

INSTITUTO SUPERIOR DE AGRONOMIA
2011/12 – ESTATÍSTICA E DELINEAMENTO
ESTATÍSTICA APLICADA AO AMBIENTE
SEGUNDO TESTE

11 de Janeiro, 2012

Uma resolução possível

I

1. Estamos no contexto do modelo ANOVA a um factor (variedade de tomate) para a variável resposta (Y) peso médio por tomateiro em cada parcela.

(a) A tabela-resumo nos modelos a um factor com k níveis, n_i observações por nível (para um total de n observações), médias amostrais de nível \bar{y}_i , média global $\bar{y}_..$, variâncias amostrais de nível s_i^2 e variância global s_y^2 , tem o seguinte aspecto:

Variação	g.l.	SQs	QMs	F
Factor	$k - 1$	$SQF = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y}_..)^2$	$QMF = \frac{SQF}{k-1}$	$F = \frac{QMF}{QMRE}$
Residual	$n - k$	$SQRE = \sum_{i=1}^k (n_i - 1) s_i^2$	$QMRE = \frac{SQRE}{n-k}$	
Total	$n - 1$	$SQT = (n - 1) s_y^2$		

No nosso caso, tem-se $k = 5$, $n = 20$, com $n_i = 4$ repetições para os cinco níveis $i = 1, 2, 3, 4, 5$ (delineamento equilibrado). As médias e variâncias de nível e globais são dadas no enunciado. A Soma de Quadrados Total é $SQT = 19 \times 3529.416 = 67\,058.9$. A Soma de Quadrados Residual é $SQRE = 3 \sum_{i=1}^5 s_i^2 = 3 [520.2600 + 538.2667 + 1487.2867 + 788.5967 + 739.7967] = 12\,222.62$.

Logo, $SQF = SQT - SQRE = 67\,058.9 - 12\,222.62 = 54\,836.28$. Assim, $QMF = \frac{SQF}{k-1} = \frac{54\,836.28}{4} = 13\,709.07$ e $QMRE = \frac{SQRE}{n-k} = \frac{12\,222.62}{15} = 814.8413$. Finalmente, $F_{calc} = \frac{QMF}{QMRE} = \frac{13\,709.07}{814.8413} = 16.824$. A tabela-resumo vem:

Variação	g.l.	SQs	QMs	F
Factor	4	$SQF = 54\,836.28$	$QMF = 13\,709.07$	$F = 16.824$
Residual	15	$SQRE = 12\,222.62$	$QMRE = 814.8413$	
Total	19	$SQT = 67\,058.9$		

(b) Pedese um teste F aos efeitos do factor, que neste modelo corresponde a um teste à igualdade das médias populacionais de nível do factor, $\mu_i = \mu_1 + \alpha_i$.

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 $F_{calc} > f_{\alpha(k-1, n-k)} = f_{0.05(4, 15)} = 3.06$.

Conclusões: Como $F_{calc} = 16.824$, rejeita-se H_0 , o que corresponde a admitir a existência de efeitos do factor variedade no peso médio do fruto por tomateiro.

(c) Pedese para avaliar hipóteses do tipo $\mu_i = \mu_3$ (sendo μ_3 o peso médio populacional da variedade 40D) para $i \neq 3$. Vamos utilizar o teste de Tukey. Podemos rejeitar uma hipótese de igualdade dum par de médias populacional, $\mu_i = \mu_3$, caso as respectivas médias amostrais difiram por mais do que o termo de comparação do teste de Tukey, i.e., se

$$|\bar{y}_i - \bar{y}_3| > q_{\alpha(k, n-k)} \cdot \sqrt{\frac{QMRE}{n_c}} = q_{0.01(5, 15)} \cdot \sqrt{\frac{814.8413}{4}} = 5.56 \cdot \sqrt{\frac{814.8413}{4}} = 79.35628 .$$

Ora, $\bar{y}_3 = 189.40$. Assim, as médias de nível significativamente diferentes são as inferiores a $\bar{y}_3 - 79.356 = 189.40 - 79.356 = 110.044$, ou as superiores a $\bar{y}_3 + 79.356 = 189.40 + 79.356 = 268.756$. Da tabela de médias conclui-se que nenhum nível tem média amostral fora do intervalo $[110.044, 268.756]$, pelo que nenhuma variedade tem peso médio significativamente diferente do peso médio da variedade 40D.

