Exercícios - Modelos Matemáticos e Aplicações
Modelo Linear - 2018-19

1 Regressão Linear

AVISO: Os conjuntos de dados de alguns exercícios desta secção encontram-se disponíveis na página *web* da disciplina (na Secção *Materiais de Apoio*, *Módulo II*, *Modelo Linear*). Para os exercícios iniciais são dadas instruções detalhadas sobre a forma de aceder a esses dados. Para exercícios posteriores, os dados encontram-se num ficheiro de nome `exerRL.RData` (a extensão indica que este ficheiro foi criado a partir duma sessão do `R`, por meio do comando `save`). Para disponibilizar estes conjuntos de dados deve:

- Descarregar o ficheiro `exerRL.RData` para a directoria onde tem a sua sessão de trabalho (por exemplo, uma pasta chamada `AulasMMA` numa *pen*).
- Executar, numa sessão do `R` nessa directoria, o comando `load("exerRL.RData")`, ou (caso esteja disponível) usar a opção `Load Workspace` do menú `Files`.

1. Com base nos dados do Instituto Nacional de Estatística (INE), foi criado um ficheiro em formato CSV (*Comma separated values*) chamado `Cereais.csv` e contendo a evolução da superfície agrícola utilizada anualmente na produção de cereais para grão (variável `area`, em km^2) em Portugal, no período de 1986 a 2011 (variável `ano`). O ficheiro `Cereais.csv` encontra-se na página *web* da UC, na secção *Materiais de Apoio* $\longrightarrow$ *Módulo II* $\longrightarrow$ *Modelo Linear* $\longrightarrow$ *Dados*. Descarregue o ficheiro `Cereais.csv` para a sua área de trabalho do `R` (que pode ser sempre identificada através do comando `getwd()`). Os dados do ficheiro ficam disponíveis se, numa sessão de trabalho do `R`, fôr dado o seguinte comando:

```
> Cereais <- read.csv("Cereais.csv")
```

   (a) Construa uma nuvem de pontos de superfície agrícola *vs.* ano e comente.

   (b) A partir do gráfico obtido na alínea anterior, sugira um valor para o coeficiente de correlação entre superfície agrícola e ano. Utilize os comandos do `R` para calcular esse mesmo coeficiente de correlação. Comente o seu significado.

   (c) Ajuste uma recta de regressão de superfície agrícola utilizada sobre anos. Discuta o significado dos parâmetros da recta ajustada, no contexto do problema sob estudo.

   (d) Comente a qualidade da recta obtida, calculando o respectivo coeficiente de determinação e interpretando o valor obtido.

   (e) Trace a recta de regressão ajustada em cima da nuvem de pontos e comente.

   (f) Calcule a Soma de Quadrados Total (SQT), a partir do cálculo da variância amostral de y.

   (g) Calcule o valor da Soma de Quadrados da Regressão (SQR).

   (h) Calcule a Soma de Quadrados dos Resíduos (SQRE), directamente a partir dos resíduos, e verifique numericamente a relação fundamental da Regressão Linear: SQT=SQR+SQRE.

   (i) Altere as unidades de medida da variável `area`, de km^2 para hectares (`area` $\rightarrow$ `area`$\times 100$). Ajuste novamente a regressão, após efectuar esta alteração. O que aconteceu aos parâmetros estimados e ao coeficiente de determinação R^2? Comente.

   (j) De novo a partir dos dados originais, transforme a variável `ano` num contador dos anos do estudo (`ano` $\rightarrow$ `ano`-1985). Ajuste novamente a regressão, após efectuar esta alteração. O que aconteceu aos parâmetros estimados e ao coeficiente de determinação R^2? Comente.

2. O ficheiro `Azeite.xls` encontra-se disponível na página *web* da disciplina (secção *Materiais de Apoio*, *Modelo Linear*, *Dados*). Trata-se duma folha de cálculo, comum a aplicações de escritório como o LibreOffice, OpenOffice ou MicrosoftOffice. A folha de cálculo contém dados relativos à produção de azeite em Portugal no período 1995-2010, disponibilizados pelo Instituto Nacional de Estatística (www.ine.pt). As colunas "Azeitona" e "Azeite" correspondem à produção de azeitona oleificada (em t) e azeite (em hl), respectivamente.

   (a) Abra o ficheiro `Azeite.xls` e guarde a folha de cálculo num ficheiro `Azeite.txt` (utilizando o *Save as* com a opção *Ficheiro de Texto*). Coloque esse ficheiro na pasta de trabalho do R.

   (b) Numa sessão do `R`, guarde os dados do ficheiro `Azeite.txt` (criado na alínea anterior) numa *data frame* de nome `azeite`, através do comando:

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

   (c) Crie a nuvem de pontos relacionando as produções de Azeite (eixo vertical, variável y) e Azeitona (eixo horizontal, variável x).

   (d) Com base na nuvem de pontos, sugira um valor para o coeficiente de correlação entre as duas variáveis. Avalie a sua sugestão calculando o valor de r_{xy}. Comente o valor obtido.

   (e) Calcule as estimativas de mínimos quadrados para os parâmetros da recta de regressão, e comente o seu significado.

   (f) Calcule a precisão da recta de regressão estimada de y sobre x e comente o valor obtido.

3. Mostre que, para quaisquer conjuntos de n valores, $\{x_i\}_{i=1}^n$, e $\{y_i\}_{i=1}^n$ de médias $\overline{x}$ e $\overline{y}$, respectivamente, se tem:

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

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

4. Mostre que, numa regressão linear simples, baseada em n pares de observações $\{(x_i, y_i)\}_{i=1}^n$, se tem:

   (a) a igualdade da média dos valores observados e da média dos valores ajustados de y.

   (b) a média dos resíduos ($e_i = y_i - \hat{y}_i$) é nula.

   (c) o declive da recta de regressão de y sobre x pode-se escrever em termos do desvio padrão de cada variável e do coeficiente de correlação entre as duas variáveis, sendo dado por:

   $$b_1 \;=\; r_{xy} \cdot \frac{s_y}{s_x}\ .$$

   (d) o coeficiente de determinação R^2 é igual ao quadrado do coeficiente de correlação entre as observações da variável preditora x e da variável resposta y.

   (e) o quadrado do coeficiente de correlação entre os n valores observados y_i e os n correspondentes valores ajustados, $\hat{y}_i$, é também igual ao coeficiente de determinação: $(r_{y\hat{y}})^2 = R^2$.

5. O programa `R` disponibiliza um grande número de módulos adicionais, entre os quais o módulo `MASS`, que pode ser carregado para uma sessão de trabalho mediante o comando `library(MASS)`.

   Considere o conjunto de dados `Animals`, disponível no referido módulo `MASS`, onde se listam pesos médios dos cérebros (em g) e dos corpos (em kg) para 28 espécies animais. Pretende-se estudar uma relação entre pesos do cérebro (variável resposta, y) e pesos do corpo (variável preditora, x).

   (a) Construa a nuvem de pontos de pesos do corpo (eixo horizontal) e pesos do cérebro (eixo vertical). Calcule o coeficiente de correlação correspondente e comente.

(b) Construa a nuvem de pontos *dos logaritmos* (naturais) do pesos do corpo e do cérebro. Calcule os coeficientes de correlação e de determinação associados à relação entre $\ln(x)$ e $\ln(y)$. Interprete os valores obtidos. Como se explica que o Coeficiente de Determinação não seja particularmente elevado, sendo evidente a partir da nuvem de pontos que existe uma boa relação linear entre log-peso do corpo e log-peso do cérebro para a generalidade das espécies?

(c) Considere uma relação linear entre $\ln(y)$ e $\ln(x)$. Explicite a relação de base correspondente entre as variáveis originais (não logaritmizadas). Comente.

Nas alíneas seguintes considere sempre os *dados logaritmizados*.

(d) Ajuste a recta de regressão de log-peso do cérebro sobre log-peso do corpo (utilizando a totalidade das observações). Trace essa recta sobre a nuvem de pontos e comente.

(e) Considere agora a estimativa para o declive da recta, $b_1 = 0.49599$. Qual o significado biológico deste valor, quer na relação entre variáveis logaritmizadas, quer na relação entre as variáveis originais (não logaritmizadas)?

(f) Considere a nuvem de pontos das variáveis logaritmizadas. Identifique os três pontos que se destacam na parte inferior direita da nuvem. (**NOTA:** explore o comando `identify` do R). Comente.

Nas restantes alíneas, considere apenas os dados (logaritmizados) respeitantes a espécies que *não sejam de dinossáurios*.

(g) Ajuste a recta de regressão de log-peso do cérebro sobre log-peso do corpo. Trace essa recta sobre a nuvem de pontos e comente. (NOTA: Utilize a nuvem de pontos com a totalidade das espécies, a fim de melhor compreender o efeito da exclusão das três espécies de dinossáurios sobre a recta ajustada).

(h) Compare a recta obtida na alínea 5g) com os resultados obtidos nas alíneas anteriores e comente. Comente também a elevação considerável no valor do coeficiente de determinação da recta agora obtida com a recta obtida na alínea 5d).

(i) Considere agora a estimativa para o declive da recta de regressão clássica após a exclusão das três espécies de dinossáurios, $b_1 = 0.75226$. Qual o significado biológico deste valor, quer na relação entre variáveis logaritmizadas, quer na relação entre as variáveis originais (não logaritmizadas)?

6. Num estudo sobre poluição numa grande cidade, foram efectuadas medições, em 116 dias, da quantidade de ozono no ar (em partes por mil milhões) às 14h00 e da temperatura máxima (em ^{o}C) no respectivo dia. Essas observações encontram-se num ficheiro em formato `csv` de nome `ozono.csv`, que se encontra disponível na página *web* da disciplina e que, após ser descarregado para a área de trabalho da sua sessão do `R`, pode ser guardado através do comando `read.csv`:

```
> ozono <- read.csv("ozono.csv")
```

(a) Construa a nuvem de pontos de ozono (eixo vertical) *vs.* temperatura máxima (eixo horizontal).

(b) Tendo em conta a curvatura observada no gráfico, foi sugerido o ajustamento dum modelo exponencial, da forma $y = a\,\mathrm{e}^{b\,x}$.

i. Construa a nuvem de pontos com as transformações adequadas para verificar se o modelo exponencial é, efectivamente, uma boa opção.

ii. Ajuste o modelo *linearizado* recorrendo ao comando `lm` do `R`. Determine o respectivo coeficiente de determinação e comente.

iii. Interprete os parâmetros da recta que ajustou, directamente em termos do modelo exponencial.

iv. Indique, justificando, qual o teor médio de ozono (em partes por mil milhões) estimado pelo modelo ajustado, para um dia em que a temperatura máxima seja de 25°C.

