INSTITUTO SUPERIOR DE AGRONOMIA

MODELOS MATEMÁTICOS – 2015/16

Algumas resoluções de Exercícios de Análise de (Co-)Variância

2. Análise de Variância

1. Pretende-se modelar a variável resposta numérica `concentracao`, tendo como variável explicativa apenas os quatro diferentes laboratórios.

   (a) Estamos perante um delineamento a um factor (`laboratorio`), com $k = 4$ níveis (os 4 laboratórios). Para cada nível há $n_i = 6$ observações, e sendo este número igual para todos os laboratórios estamos perante um delineamento equilibrado. O Modelo ANOVA a 1 factor correspondente é:

   i. $Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij},\ \forall\, i = 1, 2, 3, 4\ ,\ j = 1, 2, ..., 6\ ,\quad$ com $\alpha_1 = 0$, onde
      - Y_{ij} indica a concentração do produto químico para a j-ésima repetição observada no i-ésimo laboratório;
      - μ_1 indica a concentração média no primeiro laboratório ($i = 1$);
      - α_i indica o efeito (acréscimo em relação à média do primeiro laboratório) associado ao i-ésimo laboratório; e
      - ϵ_{ij} indica o erro aleatório associado à observação Y_{ij}.

   ii. $\epsilon_{ij} \cap \mathcal{N}(0, \sigma^2),\ \forall\, i, j.$

   iii. $\{\epsilon_{ij}\}_{i,j}$ constitui um conjunto de variáveis aleatórias independentes.

   (b) O quadro-resumo da ANOVA tem a seguinte estrutura:

Fonte	g.l.	SQ	QM	F_{calc}
Factor	$k-1$	$SQF = \sum_{i=1}^{k} n_i \cdot (\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2$	$QMF = \frac{SQF}{k-1}$	$\frac{QMF}{QMRE}$
Resíduos	$n-k$	$SQRE = \sum_{i=1}^{k} (n_i - 1)\, s_i^2$	$QMRE = \frac{SQRE}{n-k}$	
Total	$n-1$	$SQT = (n-1)\, s_y^2$	–	–

   No nosso caso, $k{=}4$; $n_i{=}6{=}n_c\ (\forall\, i)$; $n{=}n_c \times k{=}24$;
   $SQRE = (n_c - 1) \sum_{i=1}^{k} s_i^2 = 5 \times (4.1507 + 19.4750 + 1.1200 + 1.5867) = 131.662$; e

$$\begin{aligned} SQF &= n_c \cdot \sum_{i=1}^{k} (\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2 \\ &= 6\,[(52.3333{-}49.5375)^2 + (49.35{-}49.5375)^2 + (46.7{-}49.5375)^2 + (49.7667{-}49.5375)^2] \\ &= 95.73356. \end{aligned}$$

   Alternativamente, seria possível calcular SQT a partir da variância amostral da totalidade das observações da variável resposta ($SQT = (n{-}1)\, s_y^2 = 23{\times}9.8868 = 227.3964$) e subtrair-lhe uma das Somas de Quadrados anteriores para obter a outra. As pequenas diferenças

nos valores obtidos por estas duas abordagens resultam dos erros de arredondamento nos valores das médias e variâncias de nível dados no enunciado.
Assim, $QMRE = \frac{SQRE}{n-k} = \frac{131.662}{20} = 6.5831$ e $QMF = \frac{SQF}{k-1} = \frac{95.73356}{3} = 31.91119$.
Finalmente, $F_{calc} = \frac{QMF}{QMRE} = \frac{31.91119}{6.5831} = 4.847441$.
Os valores obtidos podem ser confirmados (a menos de erros resultantes dos arredondamentos com que são apresentadas no enunciado as médias e variâncias de nível), utilizando os dados disponíveis na *data frame* `toxicos` e os comandos do `R`.

```
> summary(aov(concentracao ~ laboratorio, data=toxicos))
            Df Sum Sq Mean Sq F value Pr(>F)
laboratorio  3  95.73   31.91   4.848 0.0108 *
Residuals   20 131.66    6.58
```

(c) Pede-se um teste F, que neste contexto significa perguntar se se deve admitir a igualdade das concentrações médias nos quatro laboratórios (H_0) ou se se opta pela hipótese alternativa (H_1). Mais concretamente:

Hipóteses: $H_0 : \alpha_i = 0 \ , \ \forall \ i = 2, 3, 4$ vs. $H_1 : \exists\, i = 2, 3, 4$ tal que $\alpha_i \neq 0$.
Estatística do teste: $F = \frac{QMF}{QMRE} \ \cap \ F_{(k-1\,,\,n-k)}$, sob H_0.
Nível de significância: $\alpha = 0.05$.
Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(3,20)} \ = \ 3.10$.
Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 4.848$. É um valor significativo ao nível $\alpha = 0.05$ e rejeita-se H_0 a favor da hipótese de que existem efeitos do Factor, ou seja, de que será necessário verificar a uniformidade dos protocolos de análise dos laboratórios.

(d) A alteração do nível de significância não implica alterações nos dois primeiros passos do teste. Quanto aos restantes,

Nível de significância: $\alpha = 0.01$.
Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.01(3,20)} \ = \ 4.94$.
Conclusões: O valor da estatística do teste não depende do nível de significância e continua a ser $F_{calc} = 4.848$. Mas ao nível $\alpha = 0.01$ este já não é um valor da região crítica, pelo que a esse nível não se rejeita a hipótese nula. O facto de a conclusão mudar entre os níveis de significância $\alpha = 0.05$ e $\alpha = 0.01$ significa que o valor de prova (*p-value*) do valor $F_{calc} = 4.848$ estará entre esses dois níveis, isto é, $0.01 < p < 0.05$, facto que se confirma no quadro-resumo produzido pelo `R`.

(e) A matriz do modelo, $\mathbf{X}$, será constituída por quatro colunas: uma coluna de $n = 24$ uns e as colunas indicatrizes dos segundo, terceiro e quarto níveis do factor($\mathcal{I}_2$, $\mathcal{I}_3$ e $\mathcal{I}_4$), como se pode confirmar através do comando referido no enunciado:

```
> toxicos.aov <- aov(concentracao ~ laboratorio, data=toxicos)
> model.matrix(toxicos.aov)
   (Intercept) laboratorio2 laboratorio3 laboratorio4
1            1            0            0            0
2            1            0            0            0
3            1            0            0            0
4            1            0            0            0
5            1            0            0            0
6            1            0            0            0
7            1            1            0            0
8            1            1            0            0
```

```
9           1           1           0           0
10          1           1           0           0
11          1           1           0           0
12          1           1           0           0
13          1           0           1           0
14          1           0           1           0
15          1           0           1           0
16          1           0           1           0
17          1           0           1           0
18          1           0           1           0
19          1           0           0           1
20          1           0           0           1
21          1           0           0           1
22          1           0           0           1
23          1           0           0           1
24          1           0           0           1
```

(f) Os valores ajustados $\hat{Y}_{ij}$, numa ANOVA a um factor, são as médias amostrais do nível a que cada observação pertence. Assim, tem-se:

```
> fitted(toxicos.aov)
       1        2        3        4        5        6        7        8
52.33333 52.33333 52.33333 52.33333 52.33333 52.33333 49.35000 49.35000
       9       10       11       12       13       14       15       16
49.35000 49.35000 49.35000 49.35000 46.70000 46.70000 46.70000 46.70000
      17       18       19       20       21       22       23       24
46.70000 46.70000 49.76667 49.76667 49.76667 49.76667 49.76667 49.76667
```

As médias aqui indicadas são as que também eram dadas (arredondadas a duas casas decimais) na penúltima linha da tabela do enunciado.

(g) O facto dos resíduos se encontrarem empilhados em quatro colunas é o reflexo natural do facto, referida na alínea anterior, que há apenas quatro diferentes valores ajustados nesta ANOVA: as quatro médias amostrais de cada laboratório. Assim, apenas há quatro diferentes valores no eixo horizontal, a que correspondem os valores ajustados $\hat{y}_{ij} = \overline{y}_{i.}$. Do gráfico não parecem surgir indicações de grandes diferenças na variância dos resíduos em cada nível, excepção feita para a segunda coluna, onde surge um resíduo atípico, de valor inferior a -8. É possível identificar o laboratório a que se refere essa observação, uma vez que a média amostral correspondente excede de pouco o valor 49: trata-se do laboratório 2 (cuja média amostral é 49.35). Em particular, trata-se da segunda observação nesse laboratório, cujo valor (40.5) é inferior em mais de oito unidades ao valor médio do laboratório. Assim, a observação a que corresponde o referido resíduo é a observação y_{22}.

2. Neste exercício sobre os grãos de café em Angola, não existem os dados originais, sendo apenas conhecida a tabela do enunciado, com as médias e variâncias amostrais de cada região.

(a) A variável resposta Y é a percentagem do peso total de grãos sem defeito. Para explicar eventuais diferenças nos valores médios populacionais desta variável, apenas se dispõe de um factor: o factor região, com $k = 6$ níveis (as seis regiões indicadas no enunciado). O modelo ANOVA correspondente é assim o modelo a um factor, semelhante ao do primeiro exercício, mas em que agora o factor tem $k = 6$ níveis, existindo ao todo $n = 66$ observações repartidas de forma equilibrada pelos seis níveis: $n_i = 11$, para qualquer $i = 1, 2, ..., 6$.

i. $Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij}$, $\forall\, i = 1, 2, 3, 4, 5, 6$, $j = 1, 2, ..., 11$, com $\alpha_1 = 0$, onde

- Y_{ij} indica a percentagem do peso de grãos sem defeito, no j-ésimo lote observado na região i;

- μ_1 indica a percentagem média de peso de grãos sem defeito na primeira região ($i = 1$) que, na ordem da tabela, é a região de Cabinda;
- α_i indica o efeito (acréscimo em relação à média do Cabinda) da região i; e
- ϵ_{ij} indica o erro aleatório associado à observação Y_{ij}.

ii. $\epsilon_{ij} \cap \mathcal{N}(0, \sigma^2), \ \forall\, i, j.$

iii. $\{\epsilon_{ij}\}_{i,j}$ constitui um conjunto de variáveis aleatórias independentes.

(b) Comecemos pelo cálculo das Somas de Quadrados. Tendo em conta as fórmulas vistas nas aulas teóricas e os valores dados no enunciado, temos:

$$SQRE = (n_c - 1) \sum_{i=1}^{6} s_i^2 = 10 \times (48.1636 + ... + 454.1161) = 18326.71 \ ;$$

$$\begin{aligned} SQF &= n_c \sum_{i=1}^{6} (\overline{y}_{i.} - \overline{y}_{..})^2 = 11 * ((44.19 - 53.25667)^2 + ... + (42.11 - 53.25667)^2) \\ &= 4068.939 \ , \end{aligned}$$

sendo necessário, para obter SQF, calcular primeiro a média geral da totalidade das $n = 66$ observações, que (uma vez que o delineamento é equilibrado) é a média simples das $k = 6$ médias regionais: $\overline{y}_{..} = (44.19 + 58.87 + ... + 42.11)/6 = 53.25667$. Logo, tem-se a seguinte tabela-resumo:

Fonte	g.l.	SQ	QM	F_{calc}
Factor	$k-1 = 5$	$SQF = 4068.939$	$QMF = \frac{SQF}{k-1} = 813.7878$	$\frac{QMF}{QMRE} = 2.6643$
Resíduos	$n-k = 60$	$SQRE = 18326.71$	$QMRE = \frac{SQRE}{n-k} = 305.4451$	

(c) Neste caso, e uma vez que não são conhecidas as variáveis originais, apenas é possível calcular a variância da totalidade das $n = 66$ observações recorrendo à decomposição da Soma de Quadrados Total correspondente a esta ANOVA:

$$s_y^2 = \frac{SQT}{n-1} = \frac{SQF + SQRE}{n-1} = \frac{4068.939 + 18326.71}{65} = \frac{22395.65}{65} = 344.55 \ .$$

Repare-se que este valor *não é* a média das variâncias amostrais de nível.

(d) Embora se possa escrever as hipóteses do teste com base nos efeitos α_i do factor (como se fez no exercício anterior), nas ANOVAs a um único factor é equivalente formular as hipóteses em termos das médias populacionais (valores esperados das observações $E[Y_{ij}] = \mu_i = \mu_1 + \alpha_i$) em cada nível do factor. Eis o teste com $\alpha = 0.05$:

Hipóteses: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ vs. $H_1 : \exists\, i, i'$ tal que $\mu_i \neq \mu_{i'}$.

Estatística do teste: $F = \frac{QMF}{QMRE} \ \cap \ F_{(k-1\,,\,n-k)}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(5,60)} = 2.37$.

Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 2.664$. É um valor significativo ao nível $\alpha = 0.05$ e rejeita-se H_0 a favor da hipótese de que existem efeitos do factor, ou seja, de que a percentagem média dos pesos de grãos defeituosos não é igual em todas as regiões.

No caso de se utilizar um nível de significância $\alpha = 0.01$, apenas muda a fronteira da região crítica, que passa a ser $f_{0.01(5,60)} = 3.34$. Assim, a estatística calculada (que continua a ser $F_{calc} = 2.664$) já não é significativa a este novo nível de significância, não sendo agora possível rejeitar a hipótese de iguais médias populacionais nas seis regiões.
O valor de prova associado à estatística calculada é (tendo em conta a natureza unilateral direita do teste) $P[F_{(5,60)} > F_{calc}] = P[F_{(5,60)} > 2.664]$. Não é possível obter este valor nas tabelas (embora já saibamos que ele se encontra entre 5% e 1%, uma vez que a conclusão dos testes muda para esses níveis de significância), mas pode calcular-se essa probabilidade com o auxílio do **R**:

```
> 1-pf(2.664, 5, 60)
[1] 0.03063001
```

Assim, tem-se $p = 0.03063$.

(e) **[Material Complementar]** Sabemos que duas médias de nível μ_i e $\mu_{i'}$ devem ser consideradas diferentes caso as respectivas médias amostrais difiram (em módulo) por mais do que o termo de comparação $q_{\alpha(k,n-k)}\sqrt{\frac{QMRE}{n_c}}$, onde $q_{\alpha(k,n-k)}$ corresponde ao valor que deixa à sua direita uma região de probabilidade α numa distribuição de Tukey de parâmetros k e $n-k$, e n_c indica o número comum de observações em cada nível do factor (o resultado que sustenta o teste de Tukey parte do pressuposto que o delineamento é equilibrado). No nosso caso tem-se $k = 6$ e $n = 66$. Trabalhando (como pedido no enunciado) com $\alpha = 0.05$, e recorrendo às tabelas da distribuição de Tukey (tabelas específicas, disponíveis na página *web* da disciplina), tem-se $q_{0.05(6,60)} = 4.16$. Um valor mais preciso pode ser obtido através do comando `qtukey` do **R**:

```
> qtukey(0.95, 6, 60)
[1] 4.163161
```

Sabemos pela alínea (b) que $QMRE = 305.4451$ e também que $n_c = 11$. Logo, o termo de comparação é dado por $q_{\alpha(k,n-k)}\sqrt{\frac{QMRE}{n_c}} = 4.16 \times \sqrt{\frac{305.4451}{11}} = 21.9212$. Trata-se dum valor elevado e por uma inspecção simples das médias amostrais de nível vemos que a maior diferença entre médias amostrais de nível é a diferença entre a média do Libolo e de Amboim: $|61.96 - 42.11| = 19.85 < 21.9212$. Assim, nenhum par de médias tem diferença maior que o termo de comparação, pelo que se admite a igualdade de todos os pares de médias (logo, a igualdade de todas as médias). Este resultado é contraditório com o resultado do teste F ao nível $\alpha = 0.05$, o que pode acontecer quando se usam duas ferramentes baseadas em teoria diferente (testes F e de Tukey). No entanto, as várias conclusões desses testes estão próximas da fronteira, pelo que a discrepância de resultados não é assim tão surpreendente.