- (d) Tendo em conta a restrição usual introduzida no modelo ANOVA a um factor ($\alpha_1 = 0$), os efeitos dos restantes níveis definem-se como $\alpha_i = \mu_i - \mu_1$ e estimam-se pela correspondente diferença de médias amostrais: $\hat{\alpha}_i = \bar{y}_i - \bar{y}_1$. No caso do nível 40C ($i = 2$), tem-se: $\hat{\alpha}_2 = \bar{y}_2 - \bar{y}_1 = 110.70 - 131.80 = -21.10$. Assim, estima-se que a variedade 40C tem, em média, menos 21.10g de frutos por planta do que a variedade 40B (a variedade $i = 1$).

2. Incorporam-se agora os efeitos dos terrenos diferentes, presentes no delineamento da experiência.

- (a) O modelo ANOVA adequado é um modelo a dois factores: **variedade** (Factor A, com $a = 5$ níveis) e **terreno** (Factor B, com $b = 4$ níveis). O delineamento é factorial completo, uma vez que existem observações (parcelas) correspondentes a todas as situações experimentais, isto é, parcelas correspondentes ao cruzamento de cada terreno com cada variedade. No entanto, apenas existe *uma* parcela para cada célula (há $n = 20$ observações e $ab = 20$ células). A inexistência de repetições não permite ajustar um modelo com efeitos de interacção. Assim, ajusta-se um modelo factorial a dois factores, *sem* efeitos de interacção:

- Y_{ij} indica a (única) observação da variedade i no terreno j (o terceiro índice é dispensável, uma vez que não existem repetições). A equação do modelo é $Y_{ij} = \mu_{11} + \alpha_i + \beta_j + \epsilon_{ij}$, onde μ_{11} indica a média populacional na célula (1, 1); α_i indica o efeito da variedade i ; β_j indica o efeito do terreno j e ϵ_{ij} indica o erro aleatório associado à observação Y_{ij} . Impõem-se as restrições $\alpha_1 = 0$ e $\beta_1 = 0$.
- Os erros aleatórios ϵ_{ij} são v.a.s independentes com distribuição $\epsilon_{ij} \cap \mathcal{N}(0, \sigma^2)$, $\forall i, j$.

- (b) A tabela-resumo terá mais uma linha do que a do ponto anterior, associada aos efeitos de terreno (Factor B). No entanto, há vários aspectos em que as tabelas se relacionam. Uma vez que as $n = 20$ observações são as mesmas, a sua variância s_y^2 é igual, pelo que a Soma de Quadrados Total também é igual: $SQT = 67\,058.9$. O enunciado informa que o novo Quadrado Médio Residual é $QMRE = 941$. Como neste modelo há $a+b-1 = 8$ parâmetros, os graus de liberdade associados a $QMRE$ são $n - (a+b-1) = 12$, pelo que a Soma de Quadrados Residual é agora $SQRE = QMRE \times [n - (a+b-1)] = 941 \times 12 = 11\,292$. Sabemos que a nova Soma de Quadrados associada ao factor B obtém-se pela diferença entre a Soma de Quadrados Residual do modelo que não previa os efeitos do factor B, e a Soma de Quadrados Residual do modelo com este novo tipo de efeitos, ou seja, $SQB = SQRE_A - SQRE_{A+B} = 12\,222.62 - 11\,292 = 930.62$. Os graus de liberdade que lhe estão associados são $b-1 = 3$, pelo que o Quadrado Médio associado ao Factor B é $QMB = \frac{SQB}{b-1} = \frac{930.62}{3} = 310.21$. Finalmente, SQA mantém-se igual ao SQF do modelo em que apenas se previa os efeitos do factor A, ou seja, $SQA = 54\,836.28$. Uma vez que os respectivos graus de liberdade também não se alteram, QMA mantém o valor do QMF no modelo a um factor. Finalmente, as duas estatística F associadas aos testes à existência dos dois tipos de efeitos previstos no modelo, são dados por $F_A = \frac{QMA}{QMRE} = \frac{13\,709.07}{941} = 14.569$ e $F_B = \frac{QMB}{QMRE} = \frac{310.21}{941} = 0.3297$. Assim, a tabela-resumo é:

