Exercícios - Modelos Matemáticos
Modelo Linear - 2015-16

1 Regressão Linear

AVISO: Os conjuntos de dados de alguns exercícios desta secção encontram-se disponíveis na página *web* da disciplina. 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 (de preferência uma pasta chamada `AulasMMA` numa *pen*).
- Executar, a partir duma sessão do `R` nessa directoria, o comando `load("exerRL.RData")`.

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* $\longrightarrow$ *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. O programa `R` tem vários conjuntos de dados disponíveis. Um desses conjuntos de dados designa-se `anscombe` e pode ser visto apenas escrevendo o nome do objecto. Utilizando estes dados, determine, e comente os valores obtidos para:

   (a) As médias de cada variável x_i e y_i $(i = 1 : 4)$.

   (b) As variâncias de cada variável x_i e y_i $(i = 1 : 4)$.

   (c) O valor dos parâmetros b_0 e b_1 nas quatro rectas de regressão de y_i sobre x_i $(i = 1, 2, 3, 4)$.

   (d) Os Coeficientes de Determinação associados às quatro rectas indicadas na alínea anterior.

   Após comentar os resultados obtidos, construa as quatro nuvens de pontos $\{(x_i^{(j)}, y_i^{(j)})\}_{i=1}^{11}$, para $j = 1 : 4$. Comente esses gráficos, à luz dos valores anteriormente obtidos.

4. 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 uma 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 nuvens de pontos com as seguintes transformações de uma ou ambas as variáveis:

      i. $\ln(y)$ vs. x;
      ii. y vs. $\ln(x)$;
      iii. $\ln(y)$ vs. $\ln(x)$.

   (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) 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?

(e) 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.

(f) 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)?

(g) 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.

(h) Utilize o comando `ltsreg`, descrito nos acetatos das aulas (disponível no referido módulo `MASS`), para ajustar a recta de regressão robusta através do método LTS. Utilize diferentes opções quanto ao número de observações a considerar na soma aparada (argumento `quant` do comando `ltsreg`). Comente os resultados, tendo em atenção a recta de regressão usual.

(i) Utilize o comando `lmsreg`, descrito nos acetatos das aulas (e igualmente disponível no referido módulo `MASS`), para ajustar a recta de regressão robusta através do método LMS. Comente o resultado, tendo em atenção a recta de regressão usual e também os resultados da regressão LTS.

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

(j) 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).

(k) Compare a recta obtida na alínea 4j) 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 4e.

(l) 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)?

5. 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 25ºC.

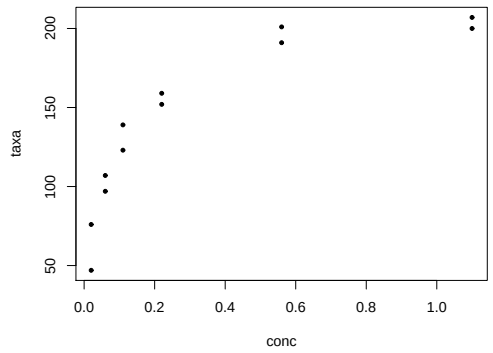
(c) Considere novamente a nuvem de pontos original. Trace a curva exponencial correspondente ao ajustamento efectuado na alínea anterior.

6. 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.

(d) Repita as alíneas anteriores, mas utilizando agora uma regressão robusta LMS. Compare os resultados obtidos e comente.

7. 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) Efectue um teste de ajustamento global do modelo de regressão linear simples do teor de flavonóides (`V8`) sobre o teor alcoólico (`V2`). Comente o resultado tendo em conta o valor do coeficiente de determinação e a nuvem de pontos das observações para essas duas variáveis. Determine o valor das três Somas de Quadrados associadas a esta regressão.

(c) 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.

(d) 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)
```

(e) 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.

8. 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 8a). 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.

9. 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.

10. 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) $\mathbf{X}^t\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}$.

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

(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

$$S_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y}) = \sum_i x_i y_i - n\overline{x}\,\overline{y}\ ; \quad \text{e}$$
$$s_x^2 = \sum_i (x_i - \overline{x})^2 = \sum_i x_i^2 - n\overline{x}^2\ .$$

11. (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.

12. 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 $y \to 10 \times y$).

13. 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 $E[\vec{\mathbf{Z}} + \vec{\mathbf{U}}] = E[\vec{\mathbf{Z}}] + E[\vec{\mathbf{U}}]$.

14. 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.

15. Considere os dados do Exercício 4 (`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 4e). 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?

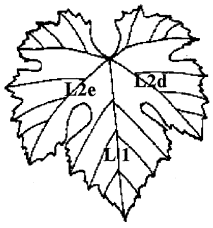
    (f) Preveja o valor esperado do log-peso do cérebro para espécies com peso de corpo igual a $250kg$. Construa um intervalo de confiança para esse valor esperado.

    (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.

16. 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.

17. 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°F	=	$\frac{5}{9}(z-32)$°C
Conversões:	1 *in*	=	25,4 mm
	1 bu/acre (milho)	=	0.06277 t ha^{-1}

18. 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?

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

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

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

20. 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.

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

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

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

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

22. Numa regressão linear tem-se:

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

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 $\mathbf{P}_{\mathbf{1}_n}$ é a matriz de projecção ortogonal sobre o subespaço $\mathcal{C}(\mathbf{1}_n)$ gerado pelo vector dos n uns, $\mathbf{1}_n$. Mostre, algebricamente, que $SQT = SQR + SQRE$.

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

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

Usando a notação matricial na formulação do modelo, a matrix $\mathbf{X}$ terá uma única coluna, composta por uns, ou seja, $\mathbf{X} = \mathbf{1}_n$. Tendo também em atenção o Exercício 21,

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

(b) Determine 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 23d) para mostrar que a estatística do teste F parcial, comparando o submodelo sem preditores com o modelo completo com p preditores, é igual à estatística do teste F de ajustamento global do modelo completo.

24. 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.