3. A variável resposta Y é, neste caso, a variação de massa (coluna `variacao.massa` na *data frame*). Existem ao todo $n = 50$ observações.

   (a) Para estudar este problema através duma ANOVA, ignora-se os valores numéricos das concentrações de dióxido de carbono, tratando cada diferente concentração apenas como um diferente tratamento. Assim, o factor CO_2 terá $k = 5$ níveis, havendo ($n_i = 10 = n_c$) observações para cada concentração de CO_2 (nível do factor). O modelo ANOVA associado a este delineamento é o seguinte:

      i. $Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij}$, $\forall\, i = 1, 2, 3, 4, 5$, $j = 1, 2, ..., 10$, com $\alpha_1 = 0$, onde
         - Y_{ij} indica a variação de massa para a j-ésima repetição associada à i-ésima concentração de CO_2;

- μ_1 indica o variação de massa média (populacional) na ausência de CO_2 ($i = 1$);
- α_i indica o efeito (acréscimo em relação à média populacional do primeiro nível) da i-ésima concentração de dióxido de carbono, isto é, $\alpha_i = \mu_i - \mu_1$; e
- ϵ_{ij} indica o erro aleatório associado à observação Y_{ij}.

ii. $\epsilon_{ij} \cap \mathcal{N}(0, \sigma^2), \ \forall\, i, j.$

iii. $\{\epsilon_{ij}\}_{i,j}$ constitui um conjunto de variáveis aleatórias independentes.

(b) Vamos construir a tabela-resumo da ANOVA com o auxílio do R, uma vez que os dados estão disponíveis na *data frame* `CO2`, com os valores da variável resposta na coluna `variacao.massa` e os diferentes níveis de CO_2 no factor `CO2.factor` (alternativamente, podem sempre usar-se as fórmulas disponíveis no formulário para SQF e $SQRE$ em delineamentos a um factor, sabendo-se também que os graus de liberdade associados ao Factor são $k - 1 = 4$ e os residuais $n - k = 45$):

```
> summary(aov(variacao.massa ~ CO2.factor, data=CO2))
            Df Sum Sq Mean Sq F value Pr(>F)
CO2.factor   4  11274  2818.6   101.6 <2e-16 ***
Residuals   45   1248    27.7
```

O teste F desta ANOVA diz respeito à possível existência de efeitos do Factor, ou seja,

Hipóteses: $H_0 : \alpha_i = 0 \ , \ \forall \ i = 2, 3, 4, 5$ vs. $H_1 : \exists\, i = 2, 3, 4, 5$ tal que $\alpha_i \neq 0$.

Estatística do teste: $F = \frac{QMF}{QMRE} \ \cap \ F_{(k-1\,,\,n-k)}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(4,45)} \ \approx \ 2.58$.

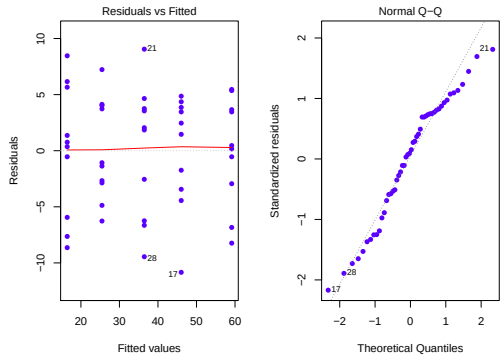
Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 101.6$. É um valor claramente significativo ao nível $\alpha = 0.05$ e rejeita-se H_0 a favor da hipótese de que existem efeitos do Factor, ou seja, que as concentrações de CO_2 estão associadas a diferentes variações médias na massa das culturas do *Pseudomonas fragi*.

(c) Como em qualquer modelo linear, o resíduo é a diferença entre cada valor observado da variável resposta e o correspondente valor ajustado pelo modelo, ou seja, e usando a notação da ANOVA a 1 Factor, $e_{ij} = y_{ij} - \hat{y}_{ij}$. Sabe-se que, num modelo ANOVA a um factor, o valor ajustado duma dada observação corresponde à média amostral das observações no mesmo nível do factor: $\hat{y}_{ij} = \overline{y}_{i.}$. Assim, todas as observações do primeiro grupo têm valor ajustado igual a $\hat{y}_{1j} = \overline{y}_{1.} = 59.14$. O resíduo da primeira observação do primeiro grupo será $e_{11} = 62.6 - 59.14 = 3.46$ e o da segunda observação desse grupo é $e_{12} = 59.6 - 59.14 = 0.46$. De forma análoga, os valores ajustados de qualquer observação no segundo grupo são dados por $\hat{y}_{2j} = \overline{y}_{2.} = 46.04$. O resíduo da terceira observação do segundo grupo é assim $e_{23} = y_{23} - \overline{y}_{2.} = 47.5 - 46.04 = 1.46$. Para calcular a totalidade dos resíduos podemos recorrer ao R (arredondando a três casas decimais):

```
> round(residuals(CO2.aov), d=3)
     1      2      3      4      5      6      7      8      9     10     11     12     13
  3.46   0.46   5.36   0.16  -0.54   5.46  -8.24  -2.94  -6.84   3.66   4.86  -1.74   1.46
    14     15     16     17     18     19     20     21     22     23     24     25     26
  3.46   2.46   4.36 -10.84   3.86  -3.44  -4.44   9.05   4.65  -6.65   1.85   3.75   2.05
    27     28     29     30     31     32     33     34     35     36     37     38     39
 -6.25  -9.45   3.55  -2.55   4.03  -2.67  -6.27  -4.87   3.73  -1.37  -2.87   7.23  -1.07
    40     41     42     43     44     45     46     47     48     49     50
  4.13   8.46   0.76  -8.64  -5.94   1.36   5.66   6.16   0.36  -0.54  -7.64
```

Com o auxílio do R, podemos obter dois dos gráficos de resíduos já considerados no estudo dos modelos de Regressão Linear, através do comando:

```
> plot(CO2.aov, which=c(1,2), pch=16, col="blue")
```



O gráfico da esquerda é o gráfico de resíduos usuais (no eixo vertical) vs. valores ajustados da variável resposta (eixo horizontal). O facto de os resíduos surgirem "empilhados" em colunas é característico numa ANOVA a um factor e resulta do já referido facto de todas as observações dum dado nível terem o mesmo valor ajustado $\hat{y}_{ij} = \overline{y}_{i.}$, logo, a mesma coordenada no eixo horizontal. Neste caso, observam-se $k = 5$ colunas. Não parece existir problema com a hipótese de homogeneidade das variâncias, uma vez que a variabilidade dos resíduos não parece diferir muito nos cinco níveis do factor.

O *qq-plot* (gráfico à direita) não indicia problemas graves com a Normalidade, dada a disposição aproximadamente linear dos pontos.

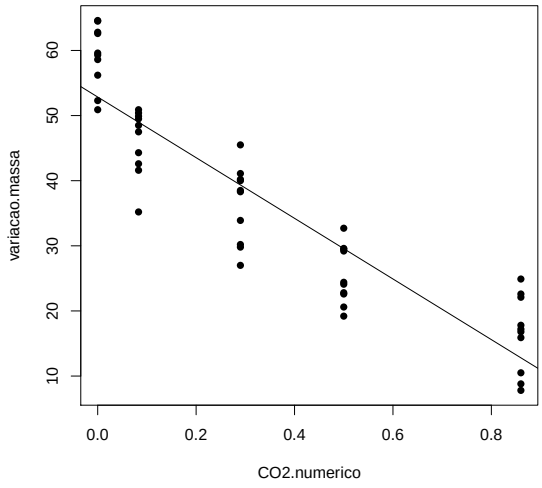
Os restantes diagnósticos que foram considerados aquando do estudo da regressão (distâncias de Cook, efeito alavanca) são geralmente de menor utilidade no contexto duma ANOVA. Em relação às distâncias de Cook, por exemplo, sabe-se de antemão qual o efeito de retirar uma observação: além de desequilibrar um delineamento equilibrado, afectará a média das observações no mesmo nível do factor (ou seja, os valores ajustados $\hat{y}$ nesse nível). Assim valores elevados da distância de Cook correspondem a observações atípicas (*outliers*) no seio dum dado nível. Mas para identificar tais observações, basta o gráfico usual de resíduos contra $\hat{y}$, não sendo necessário um diagnóstico específico. Em relação aos efeitos alavanca, é possível mostrar que o efeito alavanca de qualquer observação y_{ij} numa ANOVA a um factor é dada por $\frac{1}{n_i}$, onde n_i indica o número de observações no nível i da observação. Em delineamentos equilibrados, esse valor é igual para todas as observações (no nosso caso, todas teriam efeito alavanca igual a $\frac{1}{10}$). O gráfico obtido no `R` com a opção `which=5` tinha, na regressão linear, os valores do efeito alavanca (h_{ii}, ou *leverages*) de cada observação no eixo horizontal. No entanto, para ANOVAs com delineamentos equilibrados a um factor, o R substitui esse eixo por uma simples indicação dos diferentes níveis do factor (ordenados por ordem crescente das médias $\overline{y}_{i.}$), uma vez que um gráfico análogo ao construído na regressão linear apenas empilharia todos os resíduos numa única coluna. O gráfico alternativo produzido pelo R quando os delineamentos são equilibrados fica assim semelhante ao primeiro gráfico de resíduos, embora sem qualquer efeito de escala no eixo horizontal e com os resíduos (internamente) estandardizados no eixo vertical, em vez dos resíduos usuais.

(d) Nesta alínea pede-se para aproveitar os valores das concentrações de CO_2 utilizadas, e tratar essa variável preditora como uma variável numérica, estudando a regressão linear simples de `variacao.massa` sobre `CO2.numerico`.

i. O gráfico pedido pode ser construído com o seguinte comando do `R`.

```
> plot(variacao.massa ~ CO2.numerico, data=CO2, pch=16)
```

O resultado (já com a recta que é pedida na alínea seguinte e que foi traçada com o comando `abline(CO2.lm)`) é:



ii. A regressão linear pedida é dada por:

```
> CO2.lm <- lm(variacao.massa ~ CO2.numerico, data=CO2)
> summary(CO2.lm)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    52.849      1.408   37.52   <2e-16 ***
CO2.numerico  -46.569      3.030  -15.37   <2e-16 ***
---
Residual standard error: 6.637 on 48 degrees of freedom
Multiple R-squared: 0.8312,Adjusted R-squared: 0.8276
F-statistic: 236.3 on 1 and 48 DF,  p-value: < 2.2e-16
```

Apesar de alguma tendência para uma relação curvilinear, uma regressão linear simples pode constituir uma modelação aproximada da relação entre concentrações de dióxido de carbono e variação na massa das culturas de *Pseudomonas fragi* (repare-se como seria impossível tirar esta ilação se o número de níveis fosse mais pequeno, *e.g.*, $k = 3$). O valor do coeficiente de determinação é claramente significativo ($p < 2.2 \times 10^{-16}$) e bastante elevado ($R^2 = 0.8312$), explicando mais de 83% da variabilidade total observada na variável resposta.

iii. Os testes F de ajustamento global do contexto regressão linear simples e do contexto ANOVA a um factor, não são os mesmos. Como se viu nas aulas teóricas, a ANOVA a um factor pode ser vista como uma espécie de regressão linear múltipla em que as variáveis preditoras são as indicatrizes dos níveis (excepto o primeiro) do factor. Assim, a informação disponível para prever os valores da variável resposta é, no caso da regressão considerada nesta alínea, a variável `CO2.numerico`, com valores numéricos diferentes em cada nível (mas repetidos para as observações dum mesmo nível). No caso da ANOVA a um factor, é o conjunto das indicatrizes de nível e o vector dos n uns. Sendo diferente a informação preditora, serão diferentes os valores ajustados e os valores dos respectivos F_{calc} e coeficientes de determinação. Em relação a este último, e embora não seja hábito utilizá-lo no contexto duma ANOVA a um factor, o seu valor é aqui $R^2 = 0.9003$, superior ao que se obteve na regressão ($R^2 = 0.8312$), como se pode

constatar através do ajustamento obtido utilizando simultaneamente o comando `lm` e o factor preditor `CO2.factor`:

```
> summary(lm(variacao.massa ~ CO2.factor, data=CO2))
(...)
Residual standard error: 5.266 on 45 degrees of freedom
Multiple R-squared: 0.9003,Adjusted R-squared: 0.8915
F-statistic: 101.6 on 4 and 45 DF,  p-value: < 2.2e-16
```

Repare-se como o valor da estatística calculada, $F_{calc} = 101.6$, é o que foi obtido usando o comando `aov`.

Um comentário final: o modelo ANOVA não permite, ao contrário da regressão, fazer previsões sobre as variações de massa com concentrações de CO_2 não observadas na experiência, uma vez que os níveis de um factor não têm escala (são apenas categorias diferentes).

4. (a) A descrição da experiência corresponde a um delineamento factorial a dois factores, sendo o primeiro factor constituído pelas fases do processamento e o segundo factor constituido pelos diferentes lotes. Refira-se que na descrição da experiência dada nesta alínea, cada nível do segundo factor constitui aquilo a que, na tradição da Análise de Variância, se designa por *bloco*. Esta designação surge historicamente associada a factores cuja inclusão na experiência resulta, não tanto de se pretender estudar directamente o seu efeito sobre a variável resposta, mas sobretudo de saber que constituem uma fonte de heterogeneidade das unidades experimentais, associada a variabilidade na variável resposta. Pretende-se incorporar essa heterogeneidade no modelo, controlando-a e podendo assim filtrar a variabilidade nos valores da variável resposta que lhe está associada. Neste caso, é natural supôr que a diferentes lotes de feijão correspondam diferentes concentrações de zinco, independentemente de qualquer tratamento a que sejam submetidos[1].

   A *data frame* `zinco` tem três colunas: a variável resposta (`concentracao`), o factor com $a = 4$ níveis, cujos efeitos se pretende realmente estudar (`fase`) e o factor/bloco (`lote`), com $b = 9$ níveis, introduzido para controlar a heterogeneidade das unidades experimentais (lotes de feijão). Nas 36 células deste delineamento não há repetições de observações (ou seja, $n_c = 1$). Logo, independentemente de ser desejável, não é possível incluir efeitos de interacção no modelo. Utilizar-se-á um modelo a dois factores, sem interacção:

   i. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \epsilon_{ijk},\ \forall\, i = 1, 2, 3, 4\ ,\ j = 1, 2, ..., 9\ ,\ k = 1$ (o índice k é dispensável porque não há repetições nas células), com $\alpha_1 = 0$ e $\beta_1 = 0$, e onde
      - Y_{ijk} indica a concentração de zinco da fase i, associada ao lote de feijão j;
      - μ_{11} é a concentração esperada de zinco no início do processamento, para o lote 1;
      - α_i indica o efeito da fase i;
      - β_j indica o efeito do lote j; e
      - ϵ_{ijk} indica o erro aleatório associado à observação Y_{ijk}.

   ii. $\epsilon_{ijk} \cap \mathcal{N}(0, \sigma^2)\ ,\ \forall\, i, j, k.$

[1] Seria mais adequado supôr que ao factor `lotes` correspondem *efeitos aleatórios*, expressão usada para designar o contexto em que os níveis do factor analisados não são os únicos de interesse, mas apenas uma amostra aleatória dum número muito maior de níveis. Neste caso, não é de crer que haja interesse em estudar apenas *aqueles* nove lotes usados na experiência. Mais realista será supôr que constituem uma amostra aleatória duma infinidade de potenciais lotes de feijão. Assim, seria mais adequado associar efeitos aleatórios aos lotes, continuando a associar efeitos fixos às fases do processamento (aqui sim, existe real interesse em estudar *aqueles* quatro momentos do processamento). Um modelo onde se misturam efeitos fixos e efeitos aleatórios é conhecido por *modelo misto* e será abordado mais tarde, nesta disciplina.

iii. $\{\epsilon_{ijk}\}_{i,j,k}$ constituem um conjunto de variáveis aleatórias independentes.

