Estatística e Delineamento Experimental - 2024-25

3 Regressão Linear - Abordagem Inferencial

EXERCÍCIOS PRÁTICOS

1. Considere o conjunto de dados `iris`, disponível no `R`, referente a medições morfométricas de lírios. Considere apenas as observações das quatro variáveis morfométricas: largura e comprimento de pétalas e sépalas (todas em cm) em $n = 150$ lírios. Admita que se trata duma amostra extraída aleatoriamente duma população mais vasta.

   (a) Considere 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.

      i. Ajuste a recta de regressão de largura (y) sobre comprimento (x) das pétalas.

      ii. Obtenha estimativas das variâncias e dos desvios padrão dos estimadores dos parâmetros da recta populacional, β_0 e β_1.

      iii. 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".

      iv. 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, $0.5cm$".

      v. 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 $0.5cm$".

      vi. Estime o valor esperado da largura da pétala para lírios cuja pétala tenha comprimento 4.5cm. Construa um intervalo de confiança para esse valor esperado.

      vii. 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.

   (b) Considere agora a relação entre largura da pétala (`Petal.Width` e as restantes 3 variáveis morfométricas.

      i. Construa as nuvens de pontos para cada possível par de variáveis. Comente.

      ii. Ajuste uma regressão linear múltipla da largura das pétalas sobre as restantes três variáveis preditoras. Comente o coeficiente de determinação obtido.

      iii. Interprete os valores das estimativas dos coeficientes de cada uma das variáveis preditoras.

      iv. 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 1(b)i), qual seria o sinal do declive nessa recta de regressão? Comente.

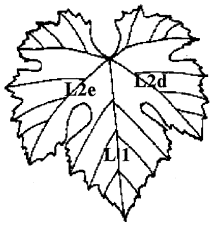
      v. Obtenha intervalos a 95% de confiança para β_1, β_2 e β_3. Comente.

      vi. 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.

   (c) Considere novamente a relação entre largura da pétala (`Petal.Width`, variável y) e comprimento da pétala (`Petal.Length`, variável x). Estude os gráficos dos resíduos para detectar a existência de eventuais problemas com os pressupostos do modelo. Comente as suas conclusões.

2. A medição rigorosa de áreas foliares faz-se através de técnicas destrutivas. Deseja-se obter um modelo que permita estimar áreas foliares de castas de videiras, utilizando variáveis preditoras que possam ser

medidas sem arrancar as folhas da videira. Para tal, na Secção de Horticultura do ISA, foram seleccionadas aleatoriamente 200 folhas de cada uma de três castas: Fernão Pires, Vital e Água Santa. Em cada folha mediu-se a área foliar (variável **Area**, em cm^2), comprimento da nervura principal da folha (variável **NP**, em cm), o comprimento da nervura lateral esquerda (variável **NLesq**, em cm), o comprimento da nervura lateral direita (variável **NLdir**, em cm). O ficheiro com esses dados encontra-se disponível na página *web* da disciplina em formato **txt** de nome **videiras.txt**. Responda às seguintes alíneas para o conjunto dos 600 pares de observações, não distinguindo as castas.



(a) Desenhe as nuvens de pontos para cada par das 4 variáveis observadas. Comente.

(b) Calcule a matriz de correlações entre todos os pares de variáveis observadas. Comente.

(c) Descreva e ajuste o Modelo de regressão linear múltipla resultante de modelar `Area` com base nos três preditores disponíveis.

(d) 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 cm^2. Repita o teste, mas agora utilizando um nível de significância $\alpha = 0.05$. Comente.

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

(f) 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.

(g) Relativamente à regressão múltipla ajustada, comente o valor do coeficiente de determinaçção. Em particular, teste o ajustamento global do modelo.

