

## Exercícios - Estatística e Delineamento - 2021-22

### 3 Regressão Linear Múltipla

**AVISO:** Alguns exercícios desta secção utilizam conjuntos dados disponibilizados no ficheiro de nome `dadosRLM.RData`. Esse ficheiro está na página *web* da disciplina. Deve ser descarregado para a directoria onde tem a sua sessão de trabalho do R e depois carregado, a partir duma sessão do R, utilizando a opção *Load Workspace* no Menu *File* ou, alternativamente, através do comando `load("dadosRLM.RData")`. Na sessão do R devem estar disponíveis (confirme com o comando `ls()`) as *data frames* `brix` (Exercício 2), e `milho` (Exercício 10), além da *data frame* `vinhos` (criada no Exercício 1).

#### EXERCÍCIOS

1. 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, que será ignorado neste Exercício), 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)
```

Há interesse em modelar o teor de flavonóides (variável `V8`), um antioxidante de medição difícil e dispendiosa.

- (a) Execute o comando `plot(vinhos)` e comente o resultado.
- (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 predictoras: `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=vinhos)
```

- (e) Ajuste uma regressão linear múltipla do teor de flavonóides (variável V8) sobre todas as restantes variáveis com o comando `summary(lm(V8 ~ . , data=vinhos[, -1]))`.
- Use o valor do coeficiente de determinação obtido com esse comando para determinar a decomposição da soma dos quadrados totais. Comente os resultados.
  - Compare os coeficientes estimados das variáveis predictoras com os correspondentes coeficientes das variáveis predictoras presentes nos modelos anteriores. Comente.
2. 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*. 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

- Construa as nuvens de pontos correspondentes a cada possível par de variáveis. Calcule os coeficientes de correlação correspondentes a cada gráfico. Comente.
- 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 predictoras. Quantos parâmetros tem este modelo?
- Determine o valor das estimativas dos parâmetros do modelo indicado na alínea anterior.
- Discuta o significado biológico do coeficiente ajustado da variável *Peso*. Quais são as unidades de medida desta estimativa?
- Discuta o significado da ordenada na origem  $b_0$  resultante do ajustamento. Comente.
- 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 2a). Comente.
- 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.

3. Considere uma regressão linear simples numa 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  $\vec{y}$  indica um vector com os  $n$  valores da variável  $Y$ ). Mostre que:

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

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

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

- (d) Deduza a partir do facto que  $\vec{b} = (\mathbf{X}^t \mathbf{X})^{-1} (\mathbf{X}^t \vec{y})$ , as fórmulas para  $b_0$  e  $b_1$  obtidas na abordagem inicial da Regressão Linear Simples.
- (e) Mostre que as variâncias e covariâncias dos estimadores  $\hat{\beta}_0$  e  $\hat{\beta}_1$  dos parâmetros da recta de regressão são dados pelos elementos da matriz  $\sigma^2 (\mathbf{X}^t \mathbf{X})^{-1}$ :

$$V[\hat{\beta}_1] = \frac{\sigma^2}{(n-1)s_x^2} \quad V[\hat{\beta}_0] = \sigma^2 \cdot \left( \frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right) \quad Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}}{(n-1)s_x^2} \cdot \sigma^2.$$

4. (a) Mostre, a partir da sua definição, que a matriz de projecção ortogonal  $\mathbf{H} = \mathbf{X}(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t$  numa regressão linear múltipla é idempotente ( $\mathbf{H}\mathbf{H} = \mathbf{H}$ ) e simétrica ( $\mathbf{H}^t = \mathbf{H}$ ).
- (b) Mostre que a projecção ortogonal sobre o subespaço das colunas da matriz  $\mathbf{X}$ ,  $\mathcal{C}(\mathbf{X})$ , de qualquer vector pertencente a esse mesmo espaço ( $\mathbf{X}\vec{a} \in \mathcal{C}(\mathbf{X})$ ) deixa esse vector invariante.
- (c) Mostre, a partir da expressão do vector dos valores ajustados de  $Y$ ,  $\vec{\hat{y}} = \mathbf{H}\vec{y}$  que, também numa regressão linear múltipla, a média amostral dos valores observados de  $Y$ ,  $\{y_i\}_{i=1}^n$ , é igual à média amostral dos valores ajustados  $\{\hat{y}_i\}_{i=1}^n$ .