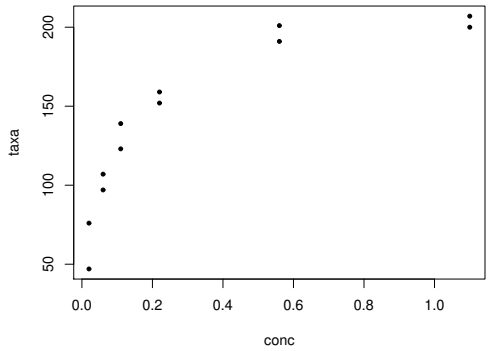
(c) Considere novamente a nuvem de pontos original. Trace a curva exponencial correspondente ao ajustamento efectuado na alínea anterior.

7. Num estudo sobre reacções enzimáticas, procura-se analisar a "velocidade" da reacção em células tratadas com Puromicina. Para diferentes concentrações do substrato (variável *conc*), medidas em partes por milhão (ppm), registou-se o número de emissões radioactivas por minuto, e a partir destas calculou-se a taxa inicial ou "velocidade" da reacção, em contagens/minuto/minuto (variável *taxa*). Os resultados obtidos são dados na tabela seguinte e encontram-se *nas duas primeiras colunas* e *doze primeiras linhas* da *data frame* `Puromycin` do `R`, com as designações `conc` e `rate`, respectivamente:

conc	0.02	0.02	0.06	0.06	0.11	0.11	0.22	0.22	0.56	0.56	1.10	1.10
taxa	76	47	97	107	123	139	159	152	191	201	207	200

A relação entre taxas da reacção e concentrações do substrato é representada no gráfico à direita. Admite-se que o modelo de Michaelis-Menten é adequado à descrição da relação referida, e decide-se usar este modelo com a seguinte parametrização (onde y representa a *taxa* e x a concentração *conc*),

$$y = \frac{ax}{b+x} \qquad (a > 0, b > 0 \quad \text{e} \quad x > 0) \ .$$



(a) Mostre que o modelo referido pode ser linearizado, indicando a relação linearizada e as transformações de variáveis necessárias.

(b) Ajuste o modelo linearizado que escolheu na alínea anterior, através do comando `lm` do `R`.

(c) Estime os parâmetros a e b na relação original no modelo de Michaelis-Menten. Como interpreta o valor estimado do parâmetro a? Trace a curva de Michaelis-Menten obtida por cima da nuvem de pontos na escala original. Comente.

8. A Floresta Experimental H.J. Andrews, no Estado norte-americano do Oregon, disponibiliza numerosos conjuntos de dados florestais (https://andrewsforest.oregonstate.edu/data). Um desses conjuntos de dados, referente à medição de nutrientes em bacias hidrográficas, tem a designação `TN025` (seguir os apontadores *Data Catalogue* e depois fazer um *Text Search* usando a designação). Os dados referem-se a 117 medições de concentração de diversos nutrientes. O ficheiro em formato `CSV`, com os dados está disponível na página da disciplina. Mais pormenores relativos às condições de recolha dos dados e à natureza das variáveis encontram-se no *website* da Floresta Experimental.

(a) Proceda à leitura dos dados para uma sessão do `R`. Inspeccione a natureza das 25 colunas do conjunto de dados. Os valores observados das concentrações de nutrientes encontram-se nas colunas 11 a 24, e são identificados pelos seus símbolos químicos (Atenção: o sódio tem a designação "`NA.`", com um ponto final, a fim de distinguir do símbolo `NA` utilizado no `R` para identificar valores omissos). Estas concentrações são todas medidas em $mg\ kg^{-1}$, excepto o azoto (`N`) e o carbono (`C`), que são dadas em percentagens. Na coluna 8, de nome `TYPE`, é indicado o tipo de material lenhoso no qual foram feitas as observações.

(b) Com base na matriz de correlações entre as diferentes concentrações de nutrientes, escolha o melhor preditor linear da concentração de fósforo (variável `P`).

i. Ajuste a recta de regressão de `P` sobre o preditor que escolheu. Discuta a qualidade desse ajustamento, a partir da listagem produzida pelo comando `lm`.

ii. Construa a nuvem dos 117 pontos e trace a recta de regressão. Comente o resultado. Em particular, identifique as observações a que corresponde a coluna de pontos que surge do lado esquerdo do gráfico. Quantas são as observações nessa coluna? Comente.

iii. Construa os gráficos de resíduos e outros diagnósticos para a regressão ajustada. Comente, tendo em conta também a sua resposta à alínea anterior. Identifique o ponto com uma muito elevada distância de Cook, e discuta-o.

(c) Mantendo a variável resposta fósforo, considere agora o preditor potássio (`K`). Construa a nuvem de pontos respectiva. Comente-a. Ajuste uma recta de regressão e comente os resultados.

(d) Considere agora uma regressão linear simples da transformação logarítmica, quer da variável resposta `P`, quer do preditor `K`.

i. Construa a nuvem de pontos correspondente e comente.

ii. Ajuste a recta de regressão linear e comente a sua qualidade, com base nos resultados produzidos pelo comando `lm`. Em particular, diga se o valor do coeficiente de determinação obtido é comparável com o valor obtido na alínea 8c).

iii. Inspeccione os gráficos de resíduos e outros diagnósticos. Comente.

iv. Deduza a que curva corresponde a recta ajustada neste modelo linearizado, quando se regressa à escala das variáveis originais (`K` e `P`). Trace essa curva sobre a nuvem de pontos obtida na alínea 8c). Comente o resultado, apontando lições de interesse geral.

9. O repositório de dados (`http://archive.ics.uci.edu/ml/`) da Universidade da Califórnia, Irvine, contém muitos conjuntos de dados em formato *comma separated value (csv)*, que podem ser facilmente lidos através do comando `read.csv` da aplicação R. Considere o conjunto de dados "Wine recognition data" desse repositório (fonte: Forina, M. et al, *PARVUS - An Extendible Package for Data Exploration, Classification and Correlation*. Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy) que contém os resultados da análise química de vinhos de três castas de uma determinada região de Itália. As 14 colunas da tabela de dados correspondem respectivamente às variáveis casta (factor V1 com 3 níveis), teor alcoólico (V2), teor de ácido málico (V3), cinzas (V4), alcalinidade das cinzas (V5), teor de magnésio (V6), índice de fenóis totais (V7), teor de flavonóides (V8), teor de outros fenóis (V9), teor de proantocianidinas (V10), intensidade de cor (V11), matiz (V12), razão de densidades ópticas em duas frequências, OD280/OD315, (V13) e teor de prolina (V14).

Proceda à leitura dos dados através do comando

```
vinhos<-read.csv("http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data",
header=FALSE)
```

e exclua da tabela de dados a primeira coluna (um factor que indica a casta) criando uma nova *data frame*, através do comando `vinho.RLM<-vinhos[,-1]`.

(a) Há interesse em modelar o teor de flavonóides (variável `V8`), um antioxidante de medição difícil e dispendiosa. Nessa perspectiva, comente o resultado do comando `plot(vinho.RLM)`.

(b) A partir da matriz de correlações entre as variáveis sob estudo, diga qual a melhor recta de regressão simples para prever o teor de flavonóides (variável `V8`). Para a regressão linear simples que escolher, determine o coeficiente de determinação e realize a correspondente decomposição da soma dos quadrados total.

(c) A variável preditora utilizada na alínea anterior também não é simples de medir, tal como sucede com as variáveis `V9` e `V10`. Foi sugerido procurar um modelo de regressão linear múltipla para a variável resposta teor de flavonóides (`V8`) que não utiliza esses preditores. Foi proposto um modelo com cinco variáveis preditoras: `V4`, `V5`, `V11`, `V12` e `V13`. Ajuste este modelo, e comente o respectivo coeficiente de determinação, comparando-o com o R^2 do modelo da alínea anterior. O comando do `R` para ajustar esta regressão linear múltipla é:

```
> lm(V8 ~ V4 + V5 + V11 + V12 + V13 , data=vinho.RLM)
```

(d) Ajuste uma regressão linear múltipla do teor de flavonóides (variável `V8`) sobre todas as restantes variáveis com o comando `summary(lm(V8 ∼ . , data=vinho.RLM))`.

i. Use o valor do coeficiente de determinação obtido com esse comando para determinar a decomposição da soma dos quadrados totais. Comente os resultados.

ii. Compare os coeficientes estimados das variáveis preditoras com os correspondentes coeficientes das variáveis preditoras presentes nos modelos anteriores. Comente.

10. Num estudo sobre framboesas realizado na Secção de Horticultura do ISA foram analisados frutos de 14 plantas diferentes, no que respeita a 6 diferentes variáveis. As variáveis observadas foram: (i) o *diâmetro* dos frutos (em *cm*); (ii) a sua *altura* (em *cm*); (iii) o seu *peso* (em *g*); (iv) o seu teor de sólidos solúveis, *brix* (em graus Brix); (v) o seu *pH*; (vi) o seu teor de *açúcar*, exceptuando a sacarose (em $g/100ml$). Os dados encontram-se na *data frame* `brix`, que se encontra no ficheiro `exerRL.RData` e pode ser obtida como indicado no aviso geral, no início destes enunciados. Os resultados médios de cada variável, para as framboesas de cada planta são:

	Diametro	Altura	Peso	Brix	pH	Acucar
1	2.0	2.1	3.71	8.4	2.78	5.12
2	2.1	2.0	3.79	8.4	2.84	5.40
3	2.0	1.7	3.65	8.7	2.89	5.38
4	2.0	1.8	3.83	8.6	2.91	5.23
5	1.8	1.8	3.95	8.0	2.84	3.44
6	2.0	1.9	4.18	8.2	3.00	3.42
7	2.1	2.2	4.37	8.1	3.00	3.48
8	1.8	1.9	3.97	8.0	2.96	3.34
9	1.8	1.8	3.43	8.2	2.75	2.02
10	1.9	1.9	3.78	8.0	2.75	2.14
11	1.9	1.9	3.42	8.0	2.73	2.06
12	2.0	1.9	3.60	8.1	2.71	2.02
13	1.9	1.7	2.87	8.4	2.94	3.86
14	2.1	1.9	3.74	8.8	3.20	3.89

(a) Construa as nuvens de pontos correspondentes a cada possível par de variáveis. Calcule os coeficientes de correlação correspondentes a cada gráfico. Comente.

