Capítulo 4

Modelo de análisis de la varianza

4.1. Introducción

En este capítulo se aplica la teoría del modelo lineal para comparar las medias de una variable real Y obtenidas en varias poblaciones. Las distintas poblaciones se definen según uno o más caracteres, que son variables cualitativas y que se llaman factores cuando son variables controladas y un modelo con p factores se llama clasificación a p entradas. Cada factor toma un número finito de alternativas llamadas niveles, categorías o modalidades. En un modelo de clasificación a p entradas, una población está definida entonces por una p-tupla de categorías, una de cada factor. El cruce de todas las categorías de los factores define celdas; a cada celda le corresponde entonces una población.

Por ejemplo se considera la cosecha anual de trigo según el tipo de suelo; el tipo de suelo es un factor que toma, por ejemplo, 4 categorías. Tenemos entonces 4 celdas o poblaciones de medias μ_1, μ_2, μ_3 y μ_4. En las celdas tenemos n_j ($j = 1, \ldots, 4$) observaciones de la cosecha anual y queremos saber si hay diferencia entre las cosechas debido al tipo de suelo. Tendremos un segundo factor si agregamos 3 tipos de fertilizante para distinguir las cosechas. En este caso tenemos 12 celdas o poblaciones de media μ_{ij} ($i = 1,2,3$; $j = 1,2,3,4$). Nos preguntamos ahora cuando hay diferencia entre las cosechas, si se debe al tipo de suelo o al tipo de fertilizante o a ambos. Se habla de los efectos de los factores.

El tratamiento del modelo para medir los efectos de los factores será distinto si los cuatro tipos de suelos constituyen una muestra representativa de todos los tipos de suelos posibles o no. Si no hay otros tipos de suelo posibles que los cuatro considerados, se habla de efecto, mientras que en el otro caso se habla de efectos aleatorios

Desarrollemos la teoría de estos modelos según el número de factores y separaremos el estudio del modelo a efectos fijos del modelo a efectos aleatorios. Antes mostraremos algunos tests para verificar los supuestos usuales de homogeneidad de las varianzas de los errores. Cabe notar que este supuesto de homogeneidad de

las varianzas es menos critico cuando los datos son casi balanceados.

4.2. Homogeneidad de las varianzas

Sean a poblaciones y σ_i^2 la varianza y n_i el tamaño de la muestra de la población i ($n = \sum n_i$). Sea y_{ik} los valores muestrales ($i = 1,...,a; k = 1,...,n_i$).

Un primer test consiste en extender el test F de comparación de dos varianzas. Es el test de Bartlett

$$H_o : \sigma_i^2 = \sigma^2, \;\; i = 1,...,a$$

La varianza intra-grupo es: $w = \frac{1}{n}\sum(n_i - 1)s_i^2$ donde $s_i^2 = \frac{1}{n_i - 1}\sum_k (y_{ik} - \bar{y}_{i\bullet})$ es la varianza en la muestra de la población i.

Definamos

$$M = \sum(n_i - 1)log(w) - \sum(n_i - 1)log(s_i^2)$$

$$C = 1 + \frac{\sum \frac{1}{n_i - 1} - \frac{1}{n - p}}{3(a - 1)}$$

Bajo H_o, la cantidad M/C tiene una distribución aproximada χ^2_{a-1}.

Este test es muy sensibles al supuesto de normalidad, especialmente cuando las distribuciones tienes colas pesadas. Un test que hace menos sensible a esta forma de distribución es el test de Levene:

Se define

$$w_i = |y_{ik} - \bar{y}_{i\bullet}|n_i$$

El estadístico de Levene es:

$$\bar{w} = \frac{\sum_i (n_i - 1)w_i}{n - p}$$

4.3. Clasificación a un factor

4.3.1. Modelo a efectos fijos

Este problema fue resuelto en el análisis de perfiles (párrafo 4.4) con la segunda pregunta relativa a la igualdad de los niveles. Pero aquí queremos plantear el problema de otra forma: a partir de un modelo lineal.

Sea A el factor que toman q categorías llamadas $A_1, A_2, \ldots, A_q$ e y_{ij} ($i = 1, 2, \ldots, q;\ j = 1, 2, \ldots, n_i$) los valores obtenidos sobre la variable respuesta Y relativos a observaciones provenientes de diferentes categorías.

Para estudiar si hay homogeneidad dentro de las categorías comparando las diferencias de una categoría a otra, podemos escribir el podemos escribir el modelo lineal:

$$y_{ij} = \mu_i + \varepsilon_{ij} \qquad (i = 1,2,\ldots,q; \quad j = 1,2,\ldots,n_i)$$

en donde los μ_i son los efectos del factor y los ε_{ij} son los errores asociados al modelo.

El estimador de los mínimos cuadrados de los efectos es: $\hat{\mu}_i = \overline{y_i}$.