5. Considere uma regressão linear múltipla.

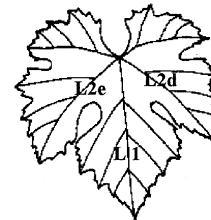
- (a) Considere o vector  $\vec{\mathbf{1}}_n \in \mathbb{R}^n$ , constituído por  $n$  uns. Construa a matriz  $\mathbf{P} = \vec{\mathbf{1}}_n (\vec{\mathbf{1}}_n^t \vec{\mathbf{1}}_n)^{-1} \vec{\mathbf{1}}_n^t$  de projecção ortogonal sobre o subespaço  $\mathcal{C}(\vec{\mathbf{1}}_n) \subset \mathbb{R}^n$  gerado pelo vector  $\vec{\mathbf{1}}_n$ . Mostre que a matriz  $\mathbf{P}$  é simétrica e idempotente.
- (b) Mostre que se verificam as seguintes igualdades:

$$\begin{aligned} SQT &= \|\vec{y} - \mathbf{P}\vec{y}\|^2 = \vec{y}^t (\mathbf{I} - \mathbf{P}) \vec{y} \\ SQR &= \|\mathbf{H}\vec{y} - \mathbf{P}\vec{y}\|^2 = \vec{y}^t (\mathbf{H} - \mathbf{P}) \vec{y} \\ SQRE &= \|\vec{y} - \mathbf{H}\vec{y}\|^2 = \vec{y}^t (\mathbf{I} - \mathbf{H}) \vec{y} \end{aligned}$$

onde  $\vec{y}$  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  $\mathbf{P}$  é a matriz de projecção ortogonal sobre o subespaço  $\mathcal{C}(\vec{\mathbf{1}}_n)$  gerado pelo vector dos  $n$  uns,  $\vec{\mathbf{1}}_n$ .

- (c) Mostre, algebricamente, que  $SQT = SQR + SQRE$ .
6. Seja  $\vec{W}_{k \times 1}$  um vector aleatório. Mostre que se verificam as seguintes propriedades:
- $E[\alpha \vec{W}] = \alpha E[\vec{W}]$ , sendo  $\alpha$  um escalar (não aleatório).
  - $E[\vec{W} + \vec{a}] = E[\vec{W}] + \vec{a}$ , sendo  $\vec{a}$  um vector não aleatório.
  - $V[\alpha \vec{W}] = \alpha^2 V[\vec{W}]$ , sendo  $\alpha$  um escalar (não aleatório).
  - $V[\vec{W} + \vec{a}] = V[\vec{W}]$ , sendo  $\vec{a}$  um vector não aleatório.
  - Considere um segundo vector aleatório  $\vec{U}_{k \times 1}$ . Mostre que  $E[\vec{W} + \vec{U}] = E[\vec{W}] + E[\vec{U}]$ .
7. Considere o conjunto de dados *iris*, disponível no R. Considere apenas as observações das quatro variáveis morfológicas: largura e comprimento de pétalas e sépalas (todas em *cm*) em  $n = 150$  lírios.
- Construa as nuvens de pontos para cada possível par de variáveis. Comente.
  - Ajuste uma regressão linear múltipla da largura das pétalas sobre as restantes três variáveis predictoras. Comente o coeficiente de determinação obtido.
  - Interprete os valores das estimativas dos coeficientes de cada uma das variáveis predictoras.
  - Considere o sinal do parâmetro  $b_j$  associado ao preditor *Sepal.Length*, na regressão linear múltipla acima ajustada. Tendo em conta a nuvem de pontos relacionando a variável resposta *Petal.Width* com o preditor *Sepal.Length*, obtida na alínea 7a), qual seria o sinal do declive nessa recta de regressão? Comente.
  - Construa os intervalos a 95% de confiança para  $\beta_1$ ,  $\beta_2$  e  $\beta_3$ . Comente.
  - Teste se é admissível considerar que um aumento no comprimento das sépalas, mantendo os restantes preditores fixos, está associado a uma diminuição na largura média das pétalas.
8. A *data frame videiras* foi já considerada no Exercício RLS 13. Além da variável *Area* (área foliar), que se deseja modelar, e do preditor *NP* (comprimento da nervura principal), usados nesse Exercício, a *data frame* contém dois outros potenciais preditores da área foliar (ambos em *cm*):