(b) Pretende-se modelar o teor de *Brix* a partir das restantes variáveis observadas. Escreva a equação de base do modelo de regressão linear múltipla com *Brix* como variável resposta e as restantes variáveis como preditoras. Quantos parâmetros tem este modelo?

(c) Determine o valor das estimativas dos parâmetros do modelo indicado na alínea anterior.

(d) Discuta o significado biológico da estimativa do coeficiente da variável *Peso*. Quais são as unidades de medida desta estimativa?

(e) Discuta o significado da estimativa do parâmetro β_0. Comente.

(f) Discuta o coeficiente de determinação do modelo. Em particular, compare o coeficiente de determinação da regressão múltipla com os coeficientes de determinação associados às regressões lineares simples (com a mesma variável resposta) da alínea 10a). Comente.

(g) Utilize o comando `model.matrix` do `R` para construir a matriz $\mathbf{X}$ do modelo. Com base nessa matriz, obtenha o vector $\vec{\mathbf{b}}$ dos parâmetros ajustados, através da sua fórmula, $\vec{\mathbf{b}} = (\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\vec{\mathbf{y}})$, onde $\vec{\mathbf{y}}$ é o vector das observações da variável resposta.

11. Para fins comerciais, é hábito estimar o peso de ameixas a partir dos seus diâmetros. A fim de se obter uma relação entre diâmetro e peso, válida para uma determinada variedade, foram calibrados (diametro em *mm*) e pesados (em *g*) $n = 41$ frutos, tendo-se obtido os valores indicados no objecto `ameixas` (disponível no ficheiro `exerRL.RData`, referido no aviso inicial).

(a) Construa a nuvem de pontos de *diametro* (X) contra *peso* (Y). Comente a relação de fundo obtida.

(b) Ajuste um polinómio de segundo grau à relação entre as duas variáveis: $y = \beta_0 + \beta_1 x + \beta_2 x^2$. Indique as estimativas dos parâmetros deste modelo. Trace a parábola ajustada por cima da nuvem de pontos obtida na alínea anterior.

(c) Inspeccione os resíduos do modelo ajustado e comente.

(d) Investigue se vale a pena considerar um polinómio de terceiro grau na relação entre diâmetro e peso dos frutos.

12. Considere uma regressão linear simples duma variável Y sobre uma variável X, com base em n pares de observações $\{(x_i, y_i)\}_{i=1}^n$. Considere ainda a notação utilizada nas aulas (em que $\mathbf{X}$ indica uma matriz com duas colunas: uma coluna de n uns, e uma coluna com os n valores x_i da variável preditora X; e $\vec{\mathbf{y}}$ indica um vector com os n valores da variável Y). Mostre que:

(a) $\mathbf{X}^t\vec{\mathbf{y}} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} = \begin{bmatrix} n\overline{y} \\ \sum_{i=1}^n x_i y_i \end{bmatrix} = \begin{bmatrix} n\overline{y} \\ (n-1)\, cov_{xy} + n\overline{x}\overline{y} \end{bmatrix}$.

(b) $\mathbf{X}^t\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} = \begin{bmatrix} n & n\,\overline{x} \\ n\,\overline{x} & (n-1)\, s_x^2 + n\,\overline{x}^2 \end{bmatrix}$.

(c) $(\mathbf{X}^t\mathbf{X})^{-1} = \frac{1}{n(n-1)\, s_x^2} \begin{bmatrix} (n-1)\, s_x^2 + n\overline{x}^2 & -n\,\overline{x} \\ -n\,\overline{x} & n \end{bmatrix}$.

(d) Deduza a partir do facto que $\vec{\mathbf{b}} = (\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\vec{\mathbf{y}})$, as fórmulas para b_0 e b_1 na Regressão Linear Simples.

NOTA: Tenha em atenção que:

$$\begin{aligned} (n-1)\, cov_{xy} &= \sum_i (x_i - \overline{x})(y_i - \overline{y}) = \sum_i x_i y_i - n\overline{x}\,\overline{y}\ ; \\ (n-1)\, s_x^2 &= \sum_i (x_i - \overline{x})^2 = \sum_i x_i^2 - n\overline{x}^2\ . \end{aligned}$$

13. (a) Mostre, a partir da sua definição, que a matriz de projecção ortogonal $\mathbf{H}$ numa regressão linear múltipla é idempotente ($\mathbf{H}\,\mathbf{H} = \mathbf{H}$) e simétrica ($\mathbf{H}^t = \mathbf{H}$).

(b) Sabendo que qualquer vector que pertence ao subespaço $\mathcal{C}(\mathbf{X})$ do espaço das colunas da matriz $\mathbf{X}$, num modelo de regressão linear múltipla, se pode escrever como o produto $\mathbf{X}\vec{\mathbf{a}}$, para algum vector de coeficientes $\vec{\mathbf{a}}$, mostre que os vectores pertencentes a $\mathcal{C}(\mathbf{X})$ permanecem invariantes quando projectados sobre esse mesmo subespaço, isto é, mostre que $\mathbf{H}\mathbf{X}\vec{\mathbf{a}} = \mathbf{X}\vec{\mathbf{a}}$.

(c) Mostre, a partir da expressão do vector dos valores ajustados de Y, $\hat{\vec{\mathbf{y}}} = \mathbf{H}\vec{\mathbf{y}}$, que a média dos valores ajustados de Y, $\{\hat{y}_i\}_{i=1}^n$, é igual à média dos valores observados, $\{y_i\}_{i=1}^n$.

(d) Mostre que a soma dos resíduos, em qualquer regressão linear, tem de ser zero.

Na resolução dos Exercícios seguintes, de natureza inferencial, admita válido o Modelo Linear.

14. Considere os dados das medições sobre lírios (*data frame* `iris`), considerando que se trata da concretização duma amostra aleatória extraída duma população mais vasta. Considere, em particular, a relação entre largura da pétala (`Petal.Width`, variável y) e comprimento da pétala (`Petal.Length`, variável x), ambas em *cm*. Responda às seguintes alíneas.

    (a) Obtenha estimativas das variâncias e desvios padrões dos estimadores dos parâmetros da recta, β_0 e β_1.

    (b) Obtenha um intervalo a 95% de confiança para o declive β_1 da correspondente recta populacional.

    (c) Obtenha um intervalo a 95% de confiança para a ordenada na origem β_0 da recta populacional.

    (d) Utilize um teste de hipóteses para validar a seguinte afirmação: "por cada centimetro a mais no comprimento da pétala, a largura da pétala cresce, em média, $0.5cm$".

    (e) Utilize um teste de hipóteses para validar a seguinte afirmação: "por cada centimetro a mais no comprimento da pétala, a largura da pétala cresce, em média, menos de $0.5cm$".

    (f) Utilize um teste de hipóteses sobre o declive da recta populacional β_1 para validar a seguinte afirmação: "não existe uma relação linear significativa entre comprimentos e larguras das pétalas, nos lírios".

    (g) Valide de novo a afirmação anterior, mas agora utilizando um teste de ajustamento global do Modelo (teste F).

    (h) Preveja o valor esperado da largura da pétala para lírios cuja pétala tenha comprimento 4.5cm. Construa um intervalo de confiança para esse valor esperado.

    (i) Construa um intervalo de predição (95%) associado à largura duma pétala cujo comprimento seja 4.5cm. Compare com o intervalo de confiança obtido na alínea anterior e comente.

    (j) Estude os gráficos dos resíduos para detectar a existência de eventuais problemas com os pressupostos do modelo. Comente as suas conclusões.

    (k) Para cada uma das seguintes transformações dos dados, verifique os efeitos sobre os parâmetros ajustados e sobre o coeficiente de determinação. Comente.

        i. os comprimentos das pétalas são dados em milímetros ($x \to 10 \times x$), mantendo-se as larguras (y) em centimetros.

        ii. as larguras das pétalas são dadas em milímetros ($y \to 10 \times y$), mantendo-se os comprimentos (x) em centimetros.

        iii. em simultâneo, larguras e comprimentos das pétalas são expressas em milimetros ($x \to 10 \times x$ e $y \to 10 \times y$).

15. Seja $\vec{\mathbf{Z}}_{k\times 1}$ um vector aleatório. Mostre que se verificam as seguintes propriedades:

    (a) $E[\alpha\vec{\mathbf{Z}}] = \alpha E[\vec{\mathbf{Z}}]$, sendo α um escalar (não aleatório).

    (b) $E[\vec{\mathbf{Z}} + \vec{\mathbf{a}}] = E[\vec{\mathbf{Z}}] + \vec{\mathbf{a}}$, sendo $\vec{\mathbf{a}}$ um vector não aleatório.

    (c) $V[\alpha\vec{\mathbf{Z}}] = \alpha^2 V[\vec{\mathbf{Z}}]$, sendo α um escalar (não aleatório).

    (d) $V[\vec{\mathbf{Z}} + \vec{\mathbf{a}}] = V[\vec{\mathbf{Z}}]$, sendo $\vec{\mathbf{a}}$ um vector não aleatório.

    (e) Considere um segundo vector aleatório $\vec{\mathbf{U}}_{k\times 1}$. Mostre que $E[\vec{\mathbf{Z}} + \vec{\mathbf{U}}] = E[\vec{\mathbf{Z}}] + E[\vec{\mathbf{U}}]$.

16. A estatística do teste de ajustamento global do modelo (teste F) é dada por $F = \frac{QMR}{QMRE}$. O Coeficiente de Determinação define-se como $R^2 = \frac{SQR}{SQT}$. Com base nestas definições, e tendo em conta as propriedades das somas de quadrados,

    (a) Mostre que a estatística F se pode escrever também como:

    $$F = \frac{n-(p+1)}{p} \cdot \frac{R^2}{1-R^2}$$

    (b) Verifique, a partir da expressão anterior, que a estatística F é (para n fixo) uma *função crescente do Coeficiente de Determinação*. Interprete esse facto, em termos do significado de R^2 e a natureza do teste de ajustamento global.

17. Considere os dados do Exercício 5 (`Animals`). Trabalhe sempre com os *dados logaritmizados*, para a totalidade das espécies.

