Exercícios AN(C)OVA - Modelos Matemáticos - 2016-17

2 Análise de Variância

AVISO: Os conjuntos de dados necessários nesta secção são `toxicos` (Exercício 1), `CO2` (Exercício 2), `terrenos` (Exercício 3) e `absorcao` (Exercício 5).

1. Um estudo sobre a presença de substâncias tóxicas no meio ambiente, associadas à utilização de um pesticida envolve análises efectuadas por quatro laboratórios diferentes. Suspeita-se que os resultados destas análises diferem devido à utilização de diferentes técnicas laboratoriais o que, a ser verdade, exigiria maior controlo das técnicas laboratoriais usadas.

   A fim de avaliar a questão, foram entregues a cada laboratório 6 pequenos contentores com solo recolhido aleatoriamente num terreno que antes fora tratado com o referido pesticida. Os resultados laboratoriais das análises químicas medem a concentração dum composto químico nocivo (em ppm). Os valores observados são indicados na tabela (disponíveis na *data frame* `toxicos`).

	Laboratórios			
	1	2	3	4
	53.2	51.0	47.4	51.0
	54.5	40.5	46.2	51.5
	52.8	50.8	46.0	48.8
	49.3	51.5	45.3	49.2
	50.4	52.4	48.2	48.3
	53.8	49.9	47.1	49.8
Média ($\overline{y}_{i.}$)	52.3333	49.3500	46.7000	49.7667
variância amostral (s_i^2)	4.1507	19.4750	1.1200	1.5867

   - A média global das observações é $\overline{y}_{..} = 49.5375$;
   - a variância amostral da totalidade das observações é $s_y^2 = 9.8868$.

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

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

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

   ii. Utilizando, no `R`, o comando `aov(concentracao ~ laboratorio , data=toxicos)`.

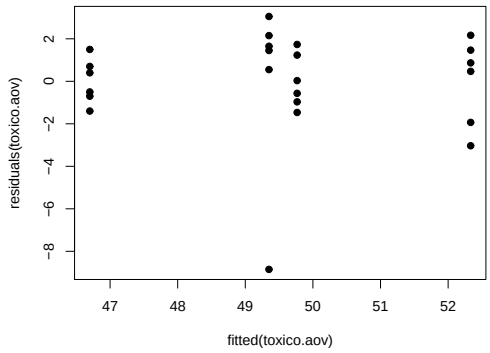
   (c) Formalize e efectue o teste F adequado ao problema acima referido, com um nível de significância de 5%. Parece-lhe necessário exigir aos laboratórios a adopção de uma técnica padronizada?

   (d) Repita agora o teste de hipóteses da alínea anterior, mas utilizando o nível de significância de 1%. Compare os resultados e comente.

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

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

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



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5. Uma experiência pretende estudar o efeito de tempo de exposição e temperatura duma dada solução sobre a dissolução dum produto químico presente num material que é submerso na solução. O estudo apenas se debruça sobre três níveis de temperatura e três tempos de exposição. Os dados recolhidos são as absorções registadas do produto químico pela solução (disponíveis na *data.frame* `absorcao`):

Quantidade (mg) absorvida do químico			
TEMPO DE EXPOSIÇÃO	TEMPERATURA (ºC)		
	T_1	T_2	T_3
E_1	35.5	91.2	70.1
	29.7	100.7	64.1
	31.5	82.4	70.1
E_2	52.5	71.0	79.4
	53.3	77.0	77.7
	55.0	75.6	75.1
E_3	85.9	87.0	83.0
	85.2	86.1	87.0
	80.2	88.1	78.5

```
Médias de células

 exposicao:temperatura
         temperatura
exposicao T1    T2    T3
       E1 32.23 91.43 68.10
       E2 53.60 74.53 77.40
       E3 83.77 87.07 82.83
```

(a) Identifique o delineamento experimental utilizado e descreva a equação do modelo ANOVA correspondente.

(b) Indique as hipóteses nulas e alternativas nos testes à existência de cada tipo de efeito previsto na equação do modelo.

(c) Efectue os testes de hipóteses referidos na alínea anterior e indique as suas conclusões ao nível de significância de 5%.

(d) Explicite estimativas da quantidade média de químico absorvido para as combinações de tratamentos que dão origem à maior e à menor absorção.

(e) Reformule o problema considerando que as nove combinações de tratamentos constituem uma Análise de Variância a um factor com nove níveis. Determine se essas nove combinações diferem ao nível de 5%. Comente.

6. Considere de novo os dados do Exercício 5. Utilize o `R`, e a *data.frame* `absorcao`, para responder às alíneas seguintes.

(a) Compare as tabelas-resumo da ANOVA resultantes de trocar a ordem dos factores na fórmula indicando o modelo factorial a dois factores, isto é, considere primeiro a fórmula

```
abs ~ temperatura * exposicao
```

e depois a fórmula

```
abs ~ exposicao * temperatura
```

Comente.

(b) Considere agora o subconjunto de dados resultante de omitir a primeira observação da primeira célula (de valor 35.5) e as duas últimas observações da última célula (de valores 87.0 e 78.5). Neste caso, estar-se-ia a trabalhar com um *delineamento desequilibrado*. Repita a alínea anterior, mas para este novo conjunto de dados. Comente.