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



- Desenhe as nuvens de pontos para cada par das 4 variáveis observadas. Comente.
- Calcule a matriz de correlações entre todos os pares de variáveis observadas. Comente.
- Descreva o Modelo de regressão linear múltipla resultante de modelar *Area* com base nos três preditores disponíveis.
- Ajuste a regressão múltipla referida na alínea anterior e comente. Em particular, teste o ajustamento global do modelo.
- 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 médio da área foliar de  $7 \text{ cm}^2$ . 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:

- o valor estimado da área foliar;
  - um intervalo de confiança (95%) para o valor esperado da área foliar associado a esses valores das variáveis predictoras;
  - um intervalo de predição (95%) para o valor da área foliar de cada folha individual.
- (h) Estude os resíduos do ajustamento efectuado, bem como os restantes diagnósticos. Comente.
- (i) Ajuste uma regressão linear múltipla análoga, mas logaritmando previamente as quatro variáveis. Diga, justificando, a qual relação de fundo entre as quatro variáveis originais corresponde o modelo agora ajustado.
- (j) Efectue o estudo dos resíduos e restantes quantidades de diagnóstico do modelo ajustado na alínea anterior. Compare com os gráficos obtidos na alínea 8h) e comente.
9. Dezanove escaravelhos da espécie *Haltica oleracea* e vinte escaravelhos da espécie *Haltica carduorum* foram sujeitos a medições morfométricas em quatro variáveis: a distância do sulco transversal à borda posterior do pró-torax (variável *TG*), o comprimento do élitro (variável *Elytra*), o comprimento do segundo segmento das antenas (variável *Second.Antenna*) e o comprimento do terceiro segmento das antenas (variável *Third.Antenna*). As unidades de todas as variáveis *excepto o comprimento do élitro* são micrómetros (milionésima parte do metro,  $\mu m$ ). O comprimento do élitro é dado em centésimas de milímetro.

Alguns dos dados obtidos são indicados na tabela seguinte.

	Species	TG	Elytra	Second.Antenna	Third.Antenna
1	oleracea	189	245	137	163
2	oleracea	192	260	132	217
3	oleracea	217	276	141	192
4	oleracea	221	299	142	213
5	oleracea	171	239	128	158
(...)					
35	carduorum	181	308	157	204
36	carduorum	192	276	154	209
37	carduorum	181	278	149	235
38	carduorum	175	271	140	192
39	carduorum	197	303	170	205
-----					
variância		196.888	502.7085	216.0445	341.8313
média		186.8205	279.2308	147.5385	197.8974

*Haltica oleracea*



Matriz de correlações:

	TG	Elytra	Second.Antenna	Third.Antenna
TG	1.0000000	0.1809792	-0.1671795	-0.07351397
Elytra	0.18097923	1.0000000	0.7265072	0.59184021

```
Second.Antenna -0.16717947 0.7265072      1.0000000    0.58674692
Third.Antenna  -0.07351397 0.5918402      0.5867469    1.00000000
```

No âmbito do estudo dos referidos escaravelhos, pretende-se estimar o comprimento do élitro como função das restantes variáveis. Ajustou-se um modelo às 39 observações, sem distinção de espécies, tendo sido obtidos os seguintes resultados.

```
> summary(flea.beetles4.lm)
Call: lm(formula = Elytra ~ TG + Second.Antenna + Third.Antenna)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -12.8302    42.1040  -0.305  0.76238
TG              0.4874     0.1598   3.050  0.00435
Second.Antenna  0.9703     0.1879   5.164 9.78e-06
Third.Antenna  0.2923     0.1477   1.979  0.05567
---
Residual standard error: 13.62291 on 35 degrees of freedom
Multiple R-Squared:  0.66,    Adjusted R-squared:  0.59
F-statistic: 13.62 on 3 and 35 DF,  p-value: 2.513e-08
```

A matriz de variâncias-covariâncias para os parâmetros estimados é a seguinte:

```
> vcov(flea.beetles4.lm)
              (Intercept)          TG Second.Antenna Third.Antenna
(Intercept)  1772.744496 -5.3085813030  -2.862555366 -1.7882851425
TG            -5.308581  0.0255422781   0.004612691 -0.0007265865
Second.Antenna -2.862555 0.0046126915   0.035306802 -0.0162119398
Third.Antenna  -1.788285 -0.0007265865  -0.016211940  0.0218088275
```