    (a) Considere a presença de erros aleatórios na relação linear entre as variáveis logaritmizadas: $\log(Y) = \beta_0 + \beta_1 \log(x) + \epsilon$. Qual a consequência para a relação entre as variáveis originais (não logaritmizadas) associada à presença dos erros aleatórios? E como se traduzem os restantes pressupostos do Modelo de Regressão Linear em termos dessa relação entre as variáveis originais (não logaritmizadas)?

    (b) Efectue um teste de ajustamento global da regressão ajustada na alínea 5d). Como se explica que o valor do coeficiente de determinação não seja particularmente bom, quando o teste F sugere que a rejeição da hipótese nula do teste de ajustamento é muito enfática?

    (c) Construa um intervalo de confiança a 95% para o declive da recta que relaciona log-peso do corpo e log-peso do cérebro. É admissível falar-se numa relação isométrica entre peso do corpo e peso do cérebro?

    (d) Estude os gráficos dos resíduos para detectar a existência de eventuais problemas com os pressupostos do modelo. Em particular, veja como a presença das três espécies de natureza diferente das restantes está a afectar estes gráficos.

    Nas restantes alíneas, considere apenas os dados (logaritmizados) respeitantes a espécies que *não sejam de dinossáurios*.

    (e) Construa um intervalo de confiança a 95% para o declive da recta que relaciona log-peso do corpo e log-peso do cérebro. Perante o novo valor de b_1, será agora admissível falar-se numa relação isométrica entre peso do corpo e peso do cérebro?

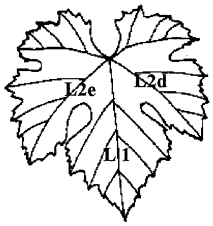
    (f) Preveja o valor esperado do log-peso do cérebro para espécies com peso de corpo igual a $250kg$. Construa um intervalo de confiança para esse valor esperado.

    (g) Construa um intervalo de predição associado ao log-peso do cérebro duma espécie cujo peso do corpo seja $250kg$. Como obter um intervalo de predição associado *ao peso* do cérebro?

    (h) Estude os gráficos dos resíduos para detectar a existência de eventuais problemas com os pressupostos do modelo. Comente as suas conclusões, tendo presente os gráficos análogos obtidos com a presença das 3 espécies de dinossáurios.

18. A medição rigorosa de áreas foliares faz-se através de técnicas que exigem que as folhas sejam arrancadas. Pretende-se estimar áreas foliares **(Área)** de castas de videiras, utilizando variáveis preditoras que possam ser medidas sem destruir as folhas. Concretamente, deseja-se prever as áreas foliares a partir de três medições em cada folha:

- o comprimento da nervura principal **(NP)**;
- o comprimento da nervura lateral esquerda **(NLesq)**; e
- o comprimento da nervura lateral direita **(NLdir)**.



Foram consideradas três diferentes **Castas** de videiras: Fernão Pires, Vital e Água Santa, mas deseja-se obter um modelo único para todas as castas. Na Secção de Horticultura do ISA foram seleccionadas 200 folhas de cada casta, e para cada folha obtiveram-se as medições de cada variável preditora (em cm), bem como a medição da área foliar (em cm^2) pela técnica mais rigorosa. Os dados obtidos constam do objecto `videiras`. As 6 primeiras linhas da `data frame` em questão são:

```
          Casta NLesq   NP NLdir Area
1  Fernao Pires  11.4 13.8  10.7  200
2  Fernao Pires   8.8  9.1   9.4  126
3  Fernao Pires  13.2 14.5  13.0  274
4  Fernao Pires  11.7 13.8  10.7  198
5  Fernao Pires   9.7 12.0  10.6  160
6  Fernao Pires  12.0 11.5  11.6  236
```

(a) Desenhe as nuvens de pontos para cada par de variáveis observadas. Comente.

(b) Calcule a matriz de correlações entre as 4 variáveis observadas. Comente.

(c) Descreva o Modelo de Regressão Linear Múltipla associado ao problema.

(d) Ajuste a regressão múltipla referida na alínea anterior e comente. Em particular, teste o ajustamento global do modelo.

(e) Admitindo a validade do modelo, teste, com um nível de significância de $\alpha = 0.01$, a hipótese de que, a cada centímetro adicional na nervura principal (e sem alterar os comprimentos das nervuras laterais) corresponda um aumento da área foliar de 7 cm^2. Repita o teste, mas agora utilizando um nível de significância $\alpha = 0.05$. Comente.

(f) Será admissível considerar que os coeficientes das duas nervuras laterais são iguais? Justifique formalmente.

(g) Foram medidas as nervuras de três novas folhas, na videira. Os resultados obtidos foram:

No. folha	NP	NLesq	NLdir
1	12.1	11.6	11.9
2	10.6	10.1	9.9
3	15.1	14.9	14.0

Para cada nova folha, calcule:

i. o valor estimado da área foliar;

ii. um intervalo de confiança (95%) para o valor esperado da área foliar associado a esses valores das variáveis preditoras;

iii. um intervalo de predição (95%) para o valor da área foliar de cada folha individual.

(h) Estude os resíduos do ajustamento efectuado. Comente.

(i) Tendo em consideração a forma das folhas e a natureza das nervuras, um investigador sugere que uma regra simples para estimar a área foliar seria a de calcular o produto do comprimento da nervura principal com a média dos comprimentos das nervuras laterais.

i. Veja se esta regra simples pode ser linearizada e, em caso afirmativo, escreva a relação de base no modelo linear resultante.

ii. Ajuste um modelo linear que permite validar a proposta do investigador. Comente as suas conclusões.

iii. Estude os resíduos do ajustamento efectuado na alínea anterior. Comente.

19. No relatório CAED – Report 17, Iowa State University, 1963, são mostrados os seguintes dados meteorológicos e de produção de milho para o estado de Iowa (EUA), nos anos 1930–1962.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	y
Ano		Prec. 'pré-estação' (in.)	Temp. Maio (^{o}F)	Prec. Junho (in.)	Temp. Junho (^{o}F)	Prec. Julho (in.)	Temp. Julho (^{o}F)	Prec. Agosto (in.)	Temp. Agosto (^{o}F)	Prod. milho (bu/acre)
1930	1	17.75	60.2	5.83	69.0	1.49	77.9	2.42	74.4	34.0
1931	2	14.76	57.5	3.83	75.0	2.72	77.2	3.30	72.6	32.9
1932	3	27.99	62.3	5.17	72.0	3.12	75.8	7.10	72.2	43.0
1933	4	16.76	60.5	1.64	77.8	3.45	76.1	3.01	70.5	40.0
1934	5	11.36	69.5	3.49	77.2	3.85	79.7	2.84	73.4	23.0
1935	6	22.71	55.0	7.00	65.9	3.35	79.4	2.42	73.6	38.4
1936	7	17.91	66.2	2.85	70.1	0.51	83.4	3.48	79.2	20.0
1937	8	23.31	61.8	3.80	69.0	2.63	75.9	3.99	77.8	44.6
1938	9	18.53	59.5	4.67	69.2	4.24	76.5	3.82	75.7	46.3
1939	10	18.56	66.4	5.32	71.4	3.15	76.2	4.72	70.7	52.2
1940	11	12.45	58.4	3.56	71.3	4.57	76.7	6.44	70.7	52.3
1941	12	16.05	66.0	6.20	70.0	2.24	75.1	1.94	75.1	51.0
1942	13	27.10	59.3	5.93	69.7	4.89	74.3	3.17	72.2	59.9
1943	14	19.05	57.5	6.16	71.6	4.56	75.4	5.07	74.0	54.7
1944	15	20.79	64.6	5.88	71.7	3.73	72.6	5.88	71.8	52.0
1945	16	21.88	55.1	4.70	64.1	2.96	72.1	3.43	72.5	43.5
1946	17	20.02	56.5	6.41	69.8	2.45	73.8	3.56	68.9	56.7
1947	18	23.17	55.6	10.39	66.3	1.72	72.8	1.49	80.6	30.5
1948	19	19.15	59.2	3.42	68.6	4.14	75.0	2.54	73.9	60.5
1949	20	18.28	63.5	5.51	72.4	3.47	76.2	2.34	73.0	46.1
1950	21	18.45	59.8	5.70	68.4	4.65	69.7	2.39	67.7	48.2
1951	22	22.00	62.2	6.11	65.2	4.45	72.1	6.21	70.5	43.1
1952	23	19.05	59.6	5.40	74.2	3.84	74.7	4.78	70.0	62.2
1953	24	15.67	60.0	5.31	73.2	3.28	74.6	2.33	73.2	52.9
1954	25	15.92	55.6	6.36	72.9	1.79	77.4	7.10	72.1	53.9
1955	26	16.75	63.6	3.07	67.2	3.29	79.8	1.79	77.2	48.4
1956	27	12.34	62.4	2.56	74.7	4.51	72.7	4.42	73.0	52.8
1957	28	15.82	59.0	4.84	68.9	3.54	77.9	3.76	72.9	62.1
1958	29	15.24	62.5	3.80	66.4	7.55	70.5	2.55	73.0	66.0
1959	30	21.72	62.8	4.11	71.5	2.29	72.3	4.92	76.3	64.2
1960	31	25.08	59.7	4.43	67.4	2.76	72.6	5.36	73.2	63.2
1961	32	17.79	57.4	3.36	69.4	5.51	72.6	3.04	72.4	75.4
1962	33	26.61	66.6	3.12	69.1	6.27	71.6	4.31	72.5	76.0

(a) Ajuste um Modelo Linear para prever a produção de milho (em *bu/acre*), utilizando a totalidade das restantes variáveis como variáveis preditoras. Comente os resultados.

(b) Determine o valor do R^2 modificado. Comente.

(c) Repita o ajustamento da primeira alínea, mas agora excluíndo a variável cronológica x_1 do conjunto de variáveis preditoras. Compare os resultados do ajustamento e o comportamento dos resíduos nos dois casos. Comente.

(d) Teste se o modelo com todas as variáveis preditoras e o modelo apenas com as variáveis preditoras que sejam conhecíveis até ao fim do mês de Junho diferem significativamente. Comente.

(e) Identifique um modelo mais parcimonioso, utilizando o método de exclusão sequencial de variáveis ($\alpha = 0.10$).