(b) Recorrendo ao R, obtém-se a tabela-resumo correspondente a este modelo:

```
> summary(aov(concentracao ~ fase + lote, data=zinco))
            Df Sum Sq Mean Sq F value   Pr(>F)
fase         3  20.60   6.866   9.736 0.000218 ***
lote         8  17.76   2.220   3.148 0.013931 *
Residuals   24  16.92   0.705
```

Repare-se que (em comparação com a tabela do modelo a um factor) existe uma nova linha na tabela, correspondente ao novo factor. Os graus de liberdade associados a cada factor são o número de níveis desse factor, menos 1 (como reflexo da imposição das restrições $\alpha_1 = 0$ e $\beta_1 = 0$), o que neste caso significa $a-1 = 3$ e $b-1 = 8$ graus de liberdade. Os graus de liberdade associados ao residual são, como de costume, o número de observações menos o número de parâmetros no modelo, ou seja, $n-(a+b-1) = 36-(4+9-1) = 24$. Uma vez que o delineamento é equilibrado, com uma única repetição por célula ($n_c = 1$) é possível utilizar as fórmulas constantes dos acetatos das aulas teóricas (e também do formulário, uma vez que as expressões para SQA e SQB são iguais às do modelo *com* interacção, no caso de delineamentos equilibrados) para calcular as restantes quantidades da tabela. Para tal, será útil dispor das concentrações médias em cada fase e de cada lote:

```
> model.tables(aov(concentracao ~  fase + lote, data=zinco), type="means")
Tables of means
Grand mean
2.847778
 fase
fase
    1     2     3     4
2.228 2.847 2.233 4.083
 lote
lote
    1     2     3     4     5     6     7     8     9
3.483 3.733 3.558 2.998 3.425 1.940 1.858 2.195 2.443
```

Assim, e como $n_c = 1$, temos: $SQA = b\,n_c \sum_{i=1}^{a} (\overline{y}_{i..} - \overline{y}_{...})^2 = 9 \times ((2.228-2.847778)^2 + (2.847-2.847778)^2 + (2.233-2.847778)^2 + (4.083-2.847778)^2) = 20.59066$, e $SQB = a\,n_c \sum_{j=1}^{b} (\overline{y}_{.j.} - \overline{y}_{...})^2 = 4 \times ((3.483-2.847778)^2 + (3.733-2.847778)^2 + ... + (2.443-2.847778)^2) = 17.76391$. Para obter a Soma de Quadrados residual, basta recordar que a Soma de Quadrados Total é o numerador da variância de todas as $n = 36$ observações. Sabendo que esta variância é:

```
> var(zinco$concentracao)
[1] 1.579458
```

pode-se deduzir que $SQT = (n-1)\,s_y^2 = 35 \times 1.579458 = 55.28102$. Logo, $SQRE = SQT - (SQA + SQB) = 55.28102 - (20.59066 + 17.76391) = 16.92645$. Os restantes valores da tabela resultam da aplicação directa das suas definições.

(c) Nesta fase apenas é pedido o teste à existência de efeitos do factor A (fases do processamento). Este teste F é indicado de seguida.

Hipóteses: $H_0 : \alpha_i = 0\ ,\ \forall\ i = 2,3,4$ vs. $H_1 : \exists\, i = 2,3,4$ tal que $\alpha_i \neq 0$.
Estatística do teste: $F = \frac{QMA}{QMRE} \cap F_{(a-1\,,\,n-(a+b-1))}$, sob H_0.
Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(3,24)} = 3.01$.

Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 9.736$. É um valor significativo ao nível $\alpha = 0.05$ e rejeita-se H_0 a favor da hipótese de que existem efeitos do Factor, ou seja, que as diferentes fases do processamento têm efeito sobre as concentrações médias de zinco.

(d) Nesta alínea, diz-se que foi ajustado um modelo apenas a um factor, o factor fases de processamento, ignorando a existência do factor (blocos) lote. O resultado obtido será:

```
> summary(aov(concentracao ~ fase , data=zinco))
            Df Sum Sq Mean Sq F value Pr(>F)
fase         3  20.60   6.866   6.334 0.0017 **
Residuals   32  34.68   1.084
```

Registem-se os seguintes factos, relativos à comparação desta tabela-resumo e da tabela-resumo do modelo a dois factores, sem interacção, ajustado nas alíneas anteriores:

- Existe uma linha comum nas duas tabelas, correspondente ao factor `fase`, e os graus de liberdade, Soma de Quadrados e Quadrado Médio do factor `fase` são idênticos aos da tabela-resumo do modelo a dois factores.
- Uma vez que a Soma de Quadrados Total é igual nos dois casos (já que $SQT = (n-1)\, s_y^2 = 35 \times 1.5795 = 55.28$ não depende do modelo ajustado) este facto tem de significar que a Soma de Quadrados Residual é aqui a soma das parcelas SQB e $SQRE$ do modelo a dois factores sem interacção. De facto, verifica-se que $SQRE_A = 34.68 = 17.76 + 16.92 = SQB + SQRE_{A+B}$. Ou seja, a não existência neste modelo de efeitos do factor B implica que a variabilidade que lhe poderia ser imputada (SQB) vai acabar por ser variabilidade residual, isto é, vai contribuir para aumentar o valor de $SQRE_A$. Neste exemplo, ao factor `lote` corresponde cerca de metade da variabilidade que é considerada residual (não explicada pelo modelo) no modelo apenas com o factor `fase`.
- Mas os graus de liberdade associados ao residual também são diferentes nos dois casos. E, mais uma vez, os graus de liberdade associados ao residual, neste modelo a um só factor, correspondem à soma dos graus de liberdade residuais e associados ao outro factor, no modelo a dois factores: $32 = 8 + 24$. Isto não acontece por acaso. Também no caso dos graus de liberdade dos modelos lineares, a soma de todas as parcelas é constante (e igual a $n-1$). Logo, a não existência, no modelo ajustado nesta alínea, de efeitos do factor `lote` significa que os graus de liberdade residuais (tal como a soma de quadrados residual) também aumentam.
- Na estatística F aos efeitos do factor `fase`, o numerador QMF (QMA, na notação para modelos a dois factores) fica igual, enquanto que o denominador $QMRE$ sofre uma dupla transformação: o seu numerador $SQRE$ é maior do que no modelo a dois factores (pois $SQRE_A = SQRE_{A+B} + SQB$), mas também o seu denominador é maior (pois $g.l.(SQRE_{A+B}) = n-(a+b-1) < n-a = g.l.(SQRE_A)$). Assim, se a estatística F é maior, ou menor, dependerá da dimensão relativa destes aumentos do numerador e denominador.
- No exemplo em questão, o $QMRE$ do modelo com dois factores é mais baixo: 0.7052 (em vez de 1.0839 no modelo só com o factor `fase`). A estatística F no teste aos efeitos do factor `fase` (que, recorde-se, continua a ter o mesmo numerador) era $F_A = 9.7361$ no modelo a dois factores e no modelo a um factor é agora $F = 6.3343$). A rejeição da hipótese de inexistência de efeitos do Factor *fase* ($H_0 : \alpha_i = 0, \forall\, i$) era mais clara no

modelo a dois factores, e embora neste caso não se altere qualitativamente a conclusão para os níveis de significância usuais, poderia dar-se esse caso.

- Caso existam realmente efeitos do novo factor, a Soma de Quadrados Residual do modelo a dois factores sem interacção, $SQRE_{A+B}$, será bastante inferior à do modelo a um factor e também $QMRE_{A+B}$ será menor, pelo que aumenta a estatística F, que tende assim a ser mais significativa. Pelo contrário, se a parcela SQB fôr relativamente pequena, pode acontecer a situação contrária, e a estatística F tornar-se menor, afastando-se assim das regiões críticas.

Conclusão: caso existam realmente efeitos dum factor adicional, que torna as unidades experimentais muito heterogeneas, a inclusão desse factor no delineamento e no modelo ANOVA contribuirá para evidenciar eventuais efeitos do outro factor, que realmente se pretende estudar. Mas no caso de ao factor adicional não corresponderem realmente efeitos importantes, a sua inclusão no delineamento e no modelo poderá até contribuir para camuflar eventuais efeitos do factor no qual estamos realmente interessados.

5. Trata-se dum delineamento factorial a dois factores (`terreno` e `variedade`), mas com uma única observação em cada célula (em cada terreno, apenas há uma parcela com cada variedade). Logo, só é possível ajustar um modelo a dois factores sem interacção, tal como no exercício 4.

   (a) A tabela-resumo correspondente é:

```
> summary(aov(rend ~ variedade + terreno, data=terrenos))
            Df Sum Sq Mean Sq F value  Pr(>F)
variedade    3  1.799  0.5997   6.145 0.00175 **
terreno     12  2.407  0.2006   2.056 0.04737 *
Residuals   36  3.513  0.0976
```

   Desta tabela depreende-se que, aos níveis de significância usuais, deve considerar-se a existência de efeitos do factor variedade:

   Hipóteses: $H_0 : \alpha_i = 0 \ , \ \forall \ i = 2,3,4$ vs. $H_1 : \exists\, i = 2,3,4$ tal que $\alpha_i \neq 0$.
   Estatística do teste: $F = \frac{QMA}{QMRE} \ \cap \ F_{(a-1\,,\,n-(a+b-1))}$, sob H_0.
   Nível de significância: $\alpha = 0.05$.
   Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(3,36)} \ \approx \ 2.87$.
   Conclusões: $F_{calc} = 6.145$, um valor significativo mesmo ao nível $\alpha = 0.005$. Logo, rejeita-se H_0 a favor da hipótese de que existem efeitos do factor. Assim, é de concluir que diferentes variedades estejam associadas a diferentes rendimentos médios.

   (b) Um teste aos efeitos do factor `terreno` permite tirar a conclusão que os efeitos deste factor são menos importantes que os efeitos do factor `variedade`, embora ao nível de significância $\alpha = 0.05$ sejam (por pouco) significativos. Assim,

   Hipóteses: $H_0 : \beta_j = 0 \ , \ \forall \ j = 2,..,13$ vs. $H_1 : \exists\, j = 2,...,13$ tal que $\beta_j \neq 0$.
   Estatística do teste: $F = \frac{QMB}{QMRE} \ \cap \ F_{(b-1\,,\,n-(a+b-1))}$, sob H_0.
   Nível de significância: $\alpha = 0.05$.
   Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(12,36)} \ \approx \ 2.04$.
   Conclusões: $F_{calc} = 2.056$, um valor significativo (por muito pouco) ao nível $\alpha = 0.05$. Logo, rejeita-se H_0 a favor da hipótese de que existem efeitos do factor `terreno`.

   NOTA: Num caso como este, em que a conclusão dependende do nível de significância usado, é especialmente importante que eventuais fontes de variabilidade, exteriores ao factor

sob estudo, mas que afectem a variável resposta, sejam tidas em conta, de forma a reduzir a variabilidade não explicada pelo modelo, isto é, o valor de $QMRE$.

6. FALTA

7. Trata-se dum delineamento factorial a dois factores, o factor A (Fósforo), com $a = 3$ níveis (Baixa, Média e Elevada dosagem de adubação) e o Factor B (Potássio), igualmente com $b = 3$ níveis (Baixa, Média e Elevada dosagem de adubação). O delineamento é equilibrado, uma vez que em cada uma das $ab = 9$ situações experimentais (células) há igual número de observações $n_{ij} = n_c = 3$. Havendo repetições nas células, é possível estudar o modelo ANOVA a 2 factores, com interacção. A equação de base deste modelo é $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, $\forall\, i = 1, 2, 3$, $j = 1, 2, 3$, $k = 1, 2, 3$, onde Y_{ijk} indica o rendimento obtido na k-ésima repetição da adubação correspondente à célula que cruza o nível i do fósforo e o nível j do potássio. Impõem-se as restrições $\alpha_1 = 0$, $\beta_1 = 0$, $(\alpha\beta)_{1j} = 0$ para qualquer j, e $(\alpha\beta)_{i1} = 0$ para qualquer i.

   (a) A tabela-resumo é dada no enunciado, mas com seis valores omissos. Os graus de liberdade do factor A (fósforo) são $a-1 = 2$. Os graus de liberdade associados aos efeitos de interacção são $(a-1)(b-1) = 4$. O Quadrado Médio associado ao factor B (potássio) é $QMB = \frac{SQB}{b-1} = \frac{18.7563}{2} = 9.37815$. O Quadrado Médio Residual é $QMRE = \frac{SQRE}{n-ab} = \frac{2.59333}{18} = 0.1440739$. O valor da estatística F para o teste aos efeitos principais do factor A é $F_A = \frac{QMA}{QMRE} = \frac{1.121481}{0.1440739} = 7.784068$. Finalmente, o valor da estatística F no teste aos efeitos principais do factor B é $F_B = \frac{QMB}{QMRE} = \frac{9.37815}{0.1440739} = 65.09264$.

   (b) Há três tipos de efeitos: principais do factor fósforo, associados às parcelas α_i; principais do factor potássio, associados às parcelas β_j; e de interacção entre os dois tipos de adubação, associados às parcelas $(\alpha\beta)_{ij}$. Existe um teste F para testar hipóteses associadas a cada um destes tipos de efeitos. Em concreto:

   Teste à interacção. As hipóteses são:

   $$H_0 : \quad (\alpha\beta)_{ij} = 0 \ , \ \forall\, i, j \qquad vs. \qquad H_1 : \quad \exists\, i, j \text{ tal que } \quad (\alpha\beta)_{ij} \neq 0 \ .$$

   Teste aos efeitos principais do factor A. As hipóteses são:

   $$H_0 : \quad \alpha_i = 0 \ , \ \forall\, i \qquad vs. \qquad H_1 : \quad \exists\, i \text{ tal que } \quad \alpha_i \neq 0 \ .$$

   Teste aos efeitos principais do factor B. As hipóteses são:

   $$H_0 : \quad \beta_j = 0 \ , \ \forall\, j \qquad vs. \qquad H_1 : \quad \exists\, j \text{ tal que } \quad \beta_j \neq 0 \ .$$

   Para cada um destes testes, as estatísticas F são definidas como $F = \frac{QMxx}{QMRE}$, onde $QMxx$ indica o quadrado médio associado ao respectivo tipo de efeitos. As distribuições destas estatísticas de teste, caso seja verdadeira cada uma das hipóteses nulas, são F com graus de liberdade dados pelos g.l. dos quadrados médios no numerador e denominador, respectivamente, da estatística correspondente. Todas as regiões críticas são unilaterais direitas. Assim, e tendo em conta os valores da tabela-resumo e utilizando o nível de significância $\alpha = 0.05$, tem-se que se rejeitam as hipóteses nulas dos três testes. De facto, rejeita-se a inexistência de efeitos de interacção, uma vez que $F_{AB_{calc}} = 3.36504 > f_{0.05(4,18)} = 2.927744$. Rejeita-se a inexistência de efeitos principais do factor fósforo uma vez que $F_{A_{calc}} = 7.784068 > f_{0.05(2,18)} = 3.554557$. Finalmente,

rejeita-se clarissimamente a inexistência de efeitos principais do factor potássio já que $F_{B_{calc}} = 65.09264 > f_{0.05(2,18)} = 3.554557$. Assim, conclui-se pela existência dos três tipos de efeitos. Estas conclusões poderiam também ser obtidas directamente a partir dos valores de prova (*p-values*) correspondentes às três estatísticas de teste, disponíveis no enunciado. O valor de prova mais elevado, no caso do teste aos efeitos de interacção ($p = 0.03187154$) indica que, ao nível de significância $\alpha = 0.01$, a conclusão já seria a não rejeição da hipótese nula, isto é, não seria possível concluir pela existência de efeitos de interacção. Já a existência de efeitos principais do factor potássio está associado a um *p-value* da ordem de 10^{-8}.