- Complete a tabela, indicando os valores em falta (graus de liberdade, valor calculado da estatística  $F$ ,  $R^2$  ajustado).
- Discuta a qualidade de ajustamento do modelo, tendo em conta a informação disponível. Na sua discussão, inclua um teste formal do ajustamento, indicando as hipóteses em confronto, a natureza da estatística do teste e os pressupostos adicionais cuja validade teve de admitir.
- Interprete o significado biológico da estimativa associada à variável  $TG$ .
- Teste formalmente se é admissível considerar que para cada micrómetro adicional no segundo segmento de antena (e mantendo fixos os restantes preditores), o comprimento do élitro aumenta, em média, menos de 10 micrómetros (*Nota*: atenção às unidades de medida).
- Teste formalmente se é admissível considerar que para cada micrómetro adicional *simultaneamente em cada um dos dois segmentos de antena* (segundo e terceiro segmentos), mantendo fixos os restantes preditores, o comprimento do élitro aumenta, em média, 10 micrómetros (*Nota*: atenção às unidades de medida).
- Teste formalmente se este modelo difere significativamente, quanto ao ajustamento, da regressão linear simples do comprimento do élitro (*Elytra*) sobre o comprimento do segundo segmento antenal (*Second.Antenna*). Comente.

- 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 (°F)	Prec. Junho (in.)	Temp. Junho (°F)	Prec. Julho (in.)	Temp. Julho (°F)	Prec. Agosto (in.)	Temp. Agosto (°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 predictoras. Comente os resultados. Estude os gráficos de resíduos e outros diagnósticos.
- (b) Determine o valor do  $R^2$  modificado. Comente.
- (c) Repita o ajustamento da primeira alínea, mas agora excluindo a variável cronológica  $x_1$  do conjunto de variáveis predictoras. Compare os resultados do ajustamento nos dois casos. Comente.
- (d) Utilize um teste  $t$  ao coeficiente  $\beta_1$  no modelo com todos os preditores, para ver se é possível concluir que os modelos com e sem o preditor  $x_1$  têm ajustamento significativamente diferente.
- (e) Utilize um teste  $F$  parcial para responder à pergunta da alínea anterior. Compare os  $p$ -values obtidos nestes dois testes e discuta a sua relação.
- (f) Com base apenas no ajustamento do modelo completo, efectuado na alínea 10a), diga, justificando:
- Qual a variável preditora cuja exclusão do modelo menos afectaria a qualidade do modelo?
  - Qual o coeficiente de determinação do submodelo resultante da exclusão dessa variável?

- (g) 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.
- (h) Identifique um modelo mais parcimonioso, utilizando o método de exclusão sequencial de variáveis baseado nos testes a  $\beta_j = 0$  ( $\alpha = 0.10$ ). Repita, usando como critério de selecção o valor do Critério de Informação de Akaike (AIC). Efectue ainda uma pesquisa completa dos subconjuntos de cada cardinalidade, usando a função `leaps` do módulo R de igual nome.
- (i) 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.

$$\begin{aligned} z^{\circ}\text{F} &= \frac{5}{9}(z - 32)^{\circ}\text{C} \\ \text{Conversões: } 1 \text{ in} &= 25,4 \text{ mm} \\ 1 \text{ bu/acre (milho)} &= 0.06277 \text{ t ha}^{-1} \end{aligned}$$

11. Pretende-se estudar a evolução de características relacionadas com a frutificação de amoras (*Rubus spp.*), e concretamente modelar o número de frutos vingados por cacho (variável  $v$ ) à custa de outras variáveis preditoras. Como potenciais preditores consideraram-se as variáveis: comprimento dos lançamentos frutíferos (variável  $cl$ , em cm); distância ao solo de cada cacho (variável  $d1$ , em cm); comprimento do raquis, ou seja, do eixo central do cacho (variável  $r$ , em cm); número de botões por cacho (variável  $b$ ). Num primeiro estudo, foram efectuadas 64 observações destas variáveis, para uma única cultivar. As médias e variâncias para cada variável, bem como a matriz de correlações amostrais observadas, foram:

	$v$	$cl$	$d1$	$b$	$r$
Médias	16.43750	440.25000	285.79688	17.53125	27.60938
Variâncias	54.85317	12187.61905	25473.40253	63.64980	139.89261