(f) No ajustamento do modelo escolhido na alínea anterior, mude as unidades de medida das variáveis como indicado de seguida e proceda a novo ajustamento do modelo. Comente eventuais

alterações nos resultados.

$$z^{\circ}\text{F} = \tfrac{5}{9}(z-32)^{\circ}\text{C}$$

Conversões:

$$1\ in = 25{,}4\ \text{mm}$$

$$1\ \text{bu/acre (milho)} = 0.06277\ \text{t ha}^{-1}$$

20. Num estudo duma espécie de árvores pretende-se estabelecer relações entre a altura dos troncos das árvores, o respectivo diâmetro à altura do peito e o volume desses troncos. Foram efectuadas medições destas variáveis em $n = 31$ árvores, sendo os resultados designados pelos nomes *Altura* (medida em pés), *Diametro* (medido em polegadas) e *Volume* (medido em pés cúbicos). Eis os valores de algumas estatísticas descritivas elementares, bem como dos coeficientes de correlação entre as variáveis:

```
> apply(arvores,2,summary)
        Diametro Altura Volume
Min.        8.30     63  10.20
1st Qu.    11.05     72  19.40
Median     12.90     76  24.20
Mean       13.25     76  30.17
3rd Qu.    15.25     80  37.30
Max.       20.60     87  77.00
```

```
> apply(arvores,2,var)
  Diametro     Altura     Volume
  9.847914  40.600000 270.202796

> cor(arvores)
          Diametro    Altura    Volume
Diametro 1.0000000 0.5192801 0.9671194
Altura   0.5192801 1.0000000 0.5982497
Volume   0.9671194 0.5982497 1.0000000
```

(a) Foi inicialmente ajustado um modelo de regressão linear múltipla para prever os volumes dos troncos, a partir das suas alturas e diâmetro, tendo sido obtidos os seguintes resultados.

```
Call: lm(formula = Volume ~ Diametro + Altura)
Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07
Diametro      4.7082     0.2643  17.816  < 2e-16
Altura        0.3393     0.1302   2.607   0.0145

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-Squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

i. Efectue o teste de ajustamento global do modelo. Discuta o resultado.

ii. Diga se é possível simplificar este modelo, obtendo uma regressão linear simples que não seja significativamente pior do que este modelo. Utilize os níveis de significância $\alpha = 0.05$ e $\alpha = 0.01$. Comente.

iii. Independentemente da sua resposta na alínea anterior indique, para cada um dos submodelos de regressão linear simples considerados, os Coeficientes de Determinação e o valor da estatística F no teste de ajustamento global.

(b) Tendo por base experiência anterior, foi sugerido que se poderia ainda melhorar o ajustamento procedendo a uma transformação logarítmica de todas as variáveis. O ajustamento resultante é indicado de seguida.

```
Call: lm(formula = log(Volume) ~ log(Diametro) + log(Altura))
Residuals:
      Min        1Q    Median        3Q       Max
```

```
-0.168561 -0.048488  0.002431  0.063637  0.129223

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.63162    0.79979  -8.292 5.06e-09 ***
log(Diametro)  1.98265    0.07501  26.432  < 2e-16 ***
log(Altura)    1.11712    0.20444   5.464 7.81e-06 ***

Residual standard error: 0.08139 on 28 degrees of freedom
Multiple R-Squared: 0.9777,     Adjusted R-squared: 0.9761
F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

i. Qual é a relação de base considerada por este modelo, em termos das variáveis originais (não logaritmizadas)?

ii. Discuta a seguinte afirmação: "o ajustamento dos dados logaritmizados é melhor, tendo em conta o maior Coeficiente de Determinação, o maior valor da estatística F e ainda os resíduos mais pequenos do que no caso dos dados não logaritmizados".

iii. Desconfiado de métodos estatísticos, um membro da equipa investigadora sugere que seria mais fácil estimar o volume dos troncos admitindo que estes eram cilíndricos. Nesse caso o volume seria dado por $v = \pi\, r^2\, h$, onde v, r e h indicam o volume, raio e altura do tronco, respectivamente *em unidades de medida comparáveis*. Teste se este modelo simples é admissível, à luz do ajustamento feito neste ponto e *tendo em conta as unidades das variáveis observadas*. **NOTA**: 1 pé corresponde a 12 polegadas e $\ln(\pi/24^2) = -5.211378$.

(c) Foi finalmente decidido experimentar um modelo (sem transformação das variáveis) em que as variáveis *Altura* e *Volume* trocam de papel em relação ao modelo inicial, ou seja, para saber se a altura dos troncos pode ser descrita, de forma adequada, a partir duma relação linear com o Diâmetro e o Volume. Foram obtidos os seguintes resultados com este modelo:

```
Call: lm(formula = Altura ~ Diametro + Volume)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.2958     9.0866   9.167 6.33e-10
Diametro     -1.8615     1.1567  -1.609   0.1188
Volume        0.5756     0.2208   2.607   0.0145

Residual standard error: 5.056 on 28 degrees of freedom
Multiple R-Squared: 0.4123,     Adjusted R-squared: 0.3703
F-statistic:  9.82 on 2 and 28 DF,  p-value: 0.0005868
```

Discuta o resultado deste teste, tendo em conta o valor relativamente baixo do Coeficiente de Determinação associado ao ajustamento. Como se pode explicar o facto de esta nova relação entre as mesmas três variáveis utilizadas no modelo da alínea inicial produzir uma muito pior qualidade do ajustamento?

21. Nas aulas foi visto que, dado o Modelo de Regressão Linear, se tem, para qualquer combinação linear $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$,

$$\frac{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} - \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}}{\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}}} \;\cap\; t_{n-(p+1)}\ ,$$

com $\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}} = \sqrt{QMRE \cdot \vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}$. A partir deste resultado, deduza a expressão para um intervalo a $(1-\alpha)\times 100\%$ de confiança para a combinação linear $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$.

22. Num estudo de maçãs Royal pretende-se relacionar o calibre das maçãs com o seu peso. Com base em 1273 frutos de calibre (em mm) entre 53 e 79, para os quais foi medido o peso (em g), ajustou-se um modelo de regressão linear, tendo-se obtido os resultados:

```
Call: lm(formula = Peso ~ Calibre, data = pesocal)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -210.3137     3.8078  -55.23   <2e-16
Calibre        5.1813     0.0577   89.79   <2e-16
---
Residual standard error: 8.525 on 1271 degrees of freedom
Multiple R-squared: 0.8638,Adjusted R-squared: 0.8637
F-statistic:  8063 on 1 and 1271 DF,  p-value: < 2.2e-16
```

(a) Qual seria a ordenada na origem natural para esta recta de regressão? Determine um intervalo a 95% de confiança para verificar se esse valor da ordenada na origem é admissível, face ao modelo ajustado. Comente as suas conclusões.

(b) Um investigador que analisou os resíduos do modelo ajustado alega que existe algum efeito de curvatura, e que seria preferível modelar o peso através de um polinómio de segundo grau no calibre. O resultado desse ajustamento foi o seguinte.

```
Call: lm(formula = Peso ~ Calibre + I(Calibre^2), data = pesocal)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  72.33140   46.76415   1.547   0.1222
Calibre      -3.38747    1.41429  -2.395   0.0168
I(Calibre^2)  0.06469    0.01067   6.064 1.75e-09
---
Residual standard error: 8.408 on 1270 degrees of freedom
Multiple R-squared: 0.8677,Adjusted R-squared: 0.8675
F-statistic:  4163 on 2 and 1270 DF,  p-value: < 2.2e-16
```

i. Indique a equação da parábola que descreve a relação ajustada.

ii. Considera que o investigador tem razão? Justifique através duma análise estatística adequada. Comente os seus resultados, tendo em atenção os valores dos R^2 de cada modelo.

23. Considere o vector $\vec{\mathbf{1}}_n \in \mathbb{R}^n$, constituido por n uns. Considere um outro qualquer vector $\vec{\mathbf{x}} = (x_1, x_2, ..., x_n)^t$ de $\mathbb{R}^n$, que consideramos um vector de n observações numa variável X.

(a) Construa a matriz $\mathbf{P} = \vec{\mathbf{1}}_n(\vec{\mathbf{1}}_n^t\vec{\mathbf{1}}_n)^{-1}\vec{\mathbf{1}}_n^t$ de projecção ortogonal sobre o subespaço $\mathcal{C}(\mathbf{1}_n) \subset \mathbb{R}^n$ gerado pelo vector $\vec{\mathbf{1}}_n$ (i.e., $\mathcal{C}(\vec{\mathbf{1}}_n)$ é o conjunto de vectores que são múltiplos escalares de $\vec{\mathbf{1}}_n$).

(b) Identifique os elementos do vector $\mathbf{P}\vec{\mathbf{x}}$ que é a projecção ortogonal do vector $\vec{\mathbf{x}}$ sobre o subespaço $\mathcal{C}(\vec{\mathbf{1}}_n)$, e comente.

(c) Mostre que a variável *centrada* $\mathbf{x}^c$, cujo elemento genérico é $x_i - \overline{x}$, se pode escrever como $\vec{\mathbf{x}} - \mathbf{P}\vec{\mathbf{x}} = (\mathbf{I} - \mathbf{P})\vec{\mathbf{x}}$, onde $\mathbf{I}$ indica a matriz identidade $n \times n$.

(d) Mostre que o *desvio padrão* das n observações da variável X é proporcional à norma (comprimento) do vector $\mathbf{x}^c$, definido na alínea anterior.

(e) Represente graficamente a situação descrita nas alíneas anteriores. Mostre que se definiu um triângulo rectângulo em $\mathbb{R}^n$. Aplique-lhe o Teorema de Pitágoras e comente.

24. Numa regressão linear tem-se:

$$\begin{aligned} SQT &= \|\mathbf{Y} - \mathbf{P}_{\vec{\mathbf{1}}_n}\mathbf{Y}\|^2 \\ SQR &= \|\mathbf{HY} - \mathbf{P}_{\vec{\mathbf{1}}_n}\mathbf{Y}\|^2 \\ SQRE &= \|\mathbf{Y} - \mathbf{HY}\|^2 \end{aligned}$$

onde $\mathbf{Y}$ indica o vector de observações da variável resposta, $\mathbf{H}$ é a matriz de projecção ortogonal sobre o subespaço $\mathcal{C}(\mathbf{X})$ gerado pelas colunas da matriz $\mathbf{X}$ e $\mathbf{P}_{\vec{\mathbf{1}}_n}$ é a matriz de projecção ortogonal sobre o subespaço $\mathcal{C}(\vec{\mathbf{1}}_n)$ gerado pelo vector dos n uns, $\vec{\mathbf{1}}_n$. Mostre, algebricamente, que $SQT = SQR + SQRE$.

