15.2.2 Estimação do Modelo

Nosso objetivo é estimar μ_1, μ_2 e σ_e^2 no modelo (15.6), para podermos testar H_0. Usaremos estimadores de mínimos quadrados. Poderíamos usar também estimadores de máxima verossimilhança, pois sabemos que nossas observações têm distribuição normal. Temos que, de (15.6), os *resíduos* são dados por

$$e_{ij} = y_{ij} - \mu_i, \tag{15.9}$$

e a soma dos quadrados dos resíduos é dada por

$$SQ(\mu_1,\mu_2) = \sum_{i=1}^{2}\sum_{j=1}^{n_i} e_{ij}^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2$$

$$= \sum_{j=1}^{n_1} (y_{1j} - \mu_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \mu_2)^2,$$

ou seja,

$$SQ(\mu_1,\mu_2) = \sum_{j=1}^{n_1} e_{1j}^2 + \sum_{j=1}^{n_2} e_{2j}^2. \tag{15.10}$$

Derivando (15.10) em relação a μ_1 e μ_2 obtemos:

$$\frac{\partial SQ(\mu_1, \mu_2)}{\partial \mu_1} = -2\sum_{j=1}^{n_i}(y_{1j} - \mu_i) = 0, \quad i = 1, 2,$$

do que segue que os estimadores são dados por

$$\hat{\mu}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} y_{1j} = \overline{y}_{1,}, \tag{15.11}$$

$$\hat{\mu}_2 = \frac{1}{n_2}\sum_{j=1}^{n2} y_{2j} = \overline{y}_{2,}, \tag{15.12}$$

que são as médias das observações dos níveis 1 e 2, respectivamente. Logo,

$$SQ(\hat{\mu}_1,\hat{\mu}_2) = \sum_{j=1}^{n_1} (y_{1j} - \overline{y}_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \overline{y}_2)^2. \tag{15.13}$$

Podemos pensar em (15.13) como a *quantidade total de informação quadrática perdida* pela adoção do modelo (15.6). Essa soma é também denominada *soma dos quadrados dos resíduos*.

Vejamos outra maneira de escrever essa soma. Dentro do grupo dos homens, a variância da subpopulação P_1 pode ser estimada por

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(y_{1j} - \overline{y}_1)^2, \tag{15.14}$$

e a variância da subpopulação P_2 das mulheres é estimada por

$$S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(y_{2j} - \overline{y}_2)^2. \tag{15.15}$$

Segue-se que

$$SQ(\hat{\mu}_1, \hat{\mu}_2) = (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2. \tag{15.16}$$

Temos, acima, dois estimadores não-viesados do mesmo parâmetro σ_e^2 e, portanto, podemos definir uma variância amostral ponderada

$$S_e^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}, \tag{15.17}$$

e, usando (15.16), podemos escrever

$$S_e^2 = \frac{\mathrm{SQ}(\hat{\mu}_1, \hat{\mu}_2)}{n - 2}, \tag{15.18}$$

Exemplo 15.1. (continuação) Para os dados da Tabela 15.1, temos:

Grupo dos Homens (nível 1): $\overline{y}_1 = 110{,}1$, $\sum_{j=1}^{10}(y_{1j} - \overline{y}_1)^2 = 670{,}9$, $S_1^2 = 74{,}54$;

Grupo das Mulheres (nível 2): $\overline{y}_2 = 104{,}9$, $\sum_{j=1}^{10}(y_{2j} - \overline{y}_2)^2 = 566{,}9$, $S_2^2 = 62{,}99$.

Segue-se que

$$S_e^2 = \frac{670{,}9 + 566{,}9}{18} = \frac{1.237{,}8}{18} = 68{,}77, \quad S_e = 8{,}29.$$

Note que a soma dos quadrados dos resíduos é

$$SQ(\hat{\mu}_1, \hat{\mu}_2) = \mathrm{SQ}(\overline{y}_1, \overline{y}_2) = 1.237{,}8.$$

15.2.3 Intervalos de Confiança

Com as suposições feitas sobre os erros, podemos escrever

$$\bar{y}_1 \sim N(\mu_1, \sigma_e^2 / n_1),\ \bar{y}_2 \sim N(\mu_2, \sigma_e^2 / n_2), \tag{15.23}$$

o que permite construir intervalos de confiança separados para os dois parâmetros μ_1 e μ_2, como já vimos anteriormente. Esses têm a forma

$$\bar{y}_i \pm t_\gamma \frac{S_e}{\sqrt{n_i}}, \quad i = 1, 2, \tag{15.24}$$

onde t_γ é o valor crítico da distribuição t de Student com $\nu = n - 2$ graus de liberdade, tal que $P(-t_\gamma < t(n-2) < t_\gamma) = \gamma$, $0 < \gamma < 1$. Observe que o número de graus de liberdade é $(n - 2)$ e não $n_i - 1$, porque

$$Z_i = \frac{(\bar{y}_i - \mu_i)\sqrt{n_i}}{\sigma_e} \sim N(0,1),$$

$$W = \frac{(n-2)S_e^2}{\sigma_e^2} \sim \chi^2(n-2)$$

e, portanto, $\dfrac{Z_i}{\sqrt{W/(n-2)}} = \dfrac{\sqrt{n_i}(\bar{y}_i - \mu_i)}{S_e}$ tem distribuição $t(n - 2)$ pelo Teorema 7.1. Daqui, obtemos (15.24).