	$cl$	$d1$	$b$	$v$	$r$
$cl$	1.0000000	0.484277382	0.132235969	0.1452308	0.4348473
$d1$	0.4842774	1.000000000	0.002753756	0.1014318	-0.1313583
$b$	0.1322360	0.002753756	1.000000000	0.9555627	0.6597847
$v$	0.1452308	0.101431793	0.955562651	1.0000000	0.5783831
$r$	0.4348473	-0.131358261	0.659784745	0.5783831	1.0000000

- (a) Considere o modelo de regressão linear múltipla para a variável resposta  $v$ , com as quatro restantes variáveis como preditoras. Qual o intervalo de menor amplitude onde pode garantir, com base na informação disponível até aqui, que está contido o coeficiente de determinação? Justifique e comente o seu resultado.
- (b) Foi ajustada uma regressão linear múltipla para a totalidade das variáveis preditoras acima referidas. Foram obtidos os seguintes resultados gerais.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.586e-01	1.186e+00	0.134	0.8940
$cl$	5.883e-05	3.599e-03	0.016	0.9870
$d1$	4.121e-03	2.218e-03	1.858	0.0681
$b$	9.307e-01	4.780e-02	19.471	<2e-16
$r$	-4.498e-02	3.930e-02	-1.145	0.2570

---

Residual standard error: 2.087 on 59 degrees of freedom

Multiple R-squared: 0.9256, Adjusted R-squared: 0.9206

F-statistic: 183.6 on 4 and 59 DF, p-value: < 2.2e-16

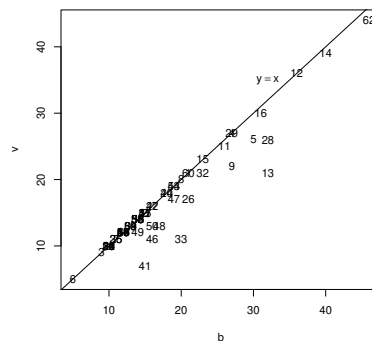
Discuta formalmente a qualidade do ajustamento do modelo.



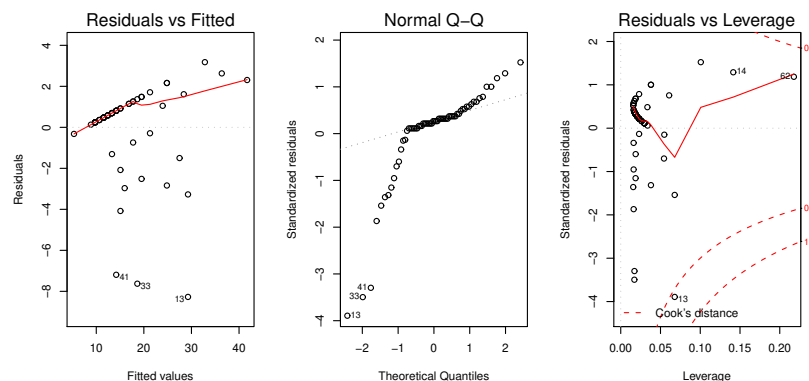
- (c) É admissível afirmar que, por cada centímetro adicional na distância ao solo dum cacho, o número de frutos vingados no cacho aumenta, em média, 0.005 unidades? Responda usando um intervalo a 95% de confiança.
- (d) Deseja-se simplificar o modelo, sem perda significativa na qualidade do ajustamento ( $\alpha = 0.10$ ).
- Justifique brevemente qual o modelo de regressão linear com três preditores que escolheria.
  - Para o modelo que acaba de escolher, calcule os valores da Soma de Quadrados Residual e do coeficiente de determinação  $R^2$ .
  - Complete o algoritmo de exclusão sequencial para determinar o mais simples submodelo possível ( $\alpha = 0.10$ ), sabendo que os coeficientes de determinação para todos os submodelos com dois preditores são os indicados na tabela seguinte. Justifique as suas afirmações.

Preditores	$R^2$	Preditores	$R^2$	Preditores	$R^2$
{c1,d1}	0.02236	{c1,b}	0.9135	{c1,r}	0.3485
{d1,b}	0.9229	{d1,r}	0.3666	{b,r}	0.9179

- (e) Considere agora a regressão linear simples de  $v$  sobre  $b$ , isto é, do número de frutos vingados sobre número de botões, por cacho.
- Diga, justificando, qual a equação da recta de regressão ajustada e qual o significado da estimativa do declive da recta, no contexto do problema em questão.
  - Um investigador chama a atenção para a relação existente entre a variável resposta ( $v$ ) e o preditor ( $b$ ), relação reflectida no seguinte gráfico (**NOTA:** a recta indicada não é a recta de regressão, mas sim a bissetriz dos quadrantes ímpares).