(h) Considere agora o problema de prever as áreas foliares (variável **Area**) apenas a partir de medições do comprimento da nervura principal (variável **NP**).

i. Construa a nuvem de pontos das áres foliares e comprimentos de nervura principal.

ii. Considere as transformações logarítmicas de ambas as variáveis e construa a nuvem de pontos usando os dados logaritmizados.

iii. Ajuste a recta de regressão de log-áreas sobre log-comprimentos da nervura principal e trace-a sobre a nuvem de pontos obtida na alínea anterior. Comente.

iv. Qual a relação não linear entre `Area` e `NP` que corresponde à recta ajustada na alínea anterior? Usando o comando `curve`, trace-a sobre a nuvem de pontos construída na alínea 2(h)i. Comente.

v. Construa um intervalo a 95% de confiança para o declive da recta de regressão populacional relacionando as variáveis log-transformadas. Interprete o significado biológico do resultado obtido.

vi. Um investigador comenta que, sendo a área foliar uma medida bidimensional e o comprimento da nervura unidimensional, seria plausível que a área seja proporcional ao quadrado do comprimento. Efectue um teste de hipóteses para avaliar a validade desta tese.

vii. Valide os pressupostos do modelo de regressão linear ajustado, analisando o comportamento dos resíduos e restantes ferramentas de diagnóstico. Comente. Considere agora a regressão linear entre as variáveis originais (não logaritmizadas) e repita o estudo dos resíduos. Comente.

3. 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. O ficheiro com esses dados encontra-se disponível na página *web* da disciplina em formato **txt** de nome **milho.txt**.

	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. Efectue o teste F de ajustamento global. Comente o resultado obtido.

(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 nos dois casos. Comente.

(d) Utilize um teste t ao coeficiente β_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 3a), diga, justificando:

i. Qual a variável preditora cuja exclusão do modelo menos afectaria a qualidade do modelo?

ii. 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).

(i) Estude os gráficos de resíduos e outros diagnósticos.

4. 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 ficheiro disponível na página *web* da disciplina em formato **txt** de nome **ameixas.txt**.

(a) Construa a nuvem de pontos de `peso` *vs* `diametro`. Comente a relação de fundo obtida. Ajuste uma regressão linear simples de `peso` sobre `diametro` e trace a recta de regressão ajustada sobre a nuvem de pontos.

(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) Teste formalmente se o modelo parabólico da alínea anterior se ajusta de forma significativamente melhor que o modelo linear inicial. Comente.

(d) Inspeccione os resíduos do modelo parabólico ajustado e comente.

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

5. 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, 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 e o maior valor da estatística F".

(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:

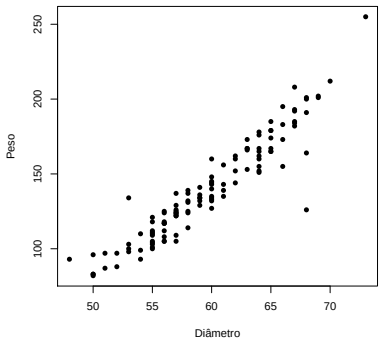
```
Call: lm(formula = Altura ~ Diametro + Volume, data=arvores)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.2958     9.0866   9.167 6.33e-10
Diametro     -1.8615     1.1567  -1.609   0.1188
Volume        0.5756     0.2208   2.607   0.0145

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?

6. Num ensaio pretende-se relacionar o peso de pêras (em g) com o seu diâmetro (em mm), tendo sido feitas medições em $n = 120$ pêras. Eis alguns indicadores relativos aos resultados obtidos, bem como a nuvem de pontos correspondente:

```
     Peso            Diâmetro
Min. :   82.0     Min. :  48.00
Mean :  139.8     Mean :  59.71
Max. :  255.0     Max. :  73.00
Var. : 1131.675   Var. :  27.4688375
```

O coeficiente de correlação entre as variáveis observadas é $r = 0.9397929$.



(a) Ajustou-se um modelo de regressão linear simples de peso sobre diâmetro.

i. Qual o coeficiente de determinação da recta de regressão ajustada e como deve ser interpretado?

ii. Teste se o modelo deve ser considerado inútil ($\alpha = 0.05$).

iii. Determine a equação da recta de regressão ajustada. A partir de que diâmetros é que o modelo ajustado prevê pesos positivos? Comente o seu resultado.

iv. Justifique qual a observação à qual corresponde o maior efeito alavanca. Para essa observação:

A. Calcule o efeito alavanca. Como se compara o valor obtido com o efeito alavanca médio?

B. Calcule o resíduo (usual).

(b) Foi seguidamente ajustado um modelo polinomial de segunda ordem, para prever o peso a partir do diâmetro, tendo sido obtido o seguinte ajustamento:

```
Call: lm(formula = Peso ~ Diâmetro + I(Diâmetro^2), data = Pera)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   134.80456  118.14750   1.141   0.2562
Diâmetro       -5.89331    3.95255  -1.491   0.1386
I(Diâmetro^2)   0.09935    0.03289   3.021   0.0031
---
Residual standard error: 11.17 on 117 degrees of freedom
Multiple R-squared: 0.8917,Adjusted R-squared: 0.8898
F-statistic: 481.5 on 2 and 117 DF,  p-value: < 2.2e-16
```

Diga, justificando brevemente, se este modelo quadrático é significativamente melhor que o modelo linear considerado antes.

7. Pretende-se estudar as características do solo que afectam a produção da biomassa aérea de *Spartina alterniflora*, uma herbácea perene comum em sistemas estuarinos-lagunares. No âmbito do estudo, foram amostrados 45 solos escolhidos ao acaso, tendo sido observadas as seguintes variáveis: biomassa (`BIO`, em $g\ m^{-2}$), salinidade (`SAL`), acidez (`pH`) e os teores em potássio (`K`), sódio (`Na`) e zinco (`Zn`), estes últimos em *ppm*.

Eis alguns indicadores de síntese dos valores observados (extremos, média, mediana, quartis e desvio padrão), bem como a matriz de correlações das variáveis observadas.

```
     BIO           SAL             pH             K               Na             Zn
Min.   : 236  Min.   :24.00  Min.   :3.200  Min.   : 350.7  Min.   : 7886  Min.   : 0.2105
1st Qu.: 416  1st Qu.:27.00  1st Qu.:3.450  1st Qu.: 528.6  1st Qu.:11345  1st Qu.:13.9852
Median : 824  Median :30.00  Median :4.450  Median : 773.3  Median :14752  Median :19.5880
Mean   :1001  Mean   :30.27  Mean   :4.602  Mean   : 797.6  Mean   :16597  Mean   :17.8752
3rd Qu.:1560  3rd Qu.:33.00  3rd Qu.:5.200  3rd Qu.: 954.1  3rd Qu.:20436  3rd Qu.:22.6758
Max.   :2436  Max.   :38.00  Max.   :7.450  Max.   :1441.7  Max.   :35186  Max.   :31.2865
sd     : 660  sd     : 3.72  sd     :1.247  sd     : 297.6  sd     : 6882  sd     : 8.2798
```

```
> cor(biomassa)
        BIO     SAL      pH       K      Na      Zn
BIO  1.0000 -0.1032  0.7742 -0.2046 -0.2721 -0.6244
SAL -0.1032  1.0000 -0.0513 -0.0206  0.1623 -0.4208
pH   0.7742 -0.0513  1.0000  0.0192 -0.0377 -0.7222
K   -0.2046 -0.0206  0.0192  1.0000  0.7921  0.0736
Na  -0.2721  0.1623 -0.0377  0.7921  1.0000  0.1170
Zn  -0.6244 -0.4208 -0.7222  0.0736  0.1170  1.0000
```

(a) Inicialmente pretende-se modelar a biomassa através duma regressão linear simples. Responda, justificando, às seguintes perguntas.

i. Qual a melhor variável preditora para modelar a biomassa através duma recta de regressão? Qual o coeficiente de determinação associado e como interpreta esse valor?

ii. Em relação à recta de regressão que escolheu no ponto anterior,

A. Calcule a equação da recta ajustada.

B. Mostre que a estimativa da variância comum dos erros aleatórios é 178 565.9.

C. Construa um intervalo de predição (95%) para uma observação individual da biomassa aérea de *Spartina alterniflora*, num local onde o solo tenha pH 4.5.

(b) Foi ajustado de seguida um modelo de regressão linear múltipla da biomassa sobre as cinco restantes variáveis observadas, com os seguintes resultados:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.252e+03  1.235e+03   1.014  0.31674
SAL         -3.029e+01  2.403e+01  -1.260  0.21508
pH           3.055e+02  8.788e+01   3.477  0.00126
K           -2.851e-01  3.484e-01  -0.818  0.41817
Na          -8.673e-03  1.593e-02  -0.544  0.58926
Zn          -2.068e+01  1.505e+01  -1.373  0.17746
---
Residual standard error: 398.3 on 39 degrees of freedom
Multiple R-squared: 0.6773,     Adjusted R-squared: 0.6359
F-statistic: 16.37 on 5 and 39 DF,  p-value: 1.082e-08
```

i. Sem efectuar cálculos, comente a seguinte afirmação: *mesmo sem efectuar testes adicionais, a análise da última coluna da tabela de resultados, relativa aos valores de prova (p-values), implica que uma regressão linear simples sobre a variável* `pH` *produziria um ajustamento que não difere significativamente do modelo com os cinco preditores.* Justifique.