Variacão	g.l.	SQs	QMs	F
Factor A	4	$SQA = 54\,836.28$	$QMA = 13\,709.07$	$F_A = 14.569$
Factor B	3	$SQB = 930.62$	$QMB = 310.21$	$F_B = 0.3297$
Residual	12	$SQRE = 11\,292$	$QMRE = 941$	
Total	19	$SQT = 67\,058.9$		

- (c) É pedido um teste aos efeitos do factor B, ou seja, aos efeitos do factor **terreno**. Tem-se:

Hipóteses: $H_0 : \beta_j = 0, \forall i = 2, 3, 4$ vs. $H_1 : \exists i = 2, 3, 4$ 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 $F_{calc} > f_{0.05(3,12)} = 3.49$.

Conclusões: Como $F_{calc} = 0.3297$, não se rejeita H_0 , o que corresponde a admitir a inexistência de efeitos do factor terreno. Esta conclusão (resultante de $SQB \ll SQRE$, o que significa que pouca da variabilidade residual pode ser explicada pela existência de diferentes terrenos) indica que os terrenos são relativamente homogêneos, não tendo uma contribuição importante para a variabilidade nos pesos médios observados.

NOTA: Neste contexto, os terrenos podem ser considerados blocos introduzidos no delineamento por se suspeitar de heterogeneidade das unidades experimentais. Este exemplo ilustra que, caso não haja heterogeneidade das unidades experimentais correspondentes ao novo factor/blocos, a sua inclusão pode até diminuir o valor da estatística F , o que tende a torná-la menos significativa. De facto, no modelo a um factor, o valor de prova (*p-value*) de $F_{calc} = 16.825$ é $p = 0.000020$; no modelo a 2 factores o *p-value* de $F_A = 14.569$ é maior: $p = 0.000148$. Assim, a existência do segundo factor (blocos) com efeitos pouco importantes contribui para esconder um pouco a significância de efeitos do factor que se deseja estudar (lembre-se que os dados são iguais nos dois casos).

II

Trata-se dum delineamento factorial a dois factores: `clone` (factor A, com $a = 5$ níveis) e `local` (factor B, igualmente com $b = 5$ níveis). O delineamento é factorial completo (uma vez que há observações sobre todos os clones, em todas as localidades) e equilibrado, havendo $n_c = 4$ parcelas (repetições) para cada combinação de clone e localidade. A variável resposta é o rendimento.

1. Uma vez que $n_c > 1$ pode ajustar-se um modelo ANOVA com interacção. A linha correspondente à interacção clone/localidade na tabela-resumo confirma ter sido esse o modelo ajustado.

(a) O modelo é dado pela equação $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, ($i = 1, \dots, 5$; $j = 1, \dots, 5$; $k = 1, 2, 3, 4$), sendo

- Y_{ijk} a observação da k -ésima parcela com o clone i , na localidade j ;
- μ_{11} a média populacional do clone AR36 em Alenquer (admitindo que a ordem dos níveis é a indicada na tabela das médias, correspondente à ordem alfabética usada no programa R);
- α_i é o efeito principal do clone i ;
- β_j é o efeito principal da localidade j ;
- $(\alpha\beta)_{ij}$ é o efeito de interacção do clone i com a localidade j ;
- ϵ_{ijk} é o erro aleatório associado à observação Y_{ijk} .

Como é hábito, consideramos as restrições $\alpha_1 = \beta_1 = 0$, e ainda $(\alpha\beta)_{ij} = 0$ se $i = 1$ e/ou $j = 1$. A descrição do modelo fica completa com os pressupostos sobre os erros aleatórios:

- $\epsilon_{ijk} \cap \mathcal{N}(0, \sigma^2)$, $\forall i, j, k$;
- $\{\epsilon_{ijk}\}_{i,j,k}$ é um conjunto de variáveis aleatórias independentes.

(b) Os quatro graus de liberdade omissos são:

- Factor A (clones) : $a - 1 = 4$;
- Factor B (localidades) : $b - 1 = 4$;

- Interacção : $(a - 1)(b - 1) = 16$;
- Residual : $n - ab = 100 - 25 = 75$.