Los efectos pueden descomponerse también como: $\forall i : \mu_i = \mu + \alpha_i$. En este caso el modelo es de rango incompleto. Hay que imponer una restricción. Por ejemplo, con la restricción $\sum_{i=1}^{p} n_i \alpha_i = 0$, se obtiene $\hat{\mu} = \overline{y}$, $\hat{\alpha}_i = \overline{y_i} - \overline{y}$.

Sean $\overline{y_1}, \overline{y_2}, \ldots, \overline{y_q}$ las medias por categorías e $\overline{y} = \frac{1}{n}\sum_i n_i \overline{y_i}$ la media total. Llamamos

$$T = \sum_{i=1}^{q}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2, \qquad B = \sum_{i=1}^{q} n_i(\overline{y_i} - \overline{y})^2, \qquad W = \sum_{i=1}^{q}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_i})^2$$

Con la hipótesis de normalidad usual sobre los errores, el error cuadrático medio es:

$$\hat{\sigma}^2 = \frac{W}{n-q} = \frac{1}{n-q}\sum_{i}^{q}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_i})^2$$

Bajo la hipótesis nula $H_o : \mu_i = \mu\ (\forall i)$ o equivalentemente $H_o : \alpha_i = 0\ (\forall i)$, $\hat{\mu}_i = \overline{y}$ y el error cuadrático medio es igual a:

$$\hat{\sigma}_o^2 = \frac{T}{n-1} = \frac{1}{n-1}\sum_{i=1}^{q}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2.$$

La diferencia entre los mínimos cuadrados bajo H_0 y bajo H_1 es igual a:

$$SC(A) = T - W = B = \sum_{i=1}^{q} n_i(\overline{y_i} - \overline{y})^2 = \sum_{i=1}^{q} n_i(\hat{\mu}_i - \overline{y})^2 = \sum_{i=1}^{q} n_i \hat{\alpha}_i^2.$$

Bajo la hipótesis nula $SC(A) \sim \sigma^2 \chi^2_{q-1}$ y $SC(E) \sim \sigma^2 \chi^2_{n-q}$. Bajo la hipótesis alternativa $SC(A)$ es decentrado: $SC(A) \sim \sigma^2 \chi^2_{q-1}\left(\frac{\sum_i n_i \alpha_i^2}{\sigma^2}\right)$.

El test de Fisher para H_o esta dado por:

$$\frac{B/(q-1)}{W/(n-q)} \sim F_{q-1,n-q}$$

Fuente de varianza	Suma de cuadrados (SC)	Grados de libertad ($g.l.$)	Cuadrados medios ($CM = SC/g.l.$)	F
Factor A	B	$q-1$	$CM(A) = B/(q-1)$	$CM(A)/CM(E)$
Errores	W	$n-q$	$CM(E) = W/(n-q)$	
Total	T	$n-1$		

Se pueden resumir los estadísticos en una tabla de análisis de la varianza (ANOVA):

Calculemos las esperanzas de los cuadrados medios: $E(CM(A))$ y $E(CM(E))$.

$$E(SC(A)) = \sum_i n_i E((\bar{y_i} - \bar{y})^2) = \sum_i n_i E(\bar{y_i^2}) - nE(\bar{y^2})$$

Como $E(\bar{y^2}) = \dfrac{\sigma^2}{n} + \mu^2$ y $E(\bar{y_i^2}) = \dfrac{\sigma^2}{n_i} + \mu_i^2 \Rightarrow E(CM(A)) = \sigma^2 + \dfrac{1}{q-1}\sum_i n_i \alpha_i^2$ y $E(CM(E)) = \sigma^2$.

Ejemplo 4.1 Tomando los datos del ejemplo 4.3 del capítulo 4, consideramos como variable respuesta Y el total de las tres repeticiones y los grupos como factor A con cuatro niveles. El valor del estadístico del test F es igual a 70,93 (tabla ANOVA 4.1), lo que da un p-valor nulo. Se concluye que existe una diferencia entre los 4 grupos. Se confirma las diferencias con el gráfico 4.1, que permite explicarlas; se diferencias los grupos salvo entre los grupos 2 y 3.

Fuente	SC	g.l.	CM	F
Factor A	2232	3	744	70,93
Errores	178,3	17	10,49	
Total	2410	20	743,9	

Cuadro 4.1: ANOVA

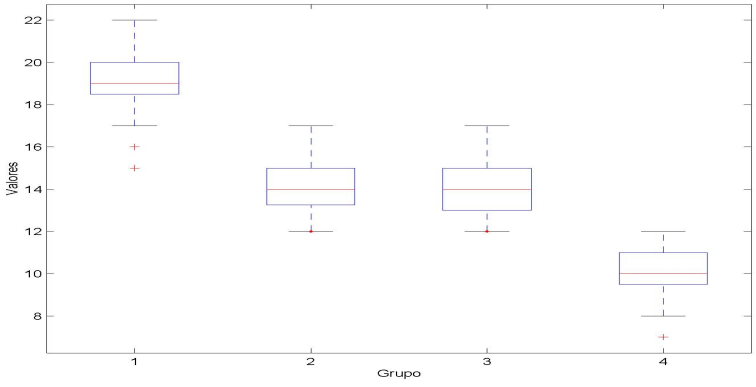

Figura 4.1: Valores por grupo

4.3.2. Modelo a efectos aleatorios

Ahora el modelo se escribe

$$y_{ij} = \mu + a_i + \varepsilon_{ij} \qquad (i = 1,2,\ldots q; \quad j = 1,2,\ldots,n_i)$$

en donde los efectos a_i y los errores ε_{ij} son aleatorios con las distribuciones:

