,$$

havendo nesta equação seis parâmetros estimados (os cinco coeficientes das variáveis preditoras e ainda a constante aditiva b_0).

(b) i. Recorrendo ao comando `lm` do `R`, tem-se:

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

ii. Tem-se:

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

A matriz do modelo é a matriz de dimensões $n \times (p+1)$, cuja primeira coluna é uma coluna de n uns e cujas p colunas seguintes são as colunas dadas pelas n observações de cada uma das variáveis preditoras. O vector $\vec{\mathbf{b}}$ dos $p+1$ parâmetros ajustados é dado pelo produto matricial do enunciado: $\vec{\mathbf{b}} = (\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\mathbf{y})$. Um produto matricial no `R` é indicado pelo operador "`%*%`", enquanto que inversa é calculada pelo comando `solve`. A transposta duma matriz é dada pelo comando `t`. Logo, o vector $\vec{\mathbf{b}}$ obtém-se da seguinte forma:

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

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

(c) $b_3 = -0.20453$. Corresponde à variação esperada no teor brix (variável resposta), associada a aumentar em uma unidade a variável preditora *Peso*, mantendo constantes os valores dos restantes preditores. Ou seja, corresponde a dizer que um aumento de $1g$ no peso dum fruto (mantendo iguais os valores dos restantes preditores) está associado a uma diminuição média do teor brix do fruto de 0.20453 graus. As unidades de medida de b_3 são graus brix/g.

(d) A interpretação de b_0 é diferente da dos restantes parâmetros, mas igual ao duma ordenada na origem num regressão linear simples: é o valor médio da variável resposta associado a todos os preditores terem valor nulo. No nosso contexto, o valor estimado $b_0 = 6.08878$ não

tem grande interesse prático ("frutos" sem peso, nem diâmetro ou altura, com valor pH fora a escala, etc...).

(e) Num contexto descritivo, a discussão da qualidade deste ajustamento faz-se com base no coeficiente de determinação $R^2 = \frac{SQR}{SQT}$. Pode calcular-se a Soma de Quadrados Total como o numerador da variância dos valores observados y_i de teor brix: $SQT = (n-1)\, s_y^2 = 13 \times 0.07565934 = 0.9835714$. A Soma de Quadrados da Regressão é calculada de forma análoga à anterior, mas com base na variância dos valores ajustados $\hat{y}_i$, obtidos a partir da regressão ajustada: $SQR = (n-1)\, s_{\hat{y}}^2 = 13 \times 0.06417822 = 0.8343169$. Logo, $R^2 = \frac{0.8343169}{0.9835714} = 0.848$. Os valores usados aqui são obtidos no `R` com os comandos:

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

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

(f) A equação do submodelo sem o preditor *Peso* é:

$$Brix \;=\; b_0^* + b_1^*\, Diametro + b_2^*\, Altura + b_4^*\, pH + b_5^*\, Acucar,$$

O valor do R^2 deste submodelo estará entre o valor obtido para a melhor regressão linear simples para prever o teor de *Brix* com cada uma das restantes variáveis preditoras (neste caso, com o preditor *Acucar*, $R_S^2 = 0.714^2 = 0.51$) e o R^2 do modelo completo (com os 5 preditores, $R_C^2 = 0.848$). Note que na procura da melhor regressão linear simples não pode ter em conta o preditor *Peso*. Como já foi referido anteriormente (ex.7), não é possível prever, a partir da equação ajustada num modelo com todos os preditores, qual será a equação ajustada num submodelo com menos preditores.