(c) Nesta alínea pede-se para considerar-se o modelo sem efeitos de interacção, ou seja, cuja equação de base é $Y_{ijk} = \mu_{11}+\alpha_i+\beta_j+\epsilon_{ijk},\ \forall\, i,j,k$, e com as restrições $\alpha_1 = \beta_1 = 0$. O facto de o modelo não prever efeitos de interacção significa que a respectiva Soma de Quadrados (indicada no enunciado) passa a englobar a Soma de Quadrados Residual (uma vez que já não corresponde a efeitos previstos pelo modelo). Tem-se agora $SQRE = 2.59333 + 1.93926 = 4.53259$. Os graus de liberdade sofrem uma transformação análoga (este modelo tem agora menos $(a-1)(b-1)$ parâmetros do que anterior, pelo que os graus de liberdade residuais aumentam nesse montante). Assim, $g.l.(SQRE) = 18 + 4 = 22$. Logo o novo Quadrado Médio Residual vem: $QMRE = \frac{4.53259}{22} = 0.2060268$. As somas de quadrados, graus de liberdade e quadrados médios associados aos efeitos principais de cada factor permanecem iguais (são calculados de forma análoga) pelo que a tabela-resumo é agora a seguinte:

variação	g.l.	SQs	QMs	F_{calc}
fosforo	2	2.24296	1.121481	5.443374
potassio	2	18.75630	9.37815	45.51908
residual	22	4.53259	0.2060268	–

Para identificar os valores de prova (*p-values*) dos novos valores das estatísticas F sobrantes, é necessário ter em conta os novos valores dos graus de liberdade residuais. Tem-se:

```
> 1-pf(5.443374, 2, 22)
[1] 0.01200658
> 1-pf(45.51908, 2, 22)
[1] 1.517658e-08
```

Assim, os dois valores calculados das estatísticas continuam a ser significativos ao nível $\alpha = 0.05$. No entanto, os efeitos do factor fósforo já não seriam considerados significativos ao nível $\alpha = 0.01$. Este exemplo ilustra o perigo de ignorar a existência de efeitos que realmente existam (neste caso, ignorar os efeitos de interacção): pode ajudar a camuflar a existência de outros tipos de efeitos, mesmo dos que são previstos no modelo, através do inflacionamento da variabilidade residual ($QMRE$).

8. (a) Trata-se dum delineamento factorial, a dois factores: tempo de exposição, com $a = 3$ níveis, e temperatura, também com $b = 3$ níveis. O delineamento é equilibrado, com $n_c = 3$ repetições em cada uma das $ab = 9$ células, para um total de $n = ab\,n_c = 27$ observações. Havendo repetições nas células, é possível ajustar um modelo a dois factores, com interacção, cuja equação de base é:

   i. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},\ \forall\, i = 1,2,3\ ,\ j = 1,2,3,\ k = 1,2,3,$
   com $\alpha_1 = 0$, $\beta_1 = 0$, $(\alpha\beta)_{1j} = 0$ para qualquer j, e $(\alpha\beta)_{i1} = 0$ para qualquer i, onde
      - Y_{ijk} indica a absorção na k-ésima repetição da situação experimental dada pela combinação do tempo de exposição i com a temperatura j.

- μ_{11} indica a absorção média (populacional) na célula definida pelo tempo de exposição E_1 com a temperatura T_1;
- α_i indica o efeito principal do tempo de exposição i;
- β_j indica o efeito principal da temperatura j;
- $(\alpha\beta)_{ij}$ indica o efeito de interacção entre o tempo de exposição i e a temperatura j; e
- ϵ_{ijk} indica o erro aleatório associado à observação Y_{ijk}.

ii. $\epsilon_{ijk} \cap \mathcal{N}(0, \sigma^2), \ \forall\, i, j, k.$

iii. $\{\epsilon_{ijk}\}_{i,j,k}$ constitui um conjunto de variáveis aleatórias independentes.

(b) Há três tipos de efeitos: principais do factor A, associados às parcelas α_i; principais do factor B, associados às parcelas β_j; e de interacção, associados às parcelas $(\alpha\beta)_{ij}$. Existe um teste F para testar hipóteses associadas a cada um destes tipos de efeitos. Em concreto:

Teste à interacção. As hipóteses são:

$$H_0: \quad (\alpha\beta)_{ij} = 0 \ , \ \forall\, i, j \qquad vs. \qquad H_1: \quad \exists\, i, j \ tal \ que \quad (\alpha\beta)_{ij} \neq 0 \ .$$

Teste aos efeitos principais do factor A. As hipóteses são:

$$H_0: \quad \alpha_i = 0 \ , \ \forall\, i \qquad vs. \qquad H_1: \quad \exists\, i \ tal \ que \quad \alpha_i \neq 0 \ .$$

Teste aos efeitos principais do factor B. As hipóteses são:

$$H_0: \quad \beta_j = 0 \ , \ \forall\, j \qquad vs. \qquad H_1: \quad \exists\, j \ tal \ que \quad \beta_j \neq 0 \ .$$

(c) Para efectuar os três testes F pedidos, vamos construir a tabela-resumo da ANOVA, utilizando para o efeito os comandos do `R`:

```
> absorcao.aov <- aov(abs ~ exposicao * temperatura, data=absorcao)
> summary(absorcao.aov)
                     Df Sum Sq Mean Sq F value   Pr(>F)
exposicao             2   2113  1056.3   63.59 6.92e-09 ***
temperatura           2   3674  1836.8  110.58 7.75e-11 ***
exposicao:temperatura 4   2704   676.1   40.70 8.74e-09 ***
Residuals            18    299    16.6
```

Repare-se na utilização do asterisco para indicar, na fórmula que define o modelo usado, que se pretende utilizar um modelo a dois factores *com* interacção.

Teste à interacção .

Hipóteses: $H_0: (\alpha\beta)_{ij} = 0 \ , \ \forall \ i = 2, 3$ e $j = 2, 3$ [não há interacção]
vs. $H_1: \exists\, i = 2, 3 \ , \ j = 2, 3$ tais que $(\alpha\beta)_{ij} \neq 0$ [há interacção].

Estatística do teste: $F = \frac{QMAB}{QMRE} \ \cap \ F_{[(a-1)(b-1)\,,\,n-ab]}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(4,18)} \ = \ 2.93$.

Conclusões: O valor da estatística do teste é $F_{calc} = 40.70$. É um valor claramente significativo ao nível $\alpha = 0.05$, rejeitando-se H_0 a favor da hipótese alternativa de que existem efeitos de interacção entre tempo de exposição e temperatura.

Teste aos efeitos principais do factor A .

Hipóteses: $H_0 : \alpha_i = 0 \ , \ \forall \ i = 2,3$ [não há efeitos principais de A] vs. $H_1 : (\alpha_2 \neq 0) \vee (\alpha_3 \neq 0)$ [há efeitos principais de A].

Estatística do teste: $F = \frac{QMA}{QMRE} \ \cap \ F_{[a-1\,,\,n-ab]}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(2,18)} \ = \ 3.55$.

Conclusões: O valor da estatística do teste é $F_{calc} = 63.59$. É um valor claramente significativo ao nível $\alpha = 0.05$, rejeitando-se H_0 a favor da hipótese alternativa de que existem efeitos principais de tempo de exposição.

Teste aos efeitos principais do factor B .

Hipóteses: $H_0 : \beta_j = 0 \ , \ \forall \ i = 2,3$ [não há efeitos principais de B] vs. $H_1 : (\beta_2 \neq 0) \vee (\beta_3 \neq 0)$ [há efeitos principais de B].

Estatística do teste: $F = \frac{QMB}{QMRE} \ \cap \ F_{[b-1\,,\,n-ab]}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(2,18)} \ = \ 3.55$.

Conclusões: O valor da estatística do teste é $F_{calc} \ = \ 110.58$. É um valor ainda mais claramente significativo do que o da estatística calculada no teste aos efeitos principais do outro factor. Rejeita-se H_0 a favor da hipótese alternativa de que existem efeitos principais de temperatura.

Assim, conclui-se que (ao nível $\alpha = 0.05$) existem os três tipos de efeitos previstos no modelo e todos contribuem para formar o valor esperado duma observação.

(d) A menor absorção média amostral verifica-se na célula associada ao tempo de exposição E_1 e temperatura T_1: $\overline{y}_{11.} = 32.23$. No outro extremo, a maior absorção média amostral observa-se quando ao mesmo tempo de exposição (E_1) se associa a temperatura T_2: $\overline{y}_{12.} = 91.43$.

(e) Para considerar as nove diferentes situações experimentais como nove níveis dum único factor, será necessário substituir as duas colunas que na *data frame* `absorcao` indicavam os dois factores (isto é, as colunas `exposicao` e `temperatura`) por um único factor (ao qual dar-se-á o nome `sit.exp`), indicando as nove situações experimentais. Para criar essa nova variável, utilizar-se-á o comando `paste` do `R`, que permite "colar" os valores de cada um dos factores originais, utilizando um ponto como símbolo de separação. O vector assim produzido será transformado em factor através do comando `as.factor`:

```
> sit.exp <- as.factor(paste(absorcao$exposicao, absorcao$temperatura, sep="."))
> sit.exp
 [1] E1.T1 E1.T1 E1.T1 E2.T1 E2.T1 E2.T1 E3.T1 E3.T1 E3.T1 E1.T2 E1.T2 E1.T2 E2.T2
[14] E2.T2 E2.T2 E3.T2 E3.T2 E3.T2 E1.T3 E1.T3 E1.T3 E2.T3 E2.T3 E2.T3 E3.T3 E3.T3
[27] E3.T3
Levels: E1.T1 E1.T2 E1.T3 E2.T1 E2.T2 E2.T3 E3.T1 E3.T2 E3.T3
```

Seguidamente, utiliza-se esse novo factor como preditor numa ANOVA a um factor:

```
> summary(aov(absorcao$abs ~ sit.exp))
            Df Sum Sq Mean Sq F value   Pr(>F)
sit.exp      8   8491  1061.3   63.89 1.22e-11 ***
Residuals   18    299    16.6
```

Em resposta directa à pergunta feita, é evidente pelo *p-value* baixíssimo que se deve considerar que as médias nas nove situações experimentais não são iguais (o que é, aliás, coerente com o que se viu acima). Note-se, no entanto, que a Soma de Quadrados Residual neste novo modelo é igual à que se havia obtido no modelo a dois factores com interacção. O que significa que a

soma de quadrados associada ao único factor agora existente tem de ser equivalente às Somas de Quadrados SQA, SQB e $SQRE$ nesse modelo a dois factores com interacção. Esta constatação reforça a ideia que a diferença entre os dois modelos não reside tanto na capacidade explicativa global, que é a mesma, mas na forma como é (dois factores), ou não (um único factor), possível atribuir essa variabilidade explicada a várias causas (interacção, factor A, factor B).

9. (a) A troca de ordem dos factores no comando do `R` não têm efeito sobre o ajustamento do modelo a dois factores com interacção (além de trocar a ordem das duas primeiras linhas da tabela-resumo), como se pode constatar comparando o ajustamento obtido na alínea 8c com o que se obtém trocando a ordem dos factores:

```
> summary(aov(abs ~ temperatura * exposicao, data=absorcao))
                     Df Sum Sq Mean Sq F value   Pr(>F)
temperatura           2   3674  1836.8  110.58 7.75e-11 ***
exposicao             2   2113  1056.3   63.59 6.92e-09 ***
temperatura:exposicao 4   2704   676.1   40.70 8.74e-09 ***
Residuals            18    299    16.6
```

   No entanto, esta invariância depende do facto de se estar a trabalhar com um delineamento equilibrado, como se verá na alínea seguinte.

   (b) Retirando a primeira e as duas últimas observações, passamos a ter um delineamento análogo, mas não equilibrado. Repare-se nas tabelas-resumo obtidas agora, trocando a ordem dos factores:

```
> summary(aov(abs ~ exposicao * temperatura, data=absorcao[-c(1,26,27),]))
                     Df Sum Sq Mean Sq F value   Pr(>F)
exposicao             2 1445.6   722.8   43.92 5.36e-07 ***
temperatura           2 3032.3  1516.2   92.14 3.76e-09 ***
exposicao:temperatura 4 2444.6   611.1   37.14 1.29e-07 ***
Residuals            15  246.8    16.5
---
> summary(aov(abs ~ temperatura * exposicao , data=absorcao[-c(1,26,27),]))
                     Df Sum Sq Mean Sq F value   Pr(>F)
temperatura           2 2700.4  1350.2   82.05 8.36e-09 ***
exposicao             2 1777.5   888.8   54.01 1.40e-07 ***
temperatura:exposicao 4 2444.6   611.1   37.14 1.29e-07 ***
Residuals            15  246.8    16.5
```

   Como se pode constatar, embora as linhas associadas à interacção e ao residual tenham idênticas somas de quadrados, graus de liberdade e quadrados médios nos dois casos, já as linhas associadas ao efeito principal de cada factor são diferentes nos dois casos. Este problema foi já referido nas aulas teóricas, no final da discussão sobre o modelo a dois factores, sem interacção.

10. (a) Trata-se dum delineamento factorial a dois factores: *localidade* (Factor A, com $a = 4$ níveis) e *cultivar* (Factor B, com $b = 9$ níveis). Existem $n_{ij} = 4 = n_c$ repetições em todas as $ab = 36$ situações experimentais (células), pelo que se trata dum delineamento equilibrado. Existem ao todo $n = abn_c = 144$ observações da variável resposta Y (rendimento, em kg/ha). O modelo ANOVA adequado é o modelo ANOVA a dois factores, com interacção, dado por:

   i. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, $\forall\, i = 1, 2, 3, 4$, $j = 1, 2, ..., 9$, $k = 1, 2, 3, 4$, com $\alpha_1 = 0$, $\beta_1 = 0$, $(\alpha\beta)_{1j} = 0$ para qualquer j, e $(\alpha\beta)_{i1} = 0$ para qualquer i, onde
      - Y_{ijk} indica o rendimento na k-ésima parcela da localidade i, associada à cultivar j;

- μ_{11} indica o rendimento médio (populacional) da cultivar *Celta*, em Elvas;
- α_i indica o efeito principal da localidade i;
- β_j indica o efeito principal da cultivar j;
- $(\alpha\beta)_{ij}$ indica o efeito de interacção entre a localidade i e a cultivar j; e
- ϵ_{ijk} indica o erro aleatório associado à observação Y_{ijk}.

ii. $\epsilon_{ijk} \cap \mathcal{N}(0, \sigma^2), \ \forall\, i, j, k.$

iii. $\{\epsilon_{ijk}\}_{i,j,k}$ constitui um conjunto de variáveis aleatórias independentes.