Exemplo 15.1. (continuação) Para o Exemplo 15.1, temos:

$$\text{IC}(\mu_1;\ 0{,}95) = 110{,}10 \pm (2{,}101)8{,}29 / \sqrt{10} =]104{,}59;\ 115{,}61[,$$

$$\text{IC}(\mu_2;\ 0{,}95) = 104{,}90 \pm (2{,}101)8{,}29 / \sqrt{10} =]99{,}39;\ 110{,}41[,$$

com $t_{0{,}95} = 2{,}101$ encontrado na Tabela V, com $\nu = 18$ graus de liberdade.

Ainda, com as suposições feitas, podemos concluir que

$$\bar{y}_1 - \bar{y}_2 \sim N(\mu_1 - \mu_2,\ \sigma_e^2 / n_1 + \sigma_e^2 / n_2), \tag{15.25}$$

de modo que a estatística

$$T = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{S_e\sqrt{1/n_1 + 1/n_2)}} \tag{15.26}$$

tem distribuição t de Student com $\nu = n_1 + n_2 - 2 = n - 2$ graus de liberdade, e um intervalo de confiança para a diferença $\mu_1 - \mu_2$ pode ser construído.

Exemplo 15.1. (continuação) Para o exemplo,

$$\text{IC}(\mu_1 - \mu_2;\ 0{,}95) = (\overline{y}_1 - \overline{y}_2) \pm t_y S_e \sqrt{1/n_1 + 1/n_2}$$

$$= (110{,}1 - 104{,}9) \pm (2{,}101)(8{,}29)\sqrt{1/10 + 1/10} =]-2{,}59;\ 12{,}99[.$$

15.2.4 Tabela de Análise de Variância

As operações processadas anteriormente podem ser resumidas num quadro, para facilitar a análise. Se (15.27) for válida, o modelo adotado será

$$y_{ij} = \mu + e_{ij},$$

e a quantidade de informação perdida (devida aos resíduos) será dada por

$$SQ(\hat{\mu}) = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \overline{y})^2, \tag{15.28}$$

que iremos chamar de *soma de quadrados total*, abreviadamente, SQTot.

Analogamente, adotado o modelo (15.4), a quantidade de informação perdida é dada por (15.13) ou (15.16), e que chamamos de *soma de quadrados dos resíduos*, abreviadamente, SQRes, ou *soma de quadrados dentro dos dois grupos*, abreviadamente, SQDen.

A *economia* obtida ao passarmos de um modelo para outro será

$$\text{SQTot} - \text{SQDen} = \text{SQEnt}, \tag{15.29}$$

que chamaremos de *soma de quadrados entre grupos*. Não é difícil provar que (veja o problema 18)

$$\text{SQEnt} = \sum_{i=1}^{2} n_i (\overline{y}_i - \overline{y})^2. \tag{15.30}$$

Observando essa expressão, vemos que ela representa a variabilidade *entre as médias amostrais*, ou seja, uma "distância" entre a média de cada grupo e a média global. Donde o nome "soma de quadrados entre grupos". Quanto mais diferentes forem as médias $\overline{y}_i$, $i = 1, 2$, maior será SQEnt e, conseqüentemente, menor será SQDen.

As quantidades

$$\text{QMTot} = \frac{\text{SQTot}}{n-1} \tag{15.31}$$

e

$$\text{QMDen} = \frac{\text{SQDen}}{n-2} \tag{15.32}$$

são chamadas *quadrado médio total* e *quadrado médio dentro* (ou residual), respectivamente.

Todas essas informações são agrupadas numa única tabela, conhecida pelo nome de ANOVA (abreviação de ANalysis Of VAriance), descrita na Tabela 15.5.

Tabela 15.5: Tabela de Análise de Variância (ANOVA).

F.V.	g.l.	SQ	QM	F
Entre	1	SQEnt	QMEnt	QMEnt/S,²
Dentro	$n-2$	SQDen	QMDen (ou S_e^2)	
Total	$n-1$	SQTot	QMTot (ou S^2)	

Na primeira coluna temos as descrições das diferentes somas de quadrados, tecnicamente indicadas por fontes de variação (F.V.). Os graus de liberdade (g.l.) da segunda coluna estão associados às respectivas somas de quadrados, sendo que o número de g.l. da SQE é obtido por subtração. Falaremos abaixo sobre QMEnt e a razão F = QMEnt/QMDen.

Exemplo 15.1. (continuação) Com os dados obtidos anteriormente para o Exemplo 15.1, podemos construir a tabela ANOVA para o modelo (15.4). O resultado está na Tabela 15.6.

Tabela 15.6: Tabela ANOVA para o Exemplo 15.1.

F.V.	g.l.	SQ	QM	F
Entre	1	135,20	135,20	1,97
Dentro	18	1.237,80	68,77	
Total	19	1.373,00	72,26	

Da ANOVA encontramos os desvios padrões residuais $S_e = \sqrt{68,77} = 8,29$ do "modelo completo" (15.4) e $S = \sqrt{72,26} = 8,50$, do "modelo reduzido" (15.19). A economia propiciada ao passar de um modelo para outro, em termos de soma de quadrados, é 135,20, e em termos de quadrados médios, comparando 72,26 e 68,77. Proporcionalmente, economizamos

$$\frac{135,20}{1.373,00} = 0,0985 \approx 9,85\%,$$

ou seja, aproximadamente 10% na SQ de resíduos. Podemos dizer que essa é *a proporção da variação explicada pelo modelo* (15.9). Essa medida é chamada *coeficiente de explicação* do modelo, denotada por

$$R^2 = \frac{\text{SQEnt}}{\text{SQTot}}. \tag{15.33}$$