Eis alguns gráficos relativos aos resíduos do ajustamento da regressão linear simples.



Comente os quatro gráficos. Que conclusões pode extrair, no que respeita à relação entre as duas variáveis, e quais as implicações para o modelo de regressão linear simples que acaba de ajustar?

12. 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), *Diâmetro* (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, data=arvores)
[...]
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) , data=arvores)
[...]
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”.
- (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, data=arvores)

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?

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

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

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

14. Considere o modelo de regressão linear *sem preditores*, ou seja, o *Modelo Nulo*:

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

Usando a notação matricial na formulação do modelo, a matrix  $\mathbf{X}$  terá uma única coluna, composta por uns, ou seja,  $\mathbf{X} = \vec{\mathbf{1}}_n$ . Tendo também em atenção a alínea 5a,

- (a) Determine o estimador de mínimos quadrados de  $\beta_0$ .
- (b) Determine o valor esperado e a variância desse estimador de  $\beta_0$ .
- (c) Determine a distribuição de probabilidades do estimador de  $\beta_0$ .
- (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 14d) 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.
15. Considere o modelo com equação base de tipo regressão linear simples, mas sem constante aditiva, também conhecido pelo modelo da *recta forçada à origem*:

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

- (a) Determine o estimador de mínimos quadrados para o parâmetro  $\beta_1$ .
- (b) Determine a distribuição de probabilidades do estimador obtido na alínea anterior, admitindo válidas as restantes hipóteses do Modelo Linear.
16. Considere um modelo de regressão linear múltipla com  $p$  variáveis predictoras, 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,  $\vec{\beta}$ , também se pode escrever como  $\vec{\beta} = \vec{\beta} + (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \vec{\epsilon}$ .
- (c) Deduza *a partir da expressão da alínea anterior*, o vector esperado e a matriz de covariâncias do vector dos estimadores,  $\vec{\beta}$ , ao abrigo do modelo de regressão linear múltipla.
17. Considere os coeficientes de determinação usual ( $R^2$ ) e modificado ( $R_{mod}^2$ ), no contexto duma regressão linear múltipla com  $p$  variáveis predictoras, ajustada com base em  $n$  observações.
- (a) Mostre que se verifica a relação  $R_{mod}^2 = 1 - (1 - R^2) \frac{n-1}{n-(p+1)}$ .
- (b) Mostre que a estatística do teste  $F$  de ajustamento global do modelo se pode escrever apenas à custa de  $R^2$  e  $R_{mod}^2$ , verificando-se  $F_{calc} = \frac{R^2}{R^2 - R_{mod}^2}$ .
- (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.

18. Um analista recebeu dados relativos a 262 eucaliptos com 9.1 anos de idade, nos quais foram medidas três variáveis: o diâmetro da árvore a 1.3 m de altura (variável  $d$ , em  $cm$ ); a altura da árvore (variável  $h$ , em  $m$ ), e volume total, com casca, da árvore (variável  $vcc$ , em  $m^3$ ). Pretende-se modelar a variável  $vcc$ , por ser a de mais difícil medição.

Após uma análise exploratória, foi ajustado um modelo de regressão linear múltipla com as três variáveis logaritimizadas e com as seguintes estimativas e erros padrão correspondentes:

	Estimate	Std. Error
(Intercept)	-9.935185	4.224875e-15
log(d)	1.819071	1.398336e-15
log(h)	1.070288	2.332038e-15

Sabendo que o coeficiente de determinação foi *exactamente*  $R^2 = 1$  diga, justificando:

- (a) Qual o valor da estimativa da variância dos erros aleatórios do modelo? Comente.
- (b) Diga o que se pode concluir ao comparar, com base num teste  $F$  parcial, este modelo com qualquer seu submodelo.
- (c) Diga, justificando, qual a equação não linear entre  $vcc$ ,  $d$  e  $h$  (sem logaritmos) a que corresponde o modelo agora ajustado.
- (d) Qual a conclusão que o analista pode retirar sobre a forma como foram obtidos os valores da variável  $vcc$  que lhe foram fornecidos? Em que medida é que essa conclusão viola os pressupostos do modelo linear?