(b) i. Os nove valores em falta na tabela são dados por:

- $g.l.(SQA) = a - 1 = 3$;
- $g.l.(SQB) = b - 1 = 8$;
- $g.l.(SQAB) = (a-1)(b-1) = 3 \times 8 = 24$;
- $g.l.(SQRE) = n - ab = 144 - 36 = 108$;
- $SQB = QMB\,(b-1) = 964\,060 \times 8 = 7\,712\,480$;
- $SQAB = SQT - (SQA + SQB + SQRE) = (n-1)\,s_y^2 - 219\,628\,472 =$ $= 143 \times 1\,714\,242 - 219\,628\,472 = 25\,508\,134$;
- $QMA = \frac{SQA}{a-1} = \frac{183\,759\,916}{3} = 61\,253\,305$;
- $QMAB = \frac{SQAB}{(a-1)(b-1)} = \frac{25\,508\,134}{24} = 1\,062\,839$;
- $F_B = \frac{QMB}{QMRE} = \frac{964\,060}{260\,704} = 3.69791$.

ii. Pedem-se os três testes F para cada tipo de efeitos previstos no modelo. Efectuemos em pormenor o teste à existência de efeitos de interacção entre localidade e cultivar:

Hipóteses: $H_0 : (\alpha\beta)_{ij} = 0\ ,\ \forall\ i = 2,3,4$ e $j = 2,3,...,9$ [não há interacção] vs. $H_1 : \exists\, i = 2,3,4\ ,\ j = 2,3,...,9$ tais que $(\alpha\beta)_{ij} \neq 0$ [há interacção].

Estatística do teste: $F = \frac{QMAB}{QMRE} \ \cap\ F_{[(a-1)(b-1)\,,\,n-ab]}$, sob H_0.

Nível de significância: $\alpha = 0.01$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.01(24,108)} \approx 1.97$.

Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 4.0768$. É um valor significativo ao nível $\alpha = 0.01$, rejeitando-se H_0 a favor da hipótese alternativa de que existem efeitos de interacção entre localidade e cultivar.

No que respeita ao teste para os efeitos principais do factor *localidade*, as hipóteses em confronto são $H_0 : \alpha_i = 0\ ,\ \forall\ i = 2,3,4$ vs. $H_1 : \exists\, i = 2,3,4\ ,$ tal que $\alpha_i \neq 0$. A Região Crítica é agora dada pela rejeição de H_0 caso $F_{calc} > f_{0.01\,(3,108)} \approx 3.97$. O valor elevadíssimo da estatística calculada $F_{calc} = 234.9531$ leva à rejeição clara de H_0, concluindo-se pela existência de importantes efeitos de localidade, nos rendimentos.

Finalmente, no teste aos efeitos principais do factor *cultivar*, as hipóteses em confronto são $H_0 : \beta_j = 0\ ,\ \forall\ j = 2,3,...,9$ vs. $H_1 : \exists\, j = 2,3,...,9\ ,$ tal que $\beta_j \neq 0$. A Região Crítica é agora dada pela rejeição de H_0 caso $F_{calc} > f_{0.01\,(8,108)} \approx 2.68$. O valor da estatística calculada $F_{calc} = 3.698$ pertence à Região Crítica, levando à rejeição de H_0, concluindo-se também pela existência de efeitos de cultivar sobre os rendimentos.

Assim, conclui-se pela existência dos três tipos de efeitos, ao nível $\alpha = 0.01$, com destaque para a existência clara de efeitos de localidade.

iii. Pede-se para discutir o efeito sobre a tabela resultante de dividir a variável resposta por mil (passando o rendimento a ser expresso em t/ha). Os graus de liberdade não são, naturalmente, afectados. O mesmo não se passa com as Somas de Quadrados. À

nova variável $Y^* = Y/1000$ corresponderão novas médias de nível, de célula e global, que também resultam de dividir por mil (para ficarem em t/ha). Tendo em conta que no modelo em questão, as médias de célula definem os valores ajustados, tem-se $\hat{Y}^*_{ijk} = \hat{Y}_{ijk}/1000$. Assim, as novas Somas de Quadrados resultam de dividir as suas congéneres originais por 1000^2, ou seja, por um milhão. De facto, $SQT^* = \sum_i \sum_j \sum_k (Y^*_{ijk} - \overline{Y}^*_{...})^2 = \sum_i \sum_j \sum_k (Y_{ijk}/1000 - \overline{Y}_{...}/1000)^2 = SQT/(1000^2)$. Também $SQRE^* = \sum_i \sum_j \sum_k (Y^*_{ijk} - \hat{Y}^*_{ijk})^2 = \sum_i \sum_j \sum_k (Y_{ijk}/1000 - \hat{Y}_{ijk}/1000)^2 = SQRE/(1000^2)$. De forma análoga, e utilizando as fórmulas para delineamentos equilibrados,

$$SQA^* = bn_c \sum_{i=1}^{a} (\overline{Y}^*_{i..} - \overline{Y}^*_{...})^2 = bn_c \sum_{i=1}^{a} (\overline{Y}_{i..}/1000 - \overline{Y}_{...}/1000)^2 = SQA/(1000^2)$$

$$SQB^* = an_c \sum_{j=1}^{b} (\overline{Y}^*_{.j.} - \overline{Y}^*_{...})^2 = an_c \sum_{j=1}^{b} (\overline{Y}_{.j.}/1000 - \overline{Y}_{...}/1000)^2 = SQB/(1000^2) \ .$$

Por diferença, tem igualmente de verificar-se $SQAB^* = SQAB/(1000^2)$. Assim, toda a coluna de Somas de Quadrados na tabela será dividida por um milhão. Essa mesma transformação aplica-se à coluna de Quadrados Médios (que resulta de dividir Somas de Quadrados por graus de liberdade). Mas na coluna final, correspondente aos valores calculados das estatísticas F, o quociente de Quadrados Médios mantém-se inalterado (a transformação multiplicativa de numerador e denominador é igual). Logo, as conclusões de todos os testes (incluíndo os respectivos *p-values*) mantêm-se inalterados.

iv. **[Material Complementar]** Os dois gráficos de interacção reflectem a mesma informação, embora de formas diferentes. No gráfico da esquerda, as quatro localidades definem posições no eixo horizontal. Por cima de cada localidade encontram-se nove pontos, associados às nove cultivares. A ordenada de cada um desses nove pontos é dada pelo rendimento médio das parcelas correspondentes a essa combinação de localidade e cultivar. Os segmentos de recta unem os pontos correspondentes a cada cultivar (segundo a legenda indicada no gráfico). Embora haja algum paralelismo nas nove curvas seccionalmente lineares, para as três primeiras localidades, os rendimentos na Revilheira sugerem a existência de efeitos de interacção. Por exemplo, a cultivar *TE9110*, que regista o rendimento mais baixo em Elvas (facto que se pode confirmar na tabela de médias dada na alínea c) tem o segundo mais elevado rendimento na Revilheira. Também a cultivar *Celta*, cujo rendimento em Benavila é o terceiro mais baixo, regista o segundo maior rendimento em Elvas. Assim, há cultivares que manifestam "preferências" ou "aversões" por diferentes localidades, reflectindo efeitos de interacção. O teste à interacção efectuado na alínea anterior confirma que esses efeitos são significativos, ao nível $\alpha = 0.01$.

O gráfico da direita dá, como se disse, uma perspectiva diferente sobre a mesma informação. Agora, são as cultivares que definem nove posições no eixo horizontal. Por cima de cada uma dessas posições (cultivares) há quatro pontos, com ordenadas dadas pelos rendimentos médios da referida cultivar, nas quatro localidades consideradas no ensaio. Segmentos de recta unem os pontos correspondentes a uma mesma localidade. Neste gráfico torna-se evidente que os rendimentos são sempre bastante superiores em Elvas (no gráfico da esquerda, esse facto reflectia-se no "pico" por cima de Elvas). Essa será

a principal razão pela clara rejeição da hipótese nula no teste à existência de efeitos principais de localidade. Por outro lado, os efeitos de interacção reflectem-se na mais visível ausência de paralelismo, nomeadamente nos traços correspondentes a Elvas e Revilheira, que para várias cultivares parecem ter comportamentos quase antagónicos.

11. (a) Trata-se dum delineamento factorial a dois factores: *Temperatura de conservação* (Factor A), com $a = 2$ níveis, e *Tempo de armazenamento* (Factor B), com $b = 4$ níveis. Para modelar a variável resposta Y (*alterações no conteúdo em taninos das polpas de sapoti*), utiliza-se um modelo ANOVA a dois factores, com interacção. É possível estudar a interacção devido à presença de repetições nas $2 \times 4 = 8$ células. Sempre que possível, é desejável considerar este modelo para delineamentos factoriais a dois factores, deixando que sejam os dados a sugerir se se deve admitir a existência desse tipo de efeitos. O delineamento é equilibrado, uma vez que todas as células têm o mesmo número de repetições: $n_{ij} = 4 = n_c \ (\forall\, i, j)$, para um total de $n = 8 \times 4 = 32$ observações. O modelo é dado por:

    i. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \ \ \forall\, i = 1, 2 \ , \ j = 1, 2, 3, 4 \ , \ k = 1, 2, 3, 4,$
    com $\alpha_1 = 0$, $\beta_1 = 0$, $(\alpha\beta)_{1j} = 0$ para qualquer j, e $(\alpha\beta)_{i1} = 0$ para qualquer i, onde
    - Y_{ijk} indica a k-ésima observação (repetição) na célula definida pelo nível i do Factor A e o nível j do Factor B;
    - μ_{11} indica a média (populacional) das observações na célula $(1, 1)$, ou seja, com temperatura alta e 0 dias de armazenamento;
    - α_i indica o efeito do nível i do Factor A (*Temperatura*);
    - β_j indica o efeito do nível j do Factor B (*Tempo de armazenamento*);
    - $(\alpha\beta)_{ij}$ indica o efeito de interacção na célula (i, j); e
    - ϵ_{ijk} indica o erro aleatório associado à observação Y_{ijk}.

    ii. $\epsilon_{ijk} \cap \mathcal{N}(0, \sigma^2) \ , \ \ \forall\, i, j, k.$

    iii. $\{\epsilon_{ijk}\}_{i,j,k}$ constituem um conjunto de variáveis aleatórias independentes.

    (b) A tabela-resumo desta ANOVA terá três linhas associadas a cada tipo de efeitos previsto no modelo (ou seja, efeitos principais do Factor A, efeitos principais do Factor B e efeitos de interacção) e ainda uma linha para o residual (podendo também incluir-se a linha associada à variabilidade Total). Como em qualquer modelo ANOVA, a tabela-resumo tem as seguintes colunas: Somas de Quadrados, graus de liberdade correspondentes, Quadrados Médios e estatísticas F. Os graus de liberdade são dados por:

    - Factor A: $a - 1 = 1$;
    - Factor B: $b - 1 = 3$;
    - Interacção: $(a - 1)(b - 1) = 3$;
    - Residual: $n - ab = 32 - 8 = 24$.

    Para calcular as somas de quadrados, registamos que no enunciado é dada a Soma de Quadrados Residual $SQRE = 20.72$. É igualmente dado o Quadrado Médio do Factor B, e multiplicando pelos respectivos graus de liberdade obtém-se $SQB = QMB\,(b - 1) = 96.01 \times 3 = 288.03$. A Soma de Quadrados Total também pode ser calculada facilmente, uma vez que no enunciado á dada a variância da totalidade das observações de Y, $s_y^2 = 47.83222$, e $SQT = (n - 1)\, s_y^2 = 31 \times 47.83222 = 1482.799$. Assim, faltam as duas Somas de Quadrados relativas aos efeitos principais do factor A (SQA) e aos efeitos de interacção ($SQAB$). Utilizando a expressão para SQA, no caso de delineamentos equilibrados (disponível no formulário) e os valores das médias de nível do factor A e da média geral (disponíveis no

enunciado), tem-se $SQA = bn_c \sum_{i=1}^{a} (\overline{Y}_{i..} - \overline{Y}_{...})^2 = 16\,[(24.681 - 22.14375)^2 + (19.606 - 22.14375)^2] = 16 \times 12.87781 = 206.045$. A última Soma de Quadrados em falta ($SQAB$) pode ser calculada a partir das restantes quatro: $SQAB = SQT - (SQA + SQB + SQRE) = 1482.799 - (206.045 + 288.03 + 20.72) = 968.004$. Assim,

Variação	g.l.	SQs	QMs	F_{calc}
Factor A	1	206.045	$QMA = \frac{SQA}{a-1} = 206.045$	$F = \frac{QMA}{QMRE} = 238.6622$
Factor B	3	288.03	$QMB = \frac{SQB}{b-1} = 96.01$	$F = \frac{QMB}{QMRE} = 111.2085$
Interacção	3	968.004	$QMAB = \frac{SQAB}{(a-1)(b-1)} = 322.668$	$F = \frac{QMAB}{QMRE} = 373.7467$
Residual	24	20.72	$QMRE = \frac{SQRE}{n-ab} = 0.8633333$	–
Total	31	1482.799	–	–

(c) De acordo com o modelo, a influência do Factor B nos valores da variável resposta pode resultar de dois tipos de efeitos: os efeitos principais do Factor B (os β_j) ou os efeitos de interacção (os $(\alpha\beta)_{ij}$). Efectuaremos estes dois testes, começando pelo dos efeitos de interacção. Neste exemplo, e como o Factor A apenas tem dois níveis, o índice i nos efeitos de interacção apenas toma o valor $i = 2$.

Hipóteses: $H_0 : (\alpha\beta)_{2j} = 0\ ,\ \forall\ j = 2,3,4$ vs. $H_1 : \exists\, j = 2,3,4$ tal que $(\alpha\beta)_{2j} \neq 0$.

Estatística do teste: $F = \frac{QMAB}{QMRE} \ \cap\ F_{[(a-1)(b-1)\,,\,n-ab]}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(3,24)} = 3.01$.

Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 373.7467$. É um valor claramente significativo e rejeita-se H_0 a favor da hipótese alternativa de que existem efeitos de interacção.

Já é possível responder afirmativamente: o Factor B tem efeitos sobre os valores médios de Y. No entanto, efectuaremos também o teste aos efeitos principais do Factor B:

Hipóteses: $H_0 : \beta_j = 0\ ,\ \forall\ j = 2,3,4$ vs. $H_1 : \exists\, j = 2,3,4$ tal que $\beta_j \neq 0$.

Estatística do teste: $F = \frac{QMB}{QMRE} \ \cap\ F_{(b-1\,,\,n-ab)}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(3,24)} = 3.01$.

Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 111.2085$. É um valor claramente significativo e rejeita-se H_0 a favor da hipótese de que existem efeitos principais do Factor B.

Assim, quer pela via dos efeitos principais, quer pela via dos efeitos de interacção, o Factor B (*tempo de armazenamento*) afecta os conteúdos médios de taninos nos sapotis.

12. (a) Pede-se para mostrar que a soma dos n_i resíduos e_{ij}, correspondentes ao nível i do Factor ($i = 1, 2, ..., k$), numa ANOVA a 1 Factor, é nula. Sabemos que, neste tipo de delineamento, os valores ajustados de cada observação correspondem à média amostral das n_i observações no nível i do Factor em que essa observação foi efectuada. Assim,

$$\sum_{j=1}^{n_i} e_{ij} \ = \ \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij}) \ = \ \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i.}) \ = \ 0\ ,$$

uma vez que se trata duma soma de desvios dum conjunto de observações em relação à sua média (ou seja, do tipo $\sum_{i=1}^{n}(x_i - \overline{x})$) que têm sempre soma zero, como facilmente se constata.