25. Considere o modelo de regressão linear *sem preditores*, ou seja, o modelo nulo:

$$\begin{array}{rcl} Y_i & = & \beta_0 + \epsilon_i \ , \quad \forall\, i = 1, ..., n \\ \epsilon_i & \cap & \mathcal{N}(0, \sigma^2) \ , \quad \forall i \\ \{\epsilon_i\}_{i=1}^n & \text{v.a.} & \text{independentes} \end{array}$$

Usando a notação matricial na formulação do modelo, a matrix $\mathbf{X}$ terá uma única coluna, composta por uns, ou seja, $\mathbf{X} = \vec{\mathbf{1}}_n$. Tendo também em atenção o Exercício 23,

(a) Determine o estimador de mínimos quadrados de β_0.

(b) Determine a média e a variância desse estimador de β_0.

(c) Determine a distribuição de probabilidades do estimador de β_0.

(d) Determine as expressões para SQR e $SQRE$ neste modelo. Comente.

(e) Relacione as suas conclusões com a matéria das disciplinas introdutórias de Estatística, relativamente à estimação duma média populacional com base numa amostra aleatória.

(f) Utilize os resultados da alínea 25d) para mostrar que a estatística do teste F parcial, comparando o submodelo sem preditores com um modelo completo com p preditores, é igual à estatística do teste F de ajustamento global do modelo completo.

26. Considere o modelo com equação base sem constante aditiva,

$$Y_i \ = \ \beta_1\, x_i \ + \ \epsilon_i \qquad (i = 1, ..., n).$$

(a) Determine o estimador de mínimos quadrados para o parâmetro β_1.

(b) Determine a distribuição de probabilidades do estimador obtido na alínea anterior, admitindo válidas as restantes hipóteses do Modelo Linear.

27. Considere um modelo de regressão linear múltipla com p variáveis preditoras, ajustado com base em n observações.

(a) Descreva pormenorizadamente o modelo, *usando a notação vectorial/matricial.*

(b) Mostre que o vector de estimadores dos parâmetros do modelo, $\vec{\hat{\boldsymbol{\beta}}}$, também se pode escrever como $\vec{\hat{\boldsymbol{\beta}}} = \vec{\boldsymbol{\beta}} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\boldsymbol{\epsilon}}$.

(c) Deduza *a partir da expressão da alínea anterior*, o vector esperado e a matriz de covariâncias do vector dos estimadores, $\vec{\hat{\boldsymbol{\beta}}}$, ao abrigo do modelo de regressão linear múltipla.

28. Considere os coeficientes de determinação usual (R^2) e modificado (R^2_{mod}), no contexto duma regressão linear múltipla com p variáveis preditoras , ajustada com base em n observações.

(a) Mostre que se verifica a relação $R^2_{mod} = 1 - (1 - R^2)\frac{n-1}{n-(p+1)}$.

(b) Mostre que a estatística do teste F de ajustamento global do modelo se pode escrever apenas à custa de R^2 e R^2_{mod}, verificando-se $F_{calc} = \frac{R^2}{R^2 - R^2_{mod}}$.

(c) Mostre que o coeficiente de determinação modificado é negativo quando $R^2 < \frac{p}{n-1}$. Comente as implicações desta condição para a estatística do teste F de ajustamento global.

2 Análise de Variância

AVISO: Os conjuntos de dados necessários nesta secção são `tomate` (Exercício 1), `CO2` (Exercício 3), `terrenos` (Exercício 4) e `pinheiro` (Exercício 6).

1. No melhoramento de variedades tradicionais de tomate, uma característica importante é a resistência da película. Esta característica foi avaliada em 6 variedades de tomate. De cada variedade foram colhidos aleatoriamente tomates em cada uma de 3 parcelas de tomateiros, sendo cada observação constituída pela resistência média dos frutos de uma dada parcela (medida num texturómetro, em grama força, *gf*). Eis os valores obtidos em cada parcela (que se encontram na *data frame* `tomate`), bem como as médias e variâncias obtidos, para cada variedade, e para a totalidade das observações:

Variedade	Observações			Média	Variância
18	632.04	629.30	420.59	560.6433	14 713.08
28	253.00	219.34	252.11	241.4833	367.9434
29	223.71	374.48	274.66	290.9500	5881.921
40C	503.51	757.44	856.39	705.7800	33 132.64
Ace	375.18	376.81	379.77	377.2533	5.414433
Roma	333.05	324.82	338.45	332.1067	47.11163

- A média global das observações é $\overline{y}_{..} = 418.0361$;
- a variância amostral da totalidade das observações é $s_y^2 = 34\,517.82$.

   (a) Indique qual o tipo de delineamento experimental em causa. Explicite o modelo correspondente e todas as hipóteses adicionais que sejam necessárias à consideração do problema em estudo.

   (b) Construa a tabela-resumo da análise de variância correspondente a este caso,

      i. utilizando apenas uma máquina de calcular e a informação disponível neste enunciado;

      ii. utilizando, no `R`, o comando `summary(aov(res.pel ~ variedade , data=tomate))`.

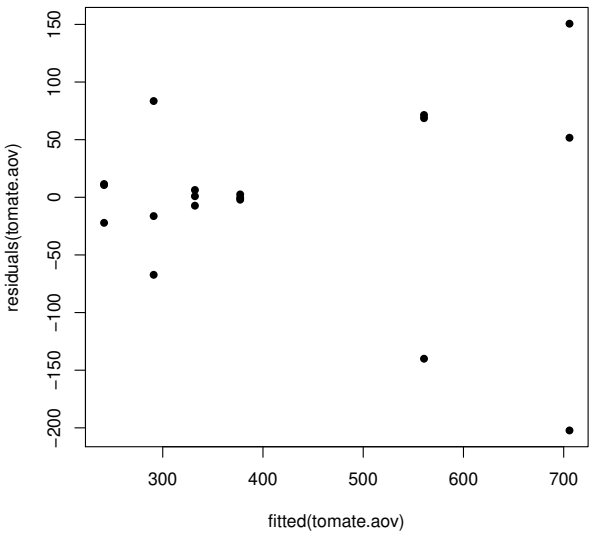
   (c) Formalize e efectue o teste F adequado ao problema acima referido, com um nível de significância de 5%. Pode afirmar-se que a resistência média da película não é sempre igual, em todas as variedades?

   (d) Qual o maior nível de significância α para o qual mudaria a sua resposta na alínea anterior? Como se designa esse valor?

   (e) Utilize o comando `model.matrix` do `R` para inspeccionar a natureza da matriz do modelo, $\mathbf{X}$, neste contexto.

   (f) Utilize o comando `fitted` do `R` para identificar os valores ajustados da variável resposta, nesta Análise de Variância.

   (g) O gráfico dos resíduos (usuais) das observações, contra os valores ajustados pelo modelo de análise de variância, é apresentado a seguir. Comente o gráfico e as suas possíveis implicações. Identifique a observação cujo resíduo é, em módulo, mais elevado.



2. Um estudo sobre três variedades de café, referenciadas por `CA`, `CL` e `PR` focou-se sobre os comprimentos dos estomas das respectivas folhas. De cada variedade foram seleccionadas 12 plantas, e para cada planta foi medido o comprimento médio dos estomas das suas folhas em condições ambientais controladas (variável `Comprimento`, em μm). São conhecidas apenas as médias e variâncias das 12 observações (plantas) de cada variedade:

	CA	CL	PR
Média	22.85833	19.49333	25.31583
Variância	13.69303	2.725424	9.388936

   (a) Explicite pormenorizadamente o modelo ANOVA adequado ao estudo do problema.

   (b) Construa a tabela-resumo da ANOVA que indicou na alínea anterior.

   (c) Qual é a variância amostral dos comprimentos dos estomas na totalidade das 36 observações?

   (d) É possível afirmar que, na população, o comprimento médio dos estomas é igual nas três variedades, para um nível de significância $\alpha = 0.05$? Responda pormenorizadamente.

   (e) **[Material Complementar]** Compare todos os pares de médias através dum teste de Tukey ($\alpha = 0.05$). Comente.

3. Sabe-se que o dióxido de carbono tem um efeito crítico no crescimento de populações microbianas; pequenas quantidades de CO_2 podem estimular o crescimento de algumas espécies enquanto que, pelo contrário, grandes concentrações têm de forma geral uma acção inibitória. Este último efeito é usado comercialmente para preservar alimentos armazenados.

   Realizou-se um estudo para investigar a acção de diferentes concentrações de CO_2 na taxa de crescimento de *Pseudomonas fragi*; os diferentes níveis (tratamentos) foram pré-fixados e a variável resposta medida foi a percentagem de variação na massa das culturas após uma hora de crescimento nas respectivas condições, originando os dados da seguinte tabela.

Concentração de CO_2				
0.0	.083	.29	.50	.86
62.6	50.9	45.5	29.5	24.9
59.6	44.3	41.1	22.8	17.2
64.5	47.5	29.8	19.2	7.8
59.3	49.5	38.3	20.6	10.5
58.6	48.5	40.2	29.2	17.8
64.6	50.4	38.5	24.1	22.1
50.9	35.2	30.2	22.6	22.6
56.2	49.9	27.0	32.7	16.8
52.3	42.6	40.0	24.4	15.9
62.8	41.6	33.9	29.6	8.8

Estes dados estão disponíveis na *data frame* `CO2`, sendo as concentrações de CO_2 repetidas em duas colunas: numa sob a forma de factor e noutra sob a forma de variável numérica.

(a) Pretende-se testar a hipótese nula $H_0 : \mu_1 = \mu_2 = ... = \mu_5$, onde μ_i indica a taxa de crescimento esperada para a i-ésima concentração de CO_2. É sugerida a utilização de uma Análise de Variância. Enuncie os pressupostos necessários para poder efectuar o teste referido.

(b) Haverá evidência suficiente para rejeitar H_0 com uma significância de $\alpha = .05$?

(c) Estude a validade dos pressupostos do modelo ANOVA.