O Quadrado Médio em falta é $QMAB = \frac{SQAB}{(a-1)(b-1)} = \frac{17.300}{16} = 1.08125$. Finalmente, o valor calculado da estatística F em falta é $F_A = \frac{QMA}{QMRE} = \frac{1.1834}{0.2237} = 5.29$.

- (c) O valor ajustado para qualquer observação numa dada célula, no modelo factorial a dois factores com efeitos de interacção, é a média amostral da respectiva célula. Assim, $\hat{y}_{45k} = \bar{y}_{45} = 1.570$, para qualquer das $n_c = 4$ observações correspondentes ao clone AR39 em Setúbal (célula (4, 5)).
- (d) Pedese para indicar o valor da variância amostral s_y^2 da totalidade das $n = 100$ observações do rendimento (variável resposta y). Tem-se:

$$s_y^2 = \frac{SQT}{n - 1} = \frac{SQA + SQB + SQAB + SQRE}{n - 1} = 1.0827 .$$

2. Pedese para efectuar os três testes F aos três tipos de efeitos previstos pelo modelo. Façamos em pormenor o teste à interacção:

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

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 $F_{calc} > f_{\alpha[(a-1)(b-1), n-ab]} = f_{0.05(16,75)} = 1.78$ (entre os valores tabelados 1.66 e 1.84).

Conclusões: Como $F_{calc} = 4.8326 > 1.78$, rejeita-se H_0 , o que corresponde a admitir a existência de efeitos de interacção entre clones e localidades.

No teste aos efeitos principais do factor **clone** (factor A) as hipóteses são $H_0 : \alpha_i = 0, \forall i > 1$ vs. $H_1 : \exists i$ tal que $\alpha_i \neq 0$. Rejeita-se H_0 se $F_{calc} > f_{\alpha(a-1, n-ab)} = f_{0.05(4,75)} = 2.49$ (entre os valores tabelados 2.45 e 2.53). Como $F_{calc} = 5.29$, rejeita-se H_0 , concluindo-se pela existência de efeitos principais do factor clone.

No teste aos efeitos principais do factor **local** (factor B) as hipóteses são $H_0 : \beta_j = 0, \forall j > 1$ vs. $H_1 : \exists j$ tal que $\beta_j \neq 0$. Rejeita-se H_0 se $F_{calc} > f_{\alpha(b-1, n-ab)} = f_{0.05(4,75)} = 2.49$ (valor igual ao do teste anterior). Como $F_{calc} = 76.4016$, rejeita-se claramente H_0 , concluindo-se pela existência de efeitos principais do factor localidade.

Assim, conclui-se pela existência dos três tipos de efeitos previstos no modelo, com destaque para os efeitos principais do factor **local**.

3. Trata-se dum gráfico de interacção, em que no eixo horizontal se indicam os cinco níveis do factor **local** (sem qualquer efeito de escala, dada a natureza categórica do factor) e no eixo vertical se indicam valores da variável resposta numérica. Por cima dos marcadores de cada local existem 4 pontos, tantos quantos os níveis do segundo factor (**clone**). A sua altura é dada pela média amostral da variável resposta na célula combinando os referidos clone e local. Caso não existam efeitos de interacção, os traços seccionalmente lineares que unem os pontos dum mesmo clone serão aproximadamente paralelos. Aqui, isso não se verifica, em particular devido à célula Felgueiras/AR38, cujo rendimento é muito inferior aos das restantes células de Felgueiras, como se confirma pela tabela das médias de célula. Também de assinalar o grande “pico” correspondente às restantes células de Felgueiras, que deve ser responsável pela conclusão, na alínea anterior, de que os efeitos principais de localidade são muito significativos.