(b) Trata-se duma situação análoga à da alínea anterior. Num modelo ANOVA a dois factores, com efeitos de interacção, sabemos que os valores ajustados $\hat{y}_{ijk}$ correspondem às médias $\overline{y}_{ij.}$ das observações da célula da referida observação. Assim, a soma dos resíduos das n_{ij} observações efectuadas na célula (i,j) é dada por:

$$\sum_{k=1}^{n_{ij}} e_{ijk} \;=\; \sum_{k=1}^{n_{ij}} (y_{ijk} - \hat{y}_{ijk}) \;=\; \sum_{k=1}^{n_{ij}} (y_{ijk} - \overline{y}_{ij.}) \;=\; 0 \;.$$

13. Está-se no contexto dum modelo ANOVA a 1 Factor, onde as observações Y_{ij} constituem n variáveis aleatórias independentes, todas com distribuição $Y_{ij} \cap \mathcal{N}(\mu_1 + \alpha_i \,,\, \sigma^2)$.

(a) Sabemos que neste modelo, os estimadores dos parâmetros μ_1 e $\alpha_i = \mu_i - \mu_1$ são dados pelas correspondentes quantidades amostrais.

- o estimador da média populacional do primeiro nível, μ_1, é dado pela média amostral das observações desse nível, $\overline{Y}_{1.} = \frac{1}{n_1} \sum_{j=1}^{n_1} Y_{ij}$. Mas, como é sabido (ver apontamentos da UC de Estatística, dos primeiros ciclos do ISA), a média $\overline{X}$ duma amostra aleatória $\{X_i\}_{i=1}^{n}$ de n variáveis aleatórias com distribuição $\mathcal{N}(\mu, \sigma^2)$, tem distribuição $\overline{X} \cap \mathcal{N}(\mu, \frac{\sigma^2}{n})$. Assim, e tendo em conta que $\alpha_1 = 0$, tem-se $Y_{1j} \cap \mathcal{N}(\mu_1 \,,\, \sigma^2)$ e $\hat{\mu}_1 = \overline{Y}_{1.} \cap \mathcal{N}(\mu_1 \,,\, \frac{\sigma^2}{n_1})$, como se quer mostrar.
- O estimador de $\alpha_i = \mu_i - \mu_1$, para $i > 1$, é dado pela correspondente diferença de médias amostrais, $\hat{\alpha}_i = \overline{Y}_{i.} - \overline{Y}_{1.}$. Viu-se na alínea anterior que a segunda parcela tem distribuição $\mathcal{N}(\mu_1 \,,\, \frac{\sigma^2}{n_1})$. Por um raciocínio análogo, a primeira parcela tem distribuição $\overline{Y}_{i.} \cap \mathcal{N}(\mu_1 + \alpha_i \,,\, \frac{\sigma^2}{n_i})$. As duas parcelas são independentes, uma vez que as parcelas que entram para o cálculo da média $\overline{Y}_{1.}$ são diferentes das que entram no cálculo da média $\overline{Y}_{i.}$. Logo, essa diferença de duas variáveis aleatórias Normais independentes tem distribuição Normal. Os parâmetros dessa distribuição são: $E[\hat{\alpha}_i] = E[\overline{Y}_{i.} - \overline{Y}_{1.}] = E[\overline{Y}_{i.}] - E[\overline{Y}_{1.}] = (\mu_1 + \alpha_i) - \mu_1 = \alpha_i$; e $V[\hat{\alpha}_i] = V[\overline{Y}_{i.} - \overline{Y}_{1.}] = V[\overline{Y}_{i.}] + V[\overline{Y}_{1.}] - 2\,\underbrace{Cov[\overline{Y}_{i.}, \overline{Y}_{1.}]}_{=0} = \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_1}$ (a covariância é nula, tendo em conta a independência de duas médias de nível diferentes). Logo, $\hat{\alpha}_i \cap \mathcal{N}\left(\alpha_i \,,\, \sigma^2(\frac{1}{n_1} + \frac{1}{n_i})\right)$, como se queria mostrar.

(b) Qualquer dos parâmetros α_i ou μ_1 são parâmetros dum modelo linear, análogos aos parâmetros β_j numa regressão linear. Vimos, no contexto da regressão linear, que os intervalos de confiança para esses parâmetros são construídos a partir quantidades do tipo $\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}}$, onde $\hat{\sigma}^2_{\hat{\beta}_j} = \hat{V}[\hat{\beta}_j]$, a variância estimada do estimador $\hat{\beta}_j$. Estes quocientes têm distribuição t, com graus de liberdade dados pelos g.l. de $SQRE$. O que faz falta é adaptar esta expressão geral, obtida no estudo da modelo linear, ao contexto específico dum modelo ANOVA a 1 Factor. No que respeita aos graus de liberdade, sabemos serem $n-k$. Falta indicar as expressões dos estimadores e das respectivas variâncias. Consideremos primeiro o caso do parâmetro μ_1 (que é a constante aditiva geral, equivalente ao parâmetro β_0 numa regressão linear).

- Já vimos na alínea anterior que $\hat{\mu}_i = \overline{Y}_{1.}$, a média amostral das observações do primeiro nível do factor.
- Também vimos na alínea anterior que $V[\hat{\mu}_1] = \frac{\sigma^2}{n_1}$. Esta quantidade é estimada por $\hat{\sigma}^2_{\hat{\mu}_1} = \frac{QMRE}{n_1}$.
- Logo, temos que $\frac{\hat{\mu}_1 - \mu_1}{\sqrt{QMRE/n_1}} \cap t_{n-k}$.

Este último resultado é o ponto de partida para a construção dum intervalo a $(1-\alpha)\times 100\%$ de confiança para o parâmetro μ_1. Designando (como de costume) por $t_{\alpha/2\,(n-k)}$ o valor que, numa distribuição t-Student com $n-k$ graus de liberdade, deixa à sua direita uma região de probabilidade $\frac{\alpha}{2}$, temos

$$
\begin{aligned}
& P\left[-t_{\alpha/2\,(n-k)} \;<\; \frac{\hat{\mu_1} - \mu_1}{\sqrt{\frac{QMRE}{n_1}}} \;<\; t_{\alpha/2\,(n-k)}\right] \;=\; 1-\alpha \\
\Leftrightarrow \quad & P\left[-t_{\alpha/2\,(n-k)} \cdot \sqrt{\frac{QMRE}{n_1}} \;<\; \hat{\mu_1} - \mu_1 \;<\; t_{\alpha/2\,(n-k)} \cdot \sqrt{\frac{QMRE}{n_1}}\right] \;=\; 1-\alpha \\
\Leftrightarrow \quad & P\left[t_{\alpha/2\,(n-k)} \cdot \sqrt{\frac{QMRE}{n_1}} \;>\; \mu_1 - \hat{\mu_1} \;>\; -t_{\alpha/2\,(n-k)} \cdot \sqrt{\frac{QMRE}{n_1}}\right] \;=\; 1-\alpha \\
\Leftrightarrow \quad & P\left[\hat{\mu_1} - t_{\alpha/2\,(n-k)} \cdot \sqrt{\frac{QMRE}{n_1}} \;<\; \mu_1 \;<\; \hat{\mu_1} + t_{\alpha/2\,(n-k)} \cdot \sqrt{\frac{QMRE}{n_1}}\right] \;=\; 1-\alpha
\end{aligned}
$$

Calculando os extremos deste intervalo de probabilidade para a nossa amostra (e recordando que $\hat{\mu}_1 = \overline{Y}_{1.}$) obtemos o intervalo de confiança referido no enunciado.

Para obter um intervalo de confiança para α_i, segue-se um raciocínio em tudo análogo ao acabado de referir. Agora,

- $\hat{\alpha}_i = \overline{Y}_{i.} - \overline{Y}_{1.}$;
- Como se viu na alínea anterior, $V[\alpha_i] = \sigma^2\,(\frac{1}{n_1} + \frac{1}{n_i})$, que é estimada por $\hat{V}[\alpha_i] = QMRE\,(\frac{1}{n_1} + \frac{1}{n_i})$;
- Logo, temos que $\frac{\hat{\alpha}_i - \alpha_i}{\sqrt{QMRE(\frac{1}{n_1}+\frac{1}{n_i})}} \cap t_{n-k}$.

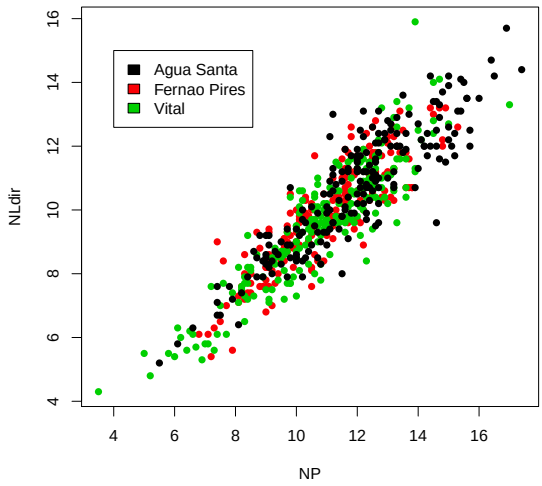
A dedução do intervalo de confiança para α_i é também em tudo análoga ao que foi feita no caso de μ_1, substituindo μ_1 por α_i, $\hat{\mu}_1$ por $\hat{\alpha}_i$ e $\sqrt{\frac{QMRE}{n_1}}$ por $\sqrt{QMRE\,\left(\frac{1}{n_1} + \frac{1}{n_i}\right)}$.

3. Análise de Covariância

1. Neste exercício consideram-se os dados da *data frame* `videiras`. A variável resposta é, em todas as alíneas, o comprimento da nervura lateral direita (`NLdir`) e o preditor, o comprimento da nervura principal (`NP`).

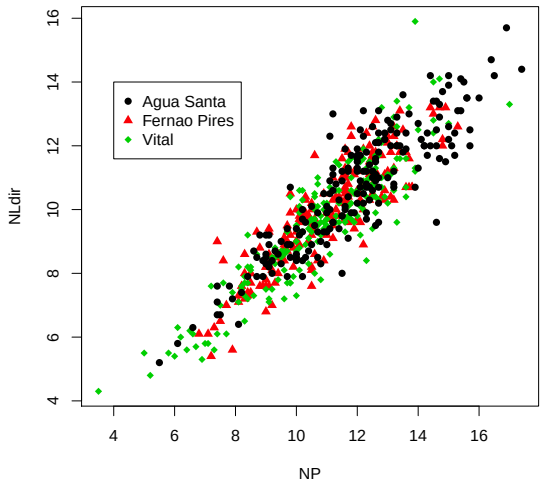
   (a) Os comandos `R` para obter a nuvem de pontos pedida, e o respectivo resultado, são:

   ```
   > plot(NLdir ~ NP, col=Casta, data=videiras, pch=16)
   > legend(4,15,legend=levels(videiras$Casta), fill=1:3)
   ```



Alternativamente, podemos também querer construir um gráfico com, não apenas cores diferentes, mas também símbolos diferentes para cada casta. Eis uma forma possível de construir um tal gráfico no R, usando os símbolos a que correspondem os códigos 16 (círculos), 17 (triângulos) e 18 (losangos), como indicado na legenda.

```
> plot(NLdir ~ NP, col=as.numeric(Casta), pch=as.numeric(Casta)+15, data=videiras)
> legend(4,14,levels(videiras$Casta),col=1:3, pch=16:18)
```



A nuvem de pontos sugere a existência duma relação linear bastante intensa, que poderá ser a mesma nas três castas consideradas. A nuvem sugere também que poderá haver dispersões maiores das observações, em torno da recta de fundo, para as folhas de maior dimensão.

(b) Eis os comandos R necessários, e os resultados numéricos correspondentes:

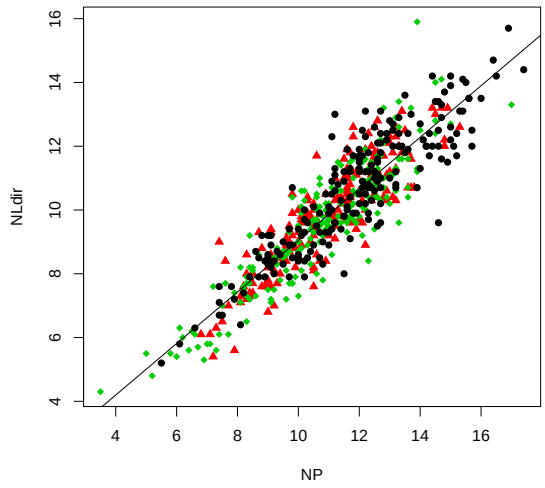
```
> videirasN.lm <- lm(NLdir ~ NP, data=videiras)
> summary(videirasN.lm)
[...]
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.96218    0.18309   5.255 2.06e-07 ***
NP           0.80841    0.01607  50.314  < 2e-16 ***
---
Residual standard error: 0.8339 on 598 degrees of freedom
Multiple R-squared:  0.8089,Adjusted R-squared:  0.8086
F-statistic:  2532 on 1 and 598 DF,  p-value: < 2.2e-16
```

```
> abline(videirasN.lm, col="blue")
```

Assim, a recta de regressão $y = 0.96218 + 0.80841\,x$ explica cerca de 81% da variabilidade observada nas nervuras laterais direitas, para o conjunto das $n = 600$ observações. Trata-se duma aproximação razoavelmente boa (como se pode constatar no gráfico), que explica cerca de 81% da variabilidade observada nas nervuras laterais direitas. Como seria de esperar, o modelo ajustado difere significativamente do modelo nulo, tendo a estatística calculada no teste F de ajustamento global um valor $F_{calc} = 2532$, cuja significância (*p-value*) correspondente é inferior à precisão de máquina, logo indistinguível de zero.

```
> abline(videirasN.lm)
```



(c) Eis os comandos R necessários, e os resultados numéricos correspondentes ao modelo ANCOVA pedido:

```
> videirasNCasta.lm <- lm(NLdir ~ NP*Casta, data=videiras)
> summary(videirasNCasta.lm)
[...]
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.39812    0.32102   4.355 1.57e-05 ***
NP                    0.77780    0.02654  29.305  < 2e-16 ***
CastaFernao Pires    -0.43069    0.48897  -0.881    0.379
CastaVital           -0.66120    0.43788  -1.510    0.132
NP:CastaFernao Pires  0.03395    0.04253   0.798    0.425
NP:CastaVital         0.04100    0.03798   1.079    0.281
---
```

```
Residual standard error: 0.8316 on 594 degrees of freedom
Multiple R-squared:  0.8112,Adjusted R-squared:  0.8096
F-statistic: 510.5 on 5 and 594 DF,  p-value: < 2.2e-16
```

A recta para a casta Água Santa (a casta correspondente ao primeiro nível do factor, o nível de referência, logo não explicitada na listagem de resultados) tem equação $y = 1.39812 + 0.77780\,x$. Para obter a equação correspondente à casta Fernão Pires, será necessário acrescentar à ordenada na origem o acréscimo estimado $\hat{\alpha}_{0:2} = -0.43069$ e ao declive, o respectivo acréscimo estimado, $\hat{\alpha}_{1:2} = 0.03395$. De forma análoga, obtém-se a recta ajustada para a casta Vital. Eis as equações das três rectas ajustadas:

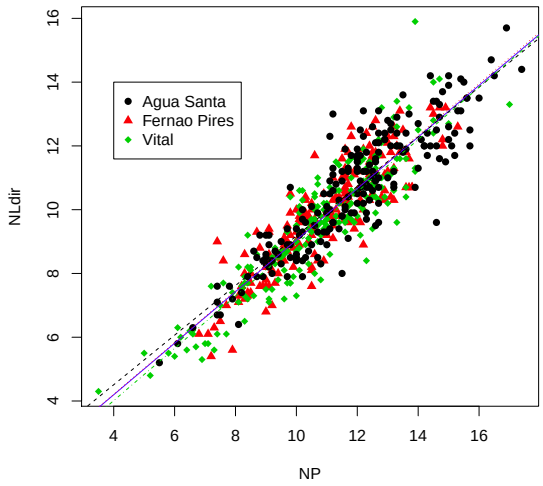
Casta Água Santa	$y = 1.39812 + 0.77780\,x$
Casta Fernão Pires	$y = (1.39812 - 0.43069) + (0.777$
Casta Vital	$y = (1.39812 - 0.66120) + (0.777$

Para traçar as rectas de cada casta na nuvem de pontos já criada, podem usar-se os seguintes comandos:

```
> coefVidCasta <- coef(videirasNCasta.lm)
> coefVidCasta
(Intercept)         NP  CastaFernao Pires  CastaVital  NP:CastaFernao Pires   NP:CastaVital
1.39811600  0.77779606        -0.43068514  -0.66119902           0.03394865      0.04100268

> abline(coefVidCasta[c(1,2)], col=1, lty=2)                      <-- recta casta Água Santa
> abline(coefVidCasta[c(1,2)]+coefVidCasta[c(3,5)],col=2,lty=3)   <-- recta casta Fernão Pires
> abline(coefVidCasta[c(1,2)]+coefVidCasta[c(4,6)],col=3,lty=4)   <-- recta casta Vital
```

Apesar das equações diferentes, as quatro rectas são difíceis de distinguir no gráfico.