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

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

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

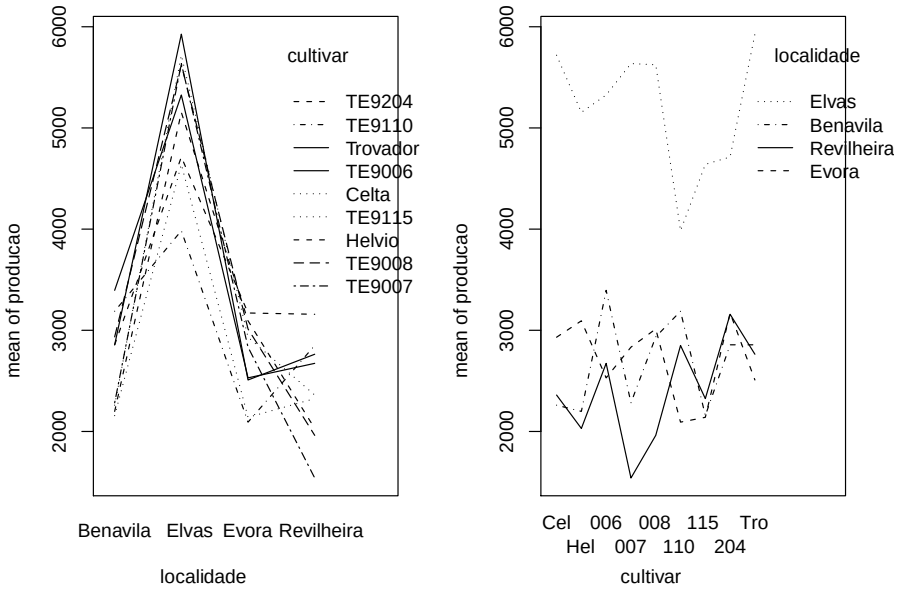
```
                    Df    Sum Sq  Mean Sq  F value
localidade         ??? 183759916      ??? 234.9531
cultivar           ???       ???   964060      ???
localidade:cultivar ???      ???      ???   4.0768
Residuals          ???  28156076   260704
```

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

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

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

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



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

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

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

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

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

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

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

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

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

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

10. Considere um Modelo ANOVA a 1 Factor. Tendo em conta as expressões dos estimadores dos parâmetros,

    (a) Mostre que os estimadores têm a seguinte distribuição:
    - $\hat{\mu}_1 \cap \mathcal{N}(\mu_1\,,\,\sigma^2/n_1)$.
    - $\hat{\alpha}_i \cap \mathcal{N}\left(\alpha_i\,,\,\sigma^2\,(\frac{1}{n_1}+\frac{1}{n_i})\right) \quad (i>1)$.

    (b) Sabendo que a Soma de Quadrados Residual, $SQRE$, é independente dos estimadores de qualquer dos parâmetros, mostre que:
    - um intervalo de confiança a $(1-\alpha)\times 100\%$ de confiança para μ_1 é dado por

$$\left]\ \overline{y}_{1\cdot} - t_{\alpha/2\,(n-k)}\cdot\sqrt{QMRE/n_1}\quad,\quad \overline{y}_{1\cdot} + t_{\alpha/2\,(n-k)}\cdot\sqrt{QMRE/n_1}\ \right[$$

    - um intervalo de confiança a $(1-\alpha)\times 100\%$ de confiança para α_i $(i>1)$ é dado por

$$\left]\ (\overline{y}_{i\cdot}-\overline{y}_{1\cdot}) - t_{\alpha/2\,(n-k)}\cdot\sqrt{QMRE\cdot\left(\tfrac{1}{n_1}+\tfrac{1}{n_i}\right)}\quad,\quad (\overline{y}_{i\cdot}-\overline{y}_{1\cdot}) + t_{\alpha/2\,(n-k)}\cdot\sqrt{QMRE\cdot\left(\tfrac{1}{n_1}+\tfrac{1}{n_i}\right)}\ \right[$$

3 Análise de Covariância

1. Considere as medições sobre folhas de videira introduzidas no Exercício 16 da Regressão Linear (*data frame* **videiras**).

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

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

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

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

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

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

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

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

2. Considere as medições sobre folhas de videira introduzidas no Exercício 16 da Regressão Linear (*data frame* **videiras**).

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

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

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

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

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

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

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

i. a recta obtida ignorando as castas de cada folha;
ii. as três rectas obtidas para cada casta (utilize cores diferentes na sua representação).

(h) Na nuvem de pontos entre as variáveis (não logaritmizadas) que obteve na alínea 2a, trace as seguintes curvas (tendo em conta o resultado das regressões lineares que ajustou):
i. a curva associada à relação entre área foliar e comprimento da nervura principal, independentemente da casta de origem de cada folha.
ii. as três curvas associadas às relações não lineares entre área foliar e comprimento da nervura principal, para cada casta.

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

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

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

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

(c) Ajuste agora um modelo ANCOVA para largura de pétalas, que cruze a regressão linear simples sobre a largura das sépalas com o factor *Species*. Em particular,
i. Desenhe as rectas de regressão linear obtidas para cada espécie, em cima da nuvem de pontos da alínea 3a).
ii. Compare o valor do coeficiente de determinação obtido agora, com o valor de R^2 obtido quando se ajustava uma única recta de regressão, independentemente das espécies. Comente.
iii. A informação disponível sugere que as rectas de regressão para as espécies *versicolor* e *virginica* são paralelas. Teste formalmente esta hipótese.

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

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

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