4. Comparemos as médias de células relativas a Felgueiras e outras médias de célula (com excepção das relativas ao clone AR38). O termo de comparação do teste de Tukey é $q_{\alpha(ab, n-ab)} \cdot \sqrt{\frac{QMRE}{n_c}} = q_{0.05(25,75)} \cdot \sqrt{\frac{0.2337}{4}} \approx 1.252$ (o valor tabelado mais próximo é $q_{0.05(20,80)} = 5.18$, próximo do verdadeiro valor $q_{0.05(25,75)} = 5.372$). O rendimento médio em Felgueiras, para qualquer casta excepto AR38 é sempre superior a 3.775. Este rendimento é significativamente superior a qualquer rendimento que não exceda $3.775 - 1.252 = 2.523$ (com o valor exacto da distribuição de Tukey seria a 2.476). Nenhuma outra média amostral de célula atinge esse rendimento, pelo que a afirmação do enunciado é legítima, ao nível de significância $\alpha = 0.05$.

III

1. Está-se no contexto dum modelo ANOVA a um factor, com k níveis.

- (a) A matriz do modelo \mathbf{X} tem n linhas (uma para cada observação) e k colunas, sendo a primeira o vector $\mathbf{1}_n$ de n uns; a segunda a indicatriz \mathcal{I}_2 das observações correspondentes ao segundo nível do factor (cujos valores são 1 para observações desse nível e 0 para as restantes); a terceira a indicatriz \mathcal{I}_3 das observações correspondentes ao terceiro nível do factor; e por aí fora, até à última (k -ésima) coluna, que é a indicatriz \mathcal{I}_k das observações no nível k do factor.
- (b) Os vectores do subespaço $\mathcal{C}(X)$ são as possíveis combinações lineares das colunas da matriz \mathbf{X} , ou seja, os vectores da forma $a_1 \mathbf{1}_n + a_2 \mathcal{I}_2 + a_3 \mathcal{I}_3 + \dots + a_k \mathcal{I}_k$. Tendo em conta a natureza das colunas de \mathbf{X} (descrita na alínea anterior), estas combinações lineares resultam sempre em vectores para os quais:
- os n_1 elementos correspondentes a observações do primeiro nível do factor têm um valor comum: a_1 ;
 - os n_2 elementos correspondentes a observações do segundo nível do factor têm um valor comum: $a_1 + a_2$;
 - em geral, os n_i elementos correspondentes a observações do i -ésimo nível do factor têm um valor comum: $a_1 + a_i$ ($i > 1$).

Assim, os vectores do subespaço $\mathcal{C}(X)$ são vectores em que os elementos associados a um mesmo nível do factor são iguais.

- (c) Sabemos que para este modelo ANOVA, a estatística do teste aos efeitos do factor é $F = \frac{QMF}{QMRE} = \frac{SQF/(k-1)}{SQRE/(n-k)}$. Pelo formulário, sabemos que $SQF = \sum_{i=1}^k n_i (\bar{y}_{i..} - \bar{y}_{...})^2 = n_c \sum_{i=1}^k (\bar{y}_{i..} - \bar{y}_{...})^2$, e $SQRE = \sum_{i=1}^k (n_i - 1) s_i^2 = (n_c - 1) \sum_{i=1}^k s_i^2$. Assim, sendo $V = \frac{1}{k-1} \sum_{i=1}^k (\bar{y}_{i..} - \bar{y}_{...})^2$ a variância das k médias amostrais de nível, $\bar{y}_{i..}$ e $M = \frac{1}{k} \sum_{i=1}^k s_i^2$ a média das k variâncias amostrais de nível s_i^2 , tem-se:

$$F = \frac{SQF/(k-1)}{SQRE/(n-k)} = \frac{n_c \cdot \frac{1}{k-1} \sum_{i=1}^k (\bar{y}_{i..} - \bar{y}_{...})^2}{\frac{n_c-1}{n-k} \sum_{i=1}^k s_i^2} = \frac{n_c \cdot V}{\frac{n_c-1}{k} \sum_{i=1}^k s_i^2} = n_c \frac{V}{M},$$

- (d) A hipótese nula do teste aos efeitos do factor pode ser escrita na forma $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$, ou seja, afirma a igualdade das médias populacionais de nível. É natural rejeitar esta hipótese

apenas se as correspondentes médias amostrais de nível forem muito diferentes entre si. Ora, a estatística do teste tem no seu numerador a variância dessas médias amostrais de nível, pelo que médias amostrais de nível muito diferentes entre si geram valores grandes da estatística. Assim a opção por uma região crítica unilateral direita é natural.