(d) A equação do modelo ANCOVA ajustado pode escrever-se da seguinte forma, utilizando a notação vectorial:

$$\vec{\mathbf{y}} \;=\; \beta_0 + \beta_1\,\vec{\mathbf{x}} + \alpha_{0:2}\boldsymbol{\mathcal{I}}_2 + \alpha_{0:3}\boldsymbol{\mathcal{I}}_3 + \alpha_{1:2}\boldsymbol{\mathcal{I}}_2 \star \vec{\mathbf{x}} + \alpha_{1:3}\boldsymbol{\mathcal{I}}_3 \star \vec{\mathbf{x}} + \boldsymbol{\epsilon}\ ,$$

sendo $\boldsymbol{\mathcal{I}}_i$ a variável indicatriz das observações da casta $i = 2, 3$ (Fernão Pires e Vital, respectivamente) e $\alpha_{j:i}$ o acréscimo no parâmetro β_j (em relação à casta de referência, a Água Santa), resultante de estarmos na casta $i = 2, 3$. O símbolo $\star$ indica um produto

elemento a elemento entre dois vectores de igual dimensão. O modelo linear ajustado acima pode agora ser visto como um submodelo deste modelo ANCOVA, associado à hipótese $\alpha_{0:2} = \alpha_{0:3} = \alpha_{1:2} = \alpha_{1:3} = 0$. Vamos efectuar um teste F parcial para testar a equivalência de modelo e submodelo.

Hipóteses: $H_0 : \alpha_{j:i} = 0, \forall j = 0, 1; i = 2, 3$ vs. $H_1 : \exists j = 0, 1; i = 2, 3$ tal que $\alpha_{j:i} \neq 0$.

Estatística do Teste: (na forma mais adequada à informação disponível)

$F = \frac{R_c^2 - R_s^2}{1 - R_c^2} \cdot \frac{n-(p+1)}{p-k} \cap F_{(p-k,n-(p+1))}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica: (unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(4,594)} \approx 2.39$.

Conclusão: Temos $F_{calc} = \frac{0.8112-0.8089}{1-0.8112} \cdot \frac{594}{4} = 1.809$. Logo, não rejeitamos H_0, isto é, não se pode dizer que o modelo ANCOVA se ajuste de forma significativamente diferente do modelo RLS com uma única recta para as três castas. Assim, não se justifica abandonar o modelo RLS, que é mais parcimonioso e tem um ajustamento considerado adequado.

Este teste F parcial, comparando o modelo ANCOVA ajustado na alínea anterior com o submodelo ajustado na alínea 1b (recta única para a totalidade das observações) obtém-se no `R` com o comando `anova`:

```
> anova(videirasN.lm, videirasNCasta.lm)
Analysis of Variance Table
Model 1: NLdir ~ NP
Model 2: NLdir ~ NP * Casta
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    598 415.80
2    594 410.81  4    4.9948 1.8055 0.1262
```

NOTA: A pequena discrepância no valor calculado da estatística de teste resulta de, na nossa resolução anterior, terem sido usados valores de R^2 arredondados a 4 casas decimais.

(e) Eis os três ajustamentos "mono-casta" pedidos.

i. Tendo em atenção que as $n_1 = 200$ observações da casta Água Santa estão nas linhas 401 a 600 da *data frame*, tem-se:

```
> summary(lm(NLdir ~ NP, data=videiras[401:600,]))
[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.39812    0.33349   4.192 4.16e-05 ***
NP           0.77780    0.02757  28.210  < 2e-16 ***
---
Residual standard error: 0.8639 on 198 degrees of freedom
Multiple R-squared:  0.8008,Adjusted R-squared:  0.7998
F-statistic: 795.8 on 1 and 198 DF,  p-value: < 2.2e-16
```

A recta de regressão obtida ($y = 1.39812 + 0.77780\,x$) é a mesma que no modelo completo (modelo ANCOVA) considerado acima. O valor do coeficiente de determinação ($R^2 = 0.8008$) é muito próximo do valor obtido com a recta única para a totalidade das $n = 600$ observações, facto que não era possível prever a partir dos ajustamentos anteriores.

ii. As $n_2 = 200$ observações da casta Fernão Pires estão nas 200 primeiras linhas do objecto `videiras`. Assim,

```
> summary(lm(NLdir ~ NP, data=videiras[1:200,]))
[...]
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.96743    0.34914   2.771  0.00612 **
NP           0.81174    0.03146  25.801  < 2e-16 ***
---
Residual standard error: 0.7872 on 198 degrees of freedom
Multiple R-squared:  0.7708,Adjusted R-squared:  0.7696
F-statistic: 665.7 on 1 and 198 DF,  p-value: < 2.2e-16
```

Também neste caso, e como teria de ser, a recta obtida ($y = 0.96743 + 0.81174\,x$) é, a menos de erros de arredondamento, a recta obtida ao ajustar o modelo ANCOVA. Também neste caso, o coeficiente de determinação $R^2 = 0.7708$ é próximo do valor obtido para a recta única, embora neste caso não tenha necessariamente de ser assim.

iii. Para as restantes observações, relativas à casta Vital, tem-se:

```
> summary(lm(NLdir ~ NP, data=videiras[201:400,]))
[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.73692    0.30147   2.444   0.0154 *
NP           0.81880    0.02751  29.769   <2e-16 ***
---
Residual standard error: 0.8418 on 198 degrees of freedom
Multiple R-squared:  0.8174,Adjusted R-squared:  0.8164
F-statistic: 886.2 on 1 and 198 DF,  p-value: < 2.2e-16
```

Confirma-se a recta de regressão $y = 0.73692 + 0.8188\,x$, e mais uma vez o valor $R^2 = 0.8174$ é próximo do obtido com uma única recta de regressão para as três castas, o que é, como para as outras castas, uma particularidade deste exemplo, associada ao facto de as três nuvens de pontos serem de configuração semelhante.

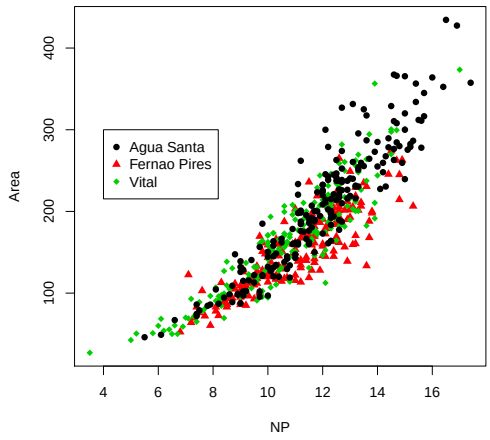
(f) O único modelo que não é de RLS é o modelo completo de ANCOVA, e será o único cuja matriz do modelo é aqui considerada. A fim de poupar no espaço, apenas se mostram as linhas correspondentes às três primeiras observações de cada casta. Recorde-se que à casta de referência (que, uma vez que o `R` ordena os níveis do factor por ordem alfabética, é a Água Santa) correspondem as últimas 200 linhas da matriz. As restantes castas estão indicadas nos nomes de coluna.

```
> model.matrix(videirasNCasta.lm)
    (Intercept)   NP CastaFernao Pires CastaVital NP:CastaFernao Pires NP:CastaVital
1             1 13.8                1          0                 13.8           0.0
2             1  9.1                1          0                  9.1           0.0
3             1 14.5                1          0                 14.5           0.0
[...]
201           1 11.7                0          1                  0.0          11.7
202           1 10.6                0          1                  0.0          10.6
203           1 11.0                0          1                  0.0          11.0
[...]
401           1 15.7                0          0                  0.0           0.0
402           1 11.7                0          0                  0.0           0.0
403           1 10.2                0          0                  0.0           0.0
[...]
```

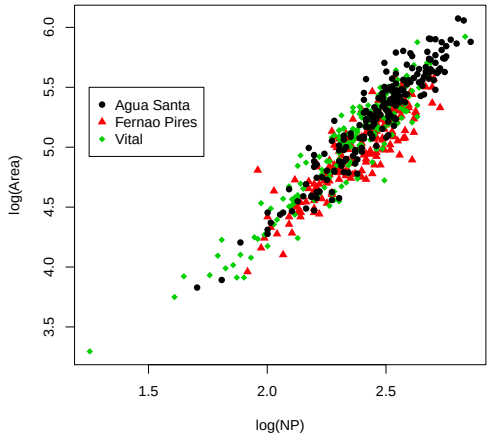
2. Neste exercício a variável resposta é `Area` e a variável preditora é `NP`.

(a) O gráfico (obtido de forma análoga ao que foi visto no Exercício 1a) torna evidente a existência duma curvatura na relação entre área foliar e comprimento da nervura principal.

esta curvatura não é de estranhar, uma vez que a área é uma característica bi-dimensional, enquanto que o comprimento é unidimensional, sugerindo que a área seja aproximadamente proporcional ao quadrado do comprimento da nervura.



(b) Com a dupla logaritmização pedida no enunciado obtém-se uma relação mais próxima da linearidade. Assim, a logaritmização de área foliar e de comprimento da nervura principal é uma boa transformação linearizante.



(c) O modelo pedido tem o seguinte ajustamento.

```
> vid.Anc2.lm <- lm(log(Area) ~ log(NP), data=videiras)
> summary(vid.Anc2.lm)
[...]
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.57869    0.07703   7.513 2.12e-13 ***
log(NP)      1.87642    0.03203  58.579  < 2e-16 ***
---
Residual standard error: 0.1597 on 598 degrees of freedom
Multiple R-squared: 0.8516,Adjusted R-squared: 0.8513
```

```
F-statistic:  3431 on 1 and 598 DF,  p-value: < 2.2e-16
```

A recta ajustada, às variáveis logaritmizadas é $\ln(Area) = 0.57869 + 1.87642\,\ln(NP)$. Em termos das variáveis originais (não logaritmizadas), esta relação corresponde a uma relação potência $Area = \mathrm{e}^{0.57869}\,NP^{1.87642}$ (ver acetatos das aulas relativos às transformações linearizantes).

(d) O modelo ANCOVA agora pedido tem o seguinte ajustamento:

```
> vid.Anc2d <- lm(log(Area) ~ log(NP)*Casta, data=videiras)
> summary(vid.Anc2d)
[...]
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)                0.33820    0.13050   2.592 0.009791 **
log(NP)                    1.99648    0.05294  37.711  < 2e-16 ***
CastaFernao Pires          0.62328    0.19914   3.130 0.001834 **
CastaVital                 0.39524    0.17007   2.324 0.020463 *
log(NP):CastaFernao Pires -0.31298    0.08232  -3.802 0.000158 ***
log(NP):CastaVital        -0.17654    0.07025  -2.513 0.012232 *
---
Residual standard error: 0.1482 on 594 degrees of freedom
Multiple R-squared: 0.8731,Adjusted R-squared: 0.872
F-statistic: 817.4 on 5 and 594 DF,  p-value: < 2.2e-16
```

O valor do coeficiente de determinação deste modelo ($R^2 = 0.8731$) é comparável com o do modelo de regressão linear simples ajustado na alínea anterior ($R^2 = 0.8516$), uma vez que em ambos os casos a escala da variável resposta é a de log-áreas. O coeficiente de determinação aumentou com o modelo ANCOVA (como tem de ser, uma vez que o modelo de uma única recta de regressão é um submodelo do modelo ANCOVA), mas o aumento não é muito acentuado (pouco mais de 2%), pelo que é legítima a dúvida se o aumento obtido com o modelo ANCOVA compensa a maior complexidade do modelo.

(e) Tendo em conta a natureza destes parâmetros estimados, resultam as seguintes relações para cada casta:

Água Santa	$\ln(Area) = 0.33820 + 1.99648\,\ln(NP)$	$\Leftrightarrow$	$Area = \mathrm{e}^{0.33820}\,NP^{1.99648}$
Fernão Pires	$\ln(Area) = 0.96148 + 1.6835\,\ln(NP)$	$\Leftrightarrow$	$Area = \mathrm{e}^{0.96148}\,NP^{1.6835}$
Vital	$\ln(Area) = 0.73344 + 1.81994\,\ln(NP)$	$\Leftrightarrow$	$Area = \mathrm{e}^{0.73344}\,NP^{1.81994}$

Em todos os casos, a área foliar é modelada como proporcional a uma potência do comprimento da nervura principal, potência essa que varia entre 1.68 e 2. Uma relação $Area = NP^2$ corresponderia a folhas de forma quadrada, com lado igual a NP. A forma irregular da folha justifica as potências menores que 2 e as constantes de proporcionalidade, que oscilam entre 1.40 (no caso da casta Água Santa) e 2.62 (casta Fernão Pires).

(f) Uma vez que os modelos das alíneas (c) e (d) são modelos encaixados, é possível usar um teste F parcial para estudar se o respectivo ajustamento é significativamente diferente. A equação do modelo ANCOVA é da forma

$$\vec{\mathbf{y}} \;=\; \beta_0 + \beta_1\,\vec{\mathbf{x}} + \alpha_{0:2}\vec{\boldsymbol{\mathcal{I}}}_2 + \alpha_{0:3}\vec{\boldsymbol{\mathcal{I}}}_3 + \alpha_{1:2}\vec{\boldsymbol{\mathcal{I}}}_2 \star \vec{\mathbf{x}} + \alpha_{1:3}\vec{\boldsymbol{\mathcal{I}}}_3 \star \vec{\mathbf{x}} + \vec{\boldsymbol{\epsilon}}\,,$$

sendo $\vec{\boldsymbol{\mathcal{I}}}_i$ a variável indicatriz das observações da casta $i = 2, 3$ (Fernão Pires e Vital, respectivamente) e $\alpha_{j:i}$ o acréscimo no parâmetro β_j (em relação à casta de referência, a Água Santa), resultante de estarmos na casta $i = 2, 3$. O símbolo $\star$ indica um produto

elemento a elemento entre dois vectores de igual dimensão. O modelo linear ajustado acima pode agora ser visto como um submodelo deste modelo ANCOVA, associado à hipótese $\alpha_{0:2}=\alpha_{0:3}=\alpha_{1:2}=\alpha_{1:3}=0$.

Hipóteses: $H_0 : \alpha_{j:i}=0, \forall j=0,1; i=2,3$ vs. $H_1 : \exists j=0,1; i=2,3$ tal que $\alpha_{j:i}\neq 0$.

Estatística do Teste: (na forma mais adequada à informação disponível)

$F = \frac{R_c^2-R_s^2}{1-R_c^2}\cdot\frac{n-(p+1)}{p-k} \cap F_{(p-k,n-(p+1))}$, sob H_0.