(d) Dada a natureza da variável preditora, também poderia ser considerada uma regressão linear das taxas de crescimento sobre as concentrações de dióxido de carbono, encaradas como uma variável numérica. Utilizando a coluna de `CO2` com as concentrações dadas como variáveis numéricas (isto é, a coluna `CO2.numerico`), responda às seguintes questões.

i. Construa a nuvem de pontos da variação de massa sobre concentração de CO2.

ii. Ajuste a regressão linear simples referida, traçando a recta de regressão sobre a nuvem de pontos. Comente.

iii. Compare os resultados do teste F de ajustamento global obtidos usando os comandos `lm` e `aov`. Comente.

4. Pretende-se comparar o rendimento obtido com quatro variedades de trigo. Identificaram-se 13 terrenos com características de solos diferentes, que correspondem aos tipos de terrenos nos quais se pretende fazer as culturas. Os 13 terrenos são então divididos em quatro parcelas de igual dimensão. Em cada terreno associa-se, de forma aleatória, uma parcela a cada uma das quatro variedades. Após a colheita registam-se os rendimentos obtidos (em t/ha) na tabela (e disponíveis na *data.frame* `terrenos`).

(a) As médias amostrais de cada variedade sugerem que há variedades com desempenho superior. Mas serão essas diferenças significativas? A fim de responder, efectue uma Análise da Variância adequada, construindo a tabela-resumo correspondente. Comente as suas conclusões.

(b) Teste se, entre terrenos, existem diferenças significativas, como seria de supôr. Comente.

Terreno	Variedade A	B	C	D
I	1.800	2.457	0.722	0.789
II	1.709	1.839	1.546	1.304
III	1.277	1.293	1.515	1.273
IV	1.675	1.745	0.800	0.846
V	1.814	1.833	1.678	1.732
VI	1.896	1.203	1.192	1.580
VII	1.078	1.689	1.583	1.168
VIII	1.740	1.518	1.050	1.305
IX	1.200	1.133	0.778	1.033
X	1.500	0.722	0.636	0.925
XI	1.932	1.700	1.203	0.850
XII	1.169	1.209	1.112	0.986
XIII	1.438	1.577	1.355	1.525
Médias	1.556	1.532	1.167	1.178
Variâncias	0.0879	0.1855	0.1266	0.0934

5. Uma experiência visa estudar o rendimento duma variedade de trigo em função de diferentes formas de aplicar dois adubos, um com fósforo (um adubo fosfatado), e outro com potássio. Consideram-se três dosagens de aplicação do adubo fosfatado, designadas por Baixa, Média e Elevada. Igualmente, consideram-se três dosagens de aplicação do fertilizante com potássio, igualmente designadas por Baixa, Média e Elevada. A experiência realiza-se num terreno com 27 parcelas de igual dimensão. Repartem-se, de forma totalmente casualizada, três parcelas por cada combinação de dosagem de um e outro fertilizante. Os resultados obtidos (em t/ha) foram os seguintes:

		Potássio (K)									Média	Variância
		Baixa			Média			Elevada				
Fósforo (P)	Baixa	4.6	4.9	4.3	6.3	6.1	6.4	6.6	6.7	6.9	5.8667	0.9775
	Média	5.4	5.6	5.2	6.8	5.7	6.7	7.5	8.0	7.3	6.4667	1.045
	Elevada	5.3	5.7	5.1	7.5	7.0	7.2	7.1	7.4	6.1	6.4889	0.88861
Média		5.1222			6.6333			7.0667				
Variância		0.20944			0.3200			0.3175				

As médias observadas para cada combinação de dosagens de cada tipo de fertilizante foram as seguintes:

		Potássio		
		Baixa	Média	Elevada
Fósforo	Baixa	4.600	6.267	6.733
	Média	5.400	6.400	7.600
	Elevada	5.367	7.233	6.867

A Tabela-Resumo associada a esta experiência é a seguinte:

```
                Df Sum of Sq  Mean Sq  F Value      Pr(F)
fosforo          ?   2.24296 1.121481    ?     0.00366530
potassio         2  18.75630    ?        ?     0.00000001
fosforo:potassio ?   1.93926 0.484815  3.36504 0.03187154
Residuals       18   2.59333    ?
```

Responda às seguintes questões, utilizando a informação disponível no enunciado.

(a) Complete a Tabela-Resumo, indicando como obtem cada um dos valores omissos.

(b) Que *tipo* de efeitos do modelo associado a este caso devem ser considerados significativos? Justifique, explicitando as hipóteses dos testes que efectuou, as estatísticas dos testes e os níveis de significância utilizados, bem como a natureza das regiões críticas, os valores obtidos e as conclusões.

(c) Ajuste agora um modelo a dois factores, mas que não preveja os efeitos de interacção. Construa a tabela-resumo correspondente. Identifique as diferenças entre esta tabela e a que se indicou acima (associada ao modelo que prevê efeitos de interacção). Comente as diferenças e identifique as consequências de não prever a existência de efeitos de interacção quando na realidade esses efeitos parecem existir.

6. Num estudo sobre características de crescimento de pinheiro manso, conduzido em Sines e em Tavira pelo Instituto Nacional de Investigação Agrária e Veterinária (INIAV), avaliou-se a altura média de pinheiros de cinco diferentes proveniências (Marrocos, Grécia, Portugal e duas proveniências de Itália), dois anos após a plantação. Quer em Sines, quer em Tavira, foram plantados seis talhões com árvores de cada proveniência, gerando assim $n=60$ valores de alturas (variável `alt2`, em cm), cuja variância amostral é $s^2=34.49584$. Eis algumas médias resultantes.

```
 prov                                                    local
Grecia Italia-1 Italia-2 Marrocos Portugal              Sines Tavira
 28.81    32.75    30.23    35.13    31.90              28.14  35.38

 prov:local
local                                                   Grand mean
 prov        Sines  Tavira                               31.76298
   Grecia    22.52   35.10
   Italia-1  31.03   34.46
   Italia-2  26.91   33.56
   Marrocos  31.16   39.09
   Portugal  29.09   34.70
```

(a) Identifique o delineamento experimental utilizado e o modelo ANOVA adequado. Descreva pormenorizadamente o modelo.

(b) Sabendo que o Quadrado Médio Residual é 16.59 e que a Soma de Quadrados associada às cinco diferentes proveniências é 280.61, construa a tabela-resumo do modelo ANOVA adequado.

(c) Use um teste F para avaliar a existência de efeitos de proveniência dos pinheiros. Comente as suas conclusões. Indique brevemente que outros tipos de efeitos devem ser considerados significativos. Considere $\alpha = 0.05$.

(d) **[Material Complementar]** Na amostra, a maior altura média em Sines é inferior à menor altura média em Tavira. Independentemente das suas respostas nas alíneas anteriores, use o teste de Tukey para indicar se igual afirmação se pode estender à população. Comente.

7. Uma engenheira agrónoma pretende seleccionar cultivares de trigo para as quatro explorações agrícolas pelas quais é responsável, que se localizam em Elvas, Évora, Benavila e Revilheira. Em cada uma destas explorações, definem-se 36 parcelas de terra, associando aleatoriamente quatro parcelas a cada uma de nove cultivares: Celta, Helvio, TE9006, TE9007, TE9008, TE9110, TE9115, TE9204 e Trovador. Em cada parcela foi medido o rendimento, em kg/ha. A variância da totalidade dos rendimentos observados é $s^2 = 1\,714\,242$.

(a) Especifique o delineamento experimental utilizado, e descreva em pormenor o modelo ANOVA adequado a esta experiência.

(b) Foi ajustado um modelo ANOVA, com o programa R. Resultados parciais desse ajustamento são dados de seguida.

```
                    Df    Sum Sq  Mean Sq  F value
localidade         ??? 183759916      ??? 234.9531
cultivar           ???       ???   964060      ???
localidade:cultivar ???      ???      ???   4.0768
Residuals          ???  28156076   260704
```

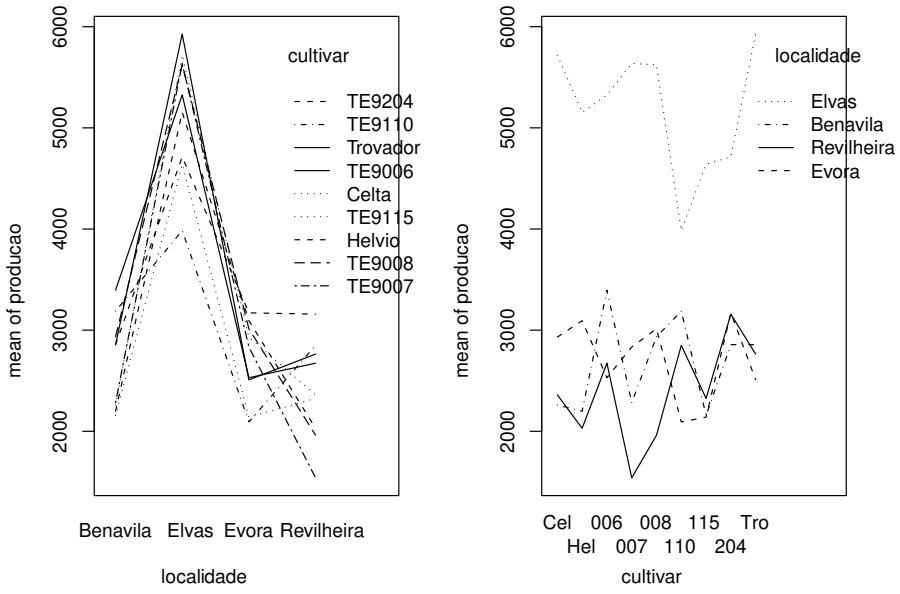
i. Complete a tabela, indicando como obtém cada um dos valores omissos.

ii. Qual o valor estimado da variância dos erros aleatórios do modelo, e quais as suas unidades de medida?

iii. Teste formalmente (a um nível de significância $\alpha = 0.01$) quais os tipos de efeitos do modelo que devem ser considerados significativos. Descreva um teste em pormenor e discuta os restantes de forma sintética.

iv. Discuta o efeito de mudar as unidades de medida da variável resposta de kg/ha para toneladas por hectare. Quais os valores da tabela que se alteram, e quais os que ficam iguais? Quais os efeitos da mudança de unidades nas conclusões dos testes F?

v. **[Material Complementar]** Os gráficos de interacção associados a esta experiência são os seguintes. Comente-os, relacionando as suas conclusões das alíneas anteriores com os gráficos.