2. Num delineamento hierarquizado, o factor subordinado B pode ter um número diferente de níveis b_i para cada nível i do factor dominante, mas neste caso tal não se verifica: há sempre $b_i = b$ níveis, para qualquer i . Assim, há ao todo ab diferentes situações experimentais (correspondentes às “folhas” terminais na representação do delineamento em dendrograma).

(a) O modelo ANOVA correspondente tem a seguinte natureza.

- A k -ésima observação no nível j do factor B, subordinado ao nível i do factor A é representado por Y_{ijk} . A equação do modelo é $Y_{ijk} = \mu_{11} + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$, onde μ_{11} indica a média populacional na célula correspondente ao primeiro nível de B sob o primeiro nível de A; α_i indica um efeito do nível i do factor dominante; e $\beta_{j(i)}$ indica o efeito do nível j do factor subordinado, no nível i do factor A. Considera-se $\alpha_1 = 0$ e $\beta_{1(i)} = 0$, $\forall i$.
- os erros aleatórios têm distribuição $\epsilon_{ijk} \cap \mathcal{N}(0, \sigma^2)$, $\forall i, j, k$.
- $\{\epsilon_{ijk}\}_{ijk}$ são variáveis aleatórias independentes.

(b) Existe 1 parâmetro μ_{11} . Quanto aos efeitos de nível do factor A, e tendo em conta a restrição $\alpha_1 = 0$, existem $a - 1$ parâmetros de tipo α_i . Finalmente, nos efeitos do factor B, e tendo em conta as restrições $\beta_{1(i)} = 0$, $\forall i$, há $b - 1$ efeitos não nulos de B, para cada nível do factor A, num total de $a(b - 1)$ efeitos de tipo $\beta_{j(i)}$. Somando, tem-se um total de $1 + (a - 1) + a(b - 1) = a + a(b - 1) = a(1 + b - 1) = ab$ parâmetros no modelo, tantos quantas as situações experimentais (tal como no modelo para delineamentos factoriais a dois factores, com interacção).

(c) Como para qualquer modelo linear, a Soma de Quadrados Residual obtém-se aqui pelos seguintes passos: cria-se a matriz do modelo \mathbf{X} , constituída por uma coluna de uns; $a - 1$ colunas de indicatrizes dos níveis do factor A associados aos efeitos α_i previstos; e finalmente $a(b - 1)$ colunas indicatrizes dos níveis com efeitos do factor B. Calcula-se a matriz de projecção ortogonal sobre o subespaço $\mathcal{C}(\mathbf{X})$ gerado pelas colunas de \mathbf{X} , i.e., a matriz $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$. Multiplicando \mathbf{H} pelo vector \mathbf{Y} dos valores observados da variável resposta obtem-se o vector dos valores ajustados do modelo, $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$. A diferença destes dois vectores é o vector dos resíduos $\mathbf{E} = \mathbf{Y} - \hat{\mathbf{Y}}$ (cujo i -ésimo elemento é $e_i = y_i - \hat{y}_i$). A norma ao quadrado do vector \mathbf{E} é a Soma de Quadrados Residual deste modelo, que podemos representar por $SQRE_{A/B}$.

Para construir a Soma de Quadrados associada aos efeitos do factor B, $SQB(A)$, tomamos a diferença entre a Soma de Quadrados Residual do modelo que apenas prevê os efeitos do Factor A, $SQRE_A$, e a Soma de Quadrados Residual do modelo hierarquizado: $SQB(A) = SQRE_A - SQRE_{A/B}$, diferença associada aos efeitos do factor B uma vez que representa a redução na variabilidade residual (não explicada pelo modelo) resultante da introdução dos efeitos desse factor. Finalmente, a Soma de Quadrados associada ao factor A define-se como a Soma de Quadrados do Factor no modelo que apenas prevê a existência do factor A: $SQA = SQF_A$. Somando estas três Somas de Quadrados obtem-se:

$$\begin{aligned} SQRE_{A/B} + SQB(A) + SQA &= \cancel{SQRE_{A/B}} + (SQRE_A - \cancel{SQRE_{A/B}}) + SQF_A \\ &= SQRE_A + SQF_A = SQT, \end{aligned}$$

sendo a igualdade final justificada por se tratar da decomposição de SQT no modelo apenas com o factor A.