Nível de significância: $\alpha=0.05$.

Região Crítica: (unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(4,594)} \approx 2.39$.

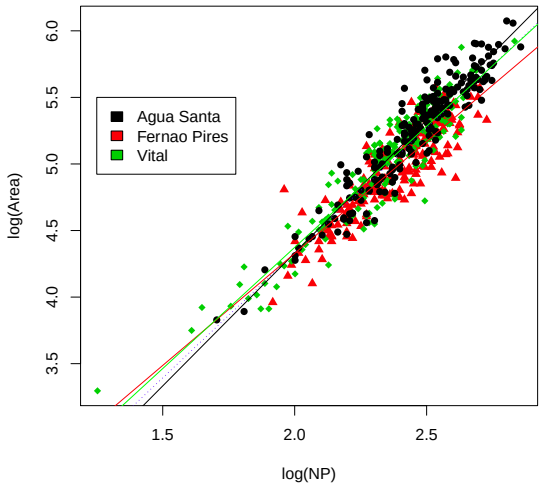
Conclusão: Temos $F_{calc}=\frac{0.8731-0.8516}{1-0.8731}\cdot\frac{594}{4}=25.15957$. Logo, neste caso rejeita-se claramente H_0, isto é, conclui-se que o ajustamento do modelo ANCOVA é significativamente diferente do ajustamento do modelo RLS com uma única recta para as três castas. Assim, do ponto de vista estatístico justifica-se a utilização do modelo ANCOVA, com rectas/curvas diferentes para cada casta.

O recurso ao comando `anova` do `R` confirma o valor calculado da estatística (arredondamentos aparte) e o valor quase nulo do *p-value* correspondente.

```
> anova(vid.Anc2.lm, vid.Anc2d)
Analysis of Variance Table
Model 1: log(Area) ~ log(NP)
Model 2: log(Area) ~ log(NP) * Casta
Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    598 15.248
2    594 13.037  4    2.2102 25.174 < 2.2e-16 ***
```

(g) O gráfico pedido é indicado em baixo, sendo a recta única para a totalidade das $n=600$ observações indicada a tracejado. O gráfico foi construído com os seguintes comandos do `R`:

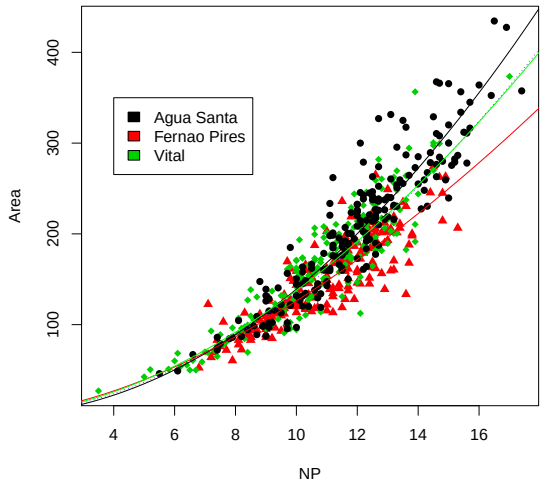
```
> plot(log(Area)~log(NP),col=as.numeric(Casta),pch=as.numeric(Casta)+15,data=videiras)
> abline(vid.Anc2.lm, col="blue", lty="dotted")
> abline(0.33820, 1.99648, col="black")
> abline(0.33820+0.62328, 1.99648-0.31298, col="red")
> abline(0.33820+0.39524, 1.99648-0.17654, col="green")
> legend(1.25,5.5, levels(videiras$Casta), fill=1:3)
```



Confirma-se o maior declive da recta associada à casta Água Santa, e o menor associado à casta Fernão Pires. Em comparação com a relação análoga estudada no Exercício 1, é visível uma maior distinção das três rectas ajustadas, que foi reflectida no facto de o teste F parcial ter considerado que o modelo ANCOVA e o modelo de regressão linear simples para as três castas em conjunto serem significativamente diferentes.

NOTA: Convém acrescentar que a signficância do teste F parcial resulta também do número bastante elevado de observações usado para ajustar estes modelos ($n = 600$). Quanto mais informação estiver disponível na amostra, mais facilmente as diferenças são consideradas significativas.

(h) O gráfico para as variáveis não logaritmizadas é o seguinte.



Foi produzido com os comandos:

```
> plot(Area ~ NP, col=as.numeric(Casta), pch=as.numeric(Casta)+15, data=videiras)
> curve(exp(0.5787)*x^(1.8764), from=0, to=18, col="blue", lty="dotted", add=TRUE)
> curve(exp(0.3382)*x^(1.9965), from=0, to=18, add=TRUE)
> curve(exp(0.3382+0.6233)*x^(1.9965-0.3130), from=0, to=18, col="red", add=TRUE)
> curve(exp(0.3382+0.3952)*x^(1.9965-0.1765), from=0, to=18, col="green", add=TRUE)
> legend(4,350, levels(videiras$Casta), fill=1:3)
```

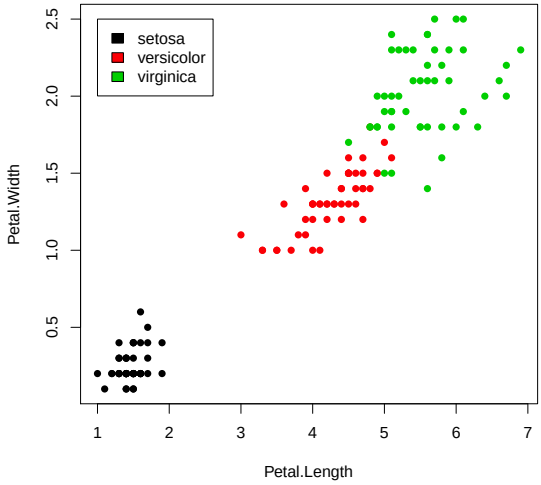
Nas escalas originais (não logaritmizadas) as diferenças entre as castas Água Santa e Fernão Pires é mais visível. A casta Vital tem um comportamento muito próximo do comportamento conjunto das três castas, sendo a sua curva ajustada quase indistinguível da curva única para as três castas (representada a ponteado).

3. FALTA

4. Neste exercício, consideram-se as $n = 150$ observações sobre lírios, com variável resposta dada pela largura das pétalas (variável `Petal.Width`) e preditor numérico comprimento das pétalas (`Petal.Length`). Será considerado também o factor espécie (`Species`), havendo $n_i = 50$ observações de cada espécie.

   (a) O gráfico pedido é obtido com os comandos seguintes. A nuvem é prometedora para uma relação linear global.

```
> plot(Petal.Width ~ Petal.Length, col=Species, data=iris, pch=16)
> legend(1,2.5, legend=levels(iris$Species), fill=1:3)
```



(b) Tem-se:

```
> iris.lm <- lm(Petal.Width ~ Petal.Length, data=iris)
> summary(iris.lm)
[...]
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.363076   0.039762  -9.131  4.7e-16 ***
Petal.Length  0.415755   0.009582  43.387  < 2e-16 ***
---
Residual standard error: 0.2065 on 148 degrees of freedom
Multiple R-squared:  0.9271,Adjusted R-squared:  0.9266
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

A recta $y = -0.363076 + 0.415755\,x$ explica quase 93% da variabilidade observada nas larguras das pétalas, para o conjunto das três espécies de lírios.

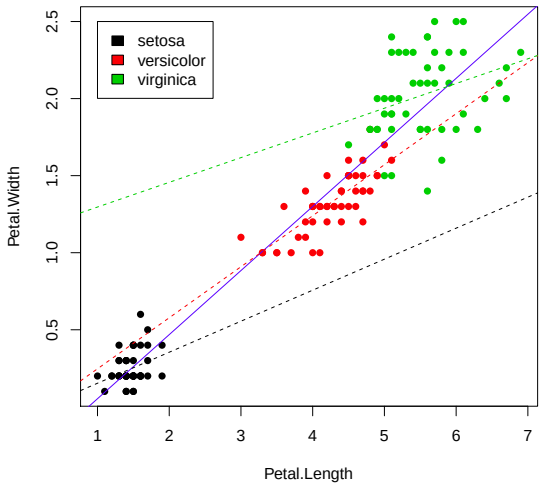
(c) O modelo completo, cruzando o preditor numérico `Petal.Length` com o factor `Species` é:

```
> irisSpecies.lm <- lm(Petal.Width ~ Petal.Length*Species, data=iris)
> summary(irisSpecies.lm)
[...]
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -0.04822    0.21472  -0.225 0.822627
Petal.Length                  0.20125    0.14586   1.380 0.169813
Speciesversicolor            -0.03607    0.31538  -0.114 0.909109
Speciesvirginica              1.18425    0.33417   3.544 0.000532 ***
Petal.Length:Speciesversicolor  0.12981    0.15550   0.835 0.405230
Petal.Length:Speciesvirginica  -0.04095    0.15291  -0.268 0.789244
---
Residual standard error: 0.1773 on 144 degrees of freedom
Multiple R-squared:  0.9477,Adjusted R-squared:  0.9459
F-statistic: 521.9 on 5 and 144 DF,  p-value: < 2.2e-16
```

Assim, as três rectas de regressão, para cada espécie individual são: $y = -0.04822 + 0.20125\,x$ para a espécie *setosa*, $y = -0.08429 + 0.33106\,x$ para a espécie *versicolor*, e $y = 1.1360 + 0.1603\,x$ para a espécie *virginica*. O valor do coeficiente de determinação do modelo ANCOVA, $R^2 = 0.9477$ é naturalmente maior do que o R^2 do submodelo constituído por uma única recta de regressão. Mas o seu valor não é de interpretação imediata, como se viu nas aulas e como se verá nas alíneas seguintes. Para traçar estas três rectas por espécie individual em cima da nuvem de pontos já anteriormente obtida, podem dar-se os seguintes comandos:

```
> coefIrisSpecies <- coef(irisSpecies.lm)
> abline(coefIrisSpecies[c(1,2)], col=1, lty=2)
> abline(coefIrisSpecies[c(1,2)]+coefIrisSpecies[c(3,5)], col=2, lty=2)
> abline(coefIrisSpecies[c(1,2)]+coefIrisSpecies[c(4,6)], col=3, lty=2)
```

Os resultados obtidos, juntamente com a recta única obtida para a totalidade das $n = 150$ observações (a azul, em traço contínuo), são indicados no gráfico seguinte.



Como se pode constatar, a situação é bem mais confusa do que no exercício 1, com duas das rectas (das espécies *setosa* e *virginica*) com declives bastante diferentes em relação aos da recta global e da recta da espécie *versicolor*. No entanto, as rectas das espécies *setosa* e *virginica* parecem ser aproximadamente paralelas, sendo os declives ajustados (0.20125 e 0.1603) próximos. No modelo completo discutido nas aulas, o declive da recta para a espécie de referência (*setosa*) é o parâmetro β_1. O declive da recta para a espécie *virginica* é a soma de β_1 com o acréscimo específico do declive da espécie *virginica*, ou seja, com o acréscimo $\alpha_{1:3}$. A hipótese de que essas duas rectas sejam paralelas corresponde à hipótese de $H_0 : \alpha_{1:3} = 0$. Esta hipótese corresponde a um teste a um parâmetro individual num modelo linear (ou seja, corresponde aos testes t usados na regressão linear para aferir possíveis valores de cada β_j). A informação necessária para efectuar esse teste está disponível na listagem de resultados obtida acima para o modelo `irisSpecies.lm`. Em particular, a estimativa desse acréscimo é -0.04095, com um erro padrão associado de $\hat{\sigma}_{\hat{\alpha}_{1:3}} = 0.15291$. Tendo em conta a hipótese nula referida, a estatística t do teste também é dada na listagem e tem valor $T_{calc} = -0.268$, a que corresponde um valor de prova $p = 0.789244$. Sendo assim, está-se muito longe de rejeitar a hipótese nula $H_0 : \alpha_{1:3} = 0$, para qualquer nível de significância usual. Assim, não se rejeita que essas duas rectas de espécie são paralelas.

5. Os três modelos individuais de espécie, ajustados apenas usando as $n_i = 50$ observações de cada espécie têm os coeficientes de determinação indicados de seguida:

```
> irisSetosa.lm <- lm(Petal.Width ~ Petal.Length, data=iris[1:50,])
> irisVersi.lm <- lm(Petal.Width ~ Petal.Length, data=iris[51:100,])
> irisVirgi.lm <- lm(Petal.Width ~ Petal.Length, data=iris[101:150,])
> summary(irisSetosa.lm)$r.sq
[1] 0.1099785
> summary(irisVersi.lm)$r.sq
[1] 0.6188467
> summary(irisVirgi.lm)$r.sq
[1] 0.1037537
```

   Assim, em todos os casos, estes R^2 por espécie individual são muito mais baixos que o R^2 global correspondente ao modelo ANCOVA completo. Como se discutiu nas aulas, tal facto corresponde a uma situação em que uma ANOVA da variável resposta `Petal.Width` sobre um único factor `Species` tem um valor elevado da Soma de Quadrados correspondente ao ajustamento do modelo, ou seja, SQF elevado. Por outras palavras, o valor elevado de $R^2 = 0.9477$ no modelo ANCOVA resulta do facto de ao factor espécie corresponderem larguras médias das pétalas bastante diferentes, e não tanto ao valor preditivo do preditor numérico `Petal.Length`. A tradução prática desse facto é visível na nuvem de pontos original, se repararmos que a forte relação linear global tem sobretudo a que ver com a separação entre os três grupos de observações correspondentes a cada espécie, e não tanto com relações lineares fortes entre as duas medições das pétalas no seio de cada espécie. Por outras palavras, a relação linear tão prometedora que parece existir entre largura e comprimento das pétalas, na nuvem da totalidade das $n = 150$ observações, é em certo sentido uma ilusão resultante de se ter considerado em conjunto as três espécies.

6. Nas aulas foi vista a fórmula que relaciona o valor de R^2 global do modelo ANCOVA com os R^2 e as Somas de Quadrados Totais para cada subconjunto de observações (por espécie), bem como o valor de SQF na ANOVA a um factor relacionando `Petal.Width` e o factor `Species`. A fórmula é

$$R^2 \;=\; \frac{\sum_{i=1}^{s} R_i^2\, SQT_i \;+\; SQF}{\sum_{i=1}^{s} SQT_i \;+\; SQF}.$$

   O valor de SQF pode obter-se da seguinte forma:

```
> summary(aov(Petal.Width ~ Species, data=iris))
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  80.41   40.21     960 <2e-16 ***
Residuals   147   6.16    0.04
```

   Por outro lado, os valores de SQT_i podem ser obtidos como o numerador das variâncias dos valores observados das larguras de pétalas em cada espécie. Tem-se $SQT_1 = 49 \times s^2_{y_1} = 0.5442$; $SQT_2 = 49 \times s^2_{y_2} = 1.9162$ e $SQT_3 = 49 \times s^2_{y_3} = 3.6962$. Logo,

$$R^2 = \frac{(0.1099785 \times 0.5442) + (0.6188467 \times 1.9162) + (0.1037537 \times 3.6962) + 80.41}{(0.5442 + 1.9162 + 3.6962) + 80.41} = 0.9477001\ .$$

Como se pode constatar, o valor de SQF sobrepõe-se ao das restantes parcelas, quer no numerador, quer no denominador, gerando um valor muito elevado do coeficiente de determinação global do modelo ANCOVA, que não corresponde a valores elevados de R^2 em nenhuma das regressões individuais de cada espécie. Confirma-se que a interpretação dos valores de R^2 em modelos ANCOVA deve ser feita com cuidado.