ii. Independentemente da sua resposta na alínea anterior, teste formalmente se o ajustamento do modelo de regressão linear múltipla com os cinco preditores difere significativamente ($\alpha = 0.10$) do ajustamento do modelo de regressão linear simples considerado na pergunta 7a. Comente.

iii. Pode admitir-se que, sendo tudo o resto igual, por cada ppm adicional no teor de zinco do solo a biomassa aérea desta planta diminui, em média, 40 $g\ m^{-2}$? Responda a esta questão utilizando um intervalo a 95% de confiança.

iv. Partindo do modelo acima, um algoritmo de exclusão sequencial baseado nos testes-t aos coeficientes β_j do modelo, seleccionou um submodelo com quatro preditores. Diga, justificando:

A. Quais são esses quatro preditores;

B. Qual o valor do coeficiente de determinação desse submodelo.

8. Num estudo sobre alfaces considerou-se útil modelar a condutividade eléctrica do solo a partir da sua composição química. Em 27 amostras de solo foram medidas, além da condutividade eléctrica (variável `CE`, em microsiemens por centímetro, $\mu S/cm$), a concentração de catiões de sódio (variável `Na`), potássio (variável `K`), cálcio (variável `Ca`) e magnésio (variável `Mg`), todas em $cmol/kg$. As correlações entre cada par de variáveis, são indicados de seguida.

```
         CE       Na        K      Ca      Mg
CE  1.00000 -0.08618 -0.05741 0.17778 0.55346
Na -0.08618  1.00000  0.33811 0.20362 0.30437
K  -0.05741  0.33811  1.00000 0.08943 0.38161
Ca  0.17778  0.20362  0.08943 1.00000 0.64198
Mg  0.55346  0.30437  0.38161 0.64198 1.00000
```

Foi ajustado um modelo de regressão linear múltipla utilizando todas as concentrações de catiões como preditores. Obtiveram-se os seguintes resultados:

```
Call: lm(formula = CE ~ Na + K + Ca + Mg, data = Alfaces)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -815.76     446.45  -1.827  0.08126 .
Na           -102.21      84.52  -1.209  0.23935
K            -405.78     207.13  -1.959  0.06291 .
Ca            -57.68      29.61  -1.948  0.06432 .
Mg            668.17     145.34   4.597  0.00014 ***
---
Residual standard error: 68.07 on 22 degrees of freedom
Multiple R-squared: 0.5147, Adjusted R-squared: 0.4264
F-statistic: 5.832 on 4 and 22 DF,  p-value: 0.002344
```

(a) Teste o ajustamento global do modelo. Discuta a qualidade desse ajustamento. Na sua discussão, tenha também em conta os valores dos coeficientes de determinação usual e modificado.

(b) Interprete, no contexto do problema sob estudo, o significado do valor -57.68 na primeira coluna da tabela, indicando as unidades de medida do referido valor.

(c) Um investigador afirma que quando aumenta a concentração de catiões de sódio, diminui a condutividade eléctrica média. Admitindo a validade do modelo, e dando o ónus da prova à hipótese do investigador, qual a conclusão a que se pode chegar com base nos dados ($\alpha = 0.05$)?

(d) Foi afirmado que com base nos resultados do modelo acima ajustado, a única de entre as quatro variáveis preditoras que é importante na modelação da condutividade eléctrica é a variável `Mg`.

i. Sem fazer quaisquer contas, diga se considera legítima esta afirmação.

ii. Independentemente da sua resposta na alínea anterior, compare formalmente o modelo acima apresentado com a regressão linear simples em que a variável resposta `CE` tem como único preditor a variável `Mg`. Comente.