8. Com o objectivo de analisar as alterações no conteúdo em taninos da polpa de sapotis (frutos do sapotizeiro, *Manilkara achras*) provocadas pela temperatura de conservação (alta/baixa) e pelo tempo de armazenamento (0, 3, 6 ou 9 dias) foi efectuado um estudo que forneceu os seguintes dados:

	Tempo							
Temperatura	0 dias		3 dias		6 dias		9 dias	
alta	20.8	19.7	26.5	27.5	26.5	26.4	26.5	26.9
	18.0	19.5	27.0	26.4	27.0	24.0	25.9	26.3
baixa	32.3	34.1	20.8	20.5	16.4	15.7	10.3	9.7
	30.7	31.8	21.0	20.9	15.9	16.0	7.8	9.8

A média e a variância do conjunto das 32 observações são 22.14375 e 47.83222, respectivamente. As médias associadas a cada tempo de armazenamento, cada temperatura e cada combinação de tempo e temperatura, são:

```
Tables of means
 tempo                          tempo:temperatura
     0      3      6      9            temperatura
25.862 23.825 20.987 17.900      tempo  alta baixa
                                    0 19.50 32.23
 temperatura                        3 26.85 20.80
  alta  baixa                       6 25.97 16.00
24.681 19.606                       9 26.40  9.40
```

(a) Identifique o delineamento experimental utilizado no estudo e descreva de forma pormenorizada o melhor modelo ANOVA que lhe está associado.

(b) Sabendo que a Soma dos Quadrados dos Resíduos é 20.72 e que o Quadrado Médio associado aos diferentes tempos de armazenamento é 96.01, construa o Quadro-Resumo da Análise de Variância associado a esta experiência.

(c) Pode considerar-se que os diferentes tempos de armazenamento influenciam o teor de taninos na polpa destes frutos? Responda a esta questão utilizando testes de hipóteses.

9. O interesse em introduzir em Palmela castas exteriores à região, conduziu à organização (numa colaboração ISA/PORVID) dum ensaio com duas castas: Malvasia Fina e Antão Vaz. Decidiu-se trabalhar com quatro genótipos de cada casta (designados MF201, MF1035, MF1420 e MF1426 no caso da Malvasia Fina e AN105, AN142, AN145 e AN510 para a Antão Vaz), sendo o objectivo escolher as combinações de casta e genótipo associadas aos maiores rendimentos na região. Foi utilizado um delineamento equilibrado com oito repetições em cada situação experimental. O rendimento médio da totalidade das observações foi $4.467625\,kg/planta$, e a respectiva variância amostral $5.389415\,(kg/planta)^2$.

   (a) Descreva em pormenor o modelo ANOVA mais adequado à experiência agora descrita.

   (b) Construa a tabela-resumo correspondente a este modelo, sabendo que o Quadrado Médio Residual é 2.873782 e que a Soma de Quadrados associada às castas é 79.73597.

   (c) Foi importante prever a possibilidade de os rendimentos serem diferentes por efeito dos genótipos? Justifique a resposta através dum teste formal.

   (d) Para além de eventuais efeitos de genótipos, pode-se falar em diferenças entre as duas castas? Justifique formalmente a sua resposta.

   (e) **[Material Complementar]** Sabendo que os rendimentos médios nas oito situações experimentais são os abaixo indicados, teste se é possível concluir que o rendimento médio do genótipo MF201 é diferente de todos os outros.

AN105	AN142	AN145	AN510	MF1035	MF1420	MF1426	MF201
2.925	2.208	3.593	4.680	3.451	5.367	5.839	7.678

10. Mostre que é nula a soma dos resíduos das observações em:

   (a) cada nível do Factor, numa ANOVA a 1 Factor;

   (b) cada célula, numa ANOVA a 2 Factores, com interacção.

11. Considere o modelo ANOVA para um delineamento a um factor, equilibrado. Tratando-se dum modelo linear, é possível calcular o respectivo coeficiente de determinação R^2, embora não seja usual fazê-lo em modelos ANOVA.

(a) Indique condições equivalentes ao valor extremo $R^2 = 0$, envolvendo as médias amostrais de nível e da totalidade das observações. Interprete essa situação em termos do teste F da ANOVA.

(b) Indique condições equivalentes ao valor extremo $R^2 = 1$, envolvendo as variâncias amostrais de nível. Interprete essa situação em termos do teste F da ANOVA.

3 Análise de Covariância

1. Considere as medições sobre folhas de videira introduzidas no Exercício 18 da Regressão Linear (*data frame* **videiras**).

   (a) Desenhe a nuvem de pontos do comprimento da nervura principal (variável *NP*), no eixo horizontal, e nervura lateral direita (variável *NLdir*) no eixo vertical, usando cores diferentes para representar as folhas de cada casta (variável *Casta*). Comente.

   (b) Ajuste uma única recta de regressão linear para prever os comprimentos das nervuras laterais direitas, a partir dos comprimentos das nervuras principais, utilizando a totalidade das $n = 600$ folhas observadas, e ignorando as Castas de origem. Trace essa recta sobre o gráfico criado na alínea anterior. Comente a qualidade desta regressão linear simples.

   (c) Ajuste um modelo de Análise de Covariância à totalidade das $n = 600$ observações, que possibilite que as folhas de cada Casta tenham uma recta de regressão linear diferente. Trace as três rectas resultantes, utilizando as cores correspondentes aos pontos da respectiva casta. Comente o resultado.

   (d) Teste formalmente se o modelo que utilizou na alínea anterior e o modelo da recta única ajustado na alínea 1b) diferem significativamente. Comente as conclusões do seu teste.

   (e) Ajuste um modelo de regressão linear simples de *NLdir* sobre *NP*, para cada um dos seguintes subconjuntos de $n_i = 200$ $(i = 1, 2, 3)$ observações:

      i. as n_1 observações da Casta Água Santa;
      ii. as n_2 observações da Casta Fernão Pires;
      iii. as n_3 observações da Casta Vital

      Comente os seus resultados. Em particular, compare os Coeficientes de Determinação de cada um destes modelos ajustados com o Coeficiente de Determinação do modelo de ANCOVA ajustado na alínea 1c).

   (f) Inspeccione a matriz $\mathbf{X}$ usada pelo programa R aquando do ajustamento de cada um dos modelos usados neste Exercício (e que é disponibilizada através da função `model.matrix`, aplicada ao objecto `lm` da regressão considerada).

2. Considere as medições sobre folhas de videira introduzidas no Exercício 18 da Regressão Linear (*data frame* **videiras**).

   (a) Desenhe a nuvem de pontos do comprimento da nervura principal (variável *NP*), no eixo horizontal, e área foliar (variável *Area*) no eixo vertical, usando cores ou símbolos diferentes para representar as folhas de cada casta (variável *Casta*). Comente.

   (b) Repita a alínea anterior, mas utilizando os logaritmos das variáveis *NP* e *Area*. Comente.

   (c) Ajuste uma única recta de regressão para modelar os logaritmos das áreas foliares com base nos logaritmos dos comprimentos das nervuras principais, independentemente das castas. Comente a qualidade do ajustamento obtido.

   (d) Ajuste um novo modelo para o logaritmo das áreas foliares, mas cruzando a relação linear sobre *log-NP* com o factor *Casta*. Comente a qualidade do novo ajustamento.

(e) Discuta o significado do modelo com ajustamento por Casta, obtido na alínea anterior, *em termos das variáveis não logaritmizadas*.

(f) Teste formalmente se a distinção de modelos linearizados por Casta é significativamente melhor.

(g) Independentemente da sua resposta na alínea anterior, desenhe as seguintes rectas, na nuvem de pontos obtida na alínea 2b:

i. a recta obtida ignorando as castas de cada folha;

ii. as três rectas obtidas para cada casta (utilize cores diferentes na sua representação).

(h) Na nuvem de pontos entre as variáveis (não logaritmizadas) que obteve na alínea 2a, trace as seguintes curvas (tendo em conta o resultado das regressões lineares que ajustou):

i. a curva associada à relação entre área foliar e comprimento da nervura principal, independentemente da casta de origem de cada folha.

ii. as três curvas associadas às relações não lineares entre área foliar e comprimento da nervura principal, para cada casta.

Compare os resultados desta alínea e da anterior, e comente.

3. Considere os dados relativos a 150 lírios (*data frame* **iris**).

(a) Construa a nuvem de pontos das medições de largura das sépalas (eixo horizontal) e largura das pétalas (eixo vertical), mas identificando a espécie a que corresponde cada observação. Comente o resultado.

(b) Independentemente do resultado da alínea anterior, ajuste uma regressão linear simples de largura das pétalas sobre largura das sépalas, para a totalidade das $n = 150$ observações. Comente os resultados obtidos.

(c) Ajuste agora um modelo ANCOVA para largura de pétalas, que cruze a regressão linear simples sobre a largura das sépalas com o factor *Species*. Em particular,

i. Desenhe as rectas de regressão linear obtidas para cada espécie, em cima da nuvem de pontos da alínea 3a).

ii. Compare o valor do coeficiente de determinação obtido agora, com o valor de R^2 obtido quando se ajustava uma única recta de regressão, independentemente das espécies. Comente.

iii. A informação disponível sugere que as rectas de regressão para as espécies *versicolor* e *virginica* são paralelas. Teste formalmente esta hipótese.

(d) Ajuste agora as 3 rectas de regressão de largura das pétalas sobre largura das sépalas, para cada espécie em separado. Compare os coeficientes de determinação obtidos com cada espécie com o coeficiente de determinação obtido ajustando o modelo ANCOVA da alínea 3c). Qual a razão para a discrepância nos valores de R^2 no modelo ANCOVA e nos modelos separados?

(e) Calcule as Somas de Quadrados para cada um dos modelos referidos na alínea anterior e confirme as fórmulas dadas nas aulas teóricas relacionando cada tipo de Somas de Quadrados e os coeficientes de determinação.

4. Repita o Exercício 3, mas utilizando agora a variável comprimento das pétalas como preditor da largura das pétalas. Comente, em particular, o valor do coeficiente de determinação do modelo único de regressão linear simples, associado aos $n = 150$ lírios. Tendo em conta o baixo valor dos R_i^2 $(i = 1, 2, 3)$ para os modelos separados de cada espécie, como se pode explicar este elevado valor do R^2 da regressão linear simples da totalidade das 150 observações? Comente as implicações duma situação deste tipo.