- $a_i \sim N(0,\sigma_a^2)$ e independientes entre si.
- $\varepsilon_{ij} \sim N(0,\sigma_e^2)$ e independientes entre si.
- los efectos a_i son independientes de los errores ε_{ij}.

En el modelo con efectos fijos se busca estimar los efectos y compararlos entre si. Aquí en el modelo de efectos aleatorios, lo fundamental es estimar las varianzas σ_a^2 y σ_e^2 y compararlas. Lo anterior se llama **estimación de las componentes de la varianza**. Observamos que en el modelo con efectos aleatorios las observaciones y_{ij} no son todas independientes entre si y $Var(y_{ij}) = \sigma_a^2 + \sigma_e^2$, $Var(\bar{y_i}) = \sigma_a^2 + \dfrac{\sigma_e^2}{n_i}$. Más precisamente:

$$y_{ij} = \mu + a_i + \varepsilon_{ij} \sim N(\mu, \sigma_a^2 + \sigma_e^2) \Rightarrow \bar{y_i} = \mu + a_i + \frac{1}{n_i}\sum_{j=1}^{n_i} \varepsilon_{ij} \sim N\left(\mu, \sigma_a^2 + \frac{\sigma_e^2}{n_i}\right)$$

$$\Rightarrow \bar{y} = \mu + \sum_{i=1}^{q} \frac{n_i}{n} a_i + \frac{1}{n}\sum_{i=1}^{q}\sum_{j=1}^{n_j} \varepsilon_{ij} \sim N\left(\mu, \sum_{i=1}^{q} \frac{n_i^2}{n^2}\sigma_a^2 + \frac{\sigma_e^2}{n}\right)$$

Para estimar las componentes de la varianza, usaremos el método de los momentos a partir de las esperanzas de los cuadrados medios

$$SC(A) = \sum_{i=1}^{q} n_i(\bar{y_i} - \bar{y})^2 \qquad \text{y} \qquad SC(E) = \sum_{i=1}^{q}\sum_{j=1}^{n_i} (y_{ij} - \bar{y_i})^2$$

Tenemos $E(SC(A)) = \sum_i n_i E((\bar{y_i} - \bar{y})^2) = \sum_i n_i E(\bar{y_i^2}) - nE(\bar{y^2})$, o sea

$$E(SC(A)) = \sum_{i=1}^{q} n_i\left(\sigma_a^2 + \frac{\sigma_e^2}{n_i}\right) + \sum_{i=1}^{q} n_i\mu^2 - \sum_{i=1}^{q}\frac{n_i}{n}\sigma_a^2 - \sigma_e^2 - n\mu^2 = (q-1)\sigma_e^2 + \frac{n^2 - \sum_i n_i^2}{n}\sigma_a^2$$

Luego:

$$E(CM(A)) = \sigma_e^2 + \frac{n^2 - \sum_i n_i^2}{n(q-1)}\sigma_a^2$$

$$E(SC(E)) = \sum_{i=1}^{q}\sum_{j=1}^{n_i} E((y_{ij}-\bar{y})^2)) - E(SC(A))$$

Como $\sum_{i=1}^{q}\sum_{j=1}^{n_i} E((y_{ij}-\bar{y})^2) = \sum_{i=1}^{q}\sum_{j=1}^{n_i} E((y_{ij})^2) - nE(\bar{y^2}) = (n-1)\sigma_e^2 + \dfrac{n^2-\sum_i n_i^2}{n}\sigma_a^2$

$$E(SC(E)) = (n-q)\sigma_e^2 \Rightarrow E(CM(E)) = \sigma_e^2$$

Se estima entonces σ_a^2 y σ_e^2 por:

$$\hat{\sigma}_e^2 = CM(E) = \frac{1}{n-q}\sum_i\sum_j (y_{ij}-\bar{y_i})^2$$

$$\hat{\sigma}_a^2 = \frac{n(q-1)(CM(A)-CM(E))}{n^2-\sum_i n_i^2}$$

Notas:

- Si el modelo es balanceado ($\forall i : n_i = m$), entonces $\hat{\sigma}_a^2 = \frac{CM(A)-CM(E)}{n/q}$
- Puede ocurrir que $\hat{\sigma}_a^2$ sea negativo; en este caso se anula.
- Se encuentran otros estimadores como el de máxima verosimilitud o Hodges-Lehmann.

El test interesante aquí es entre la hipótesis nula $H_0 : \sigma_a^2 = 0$ y la hipótesis alternativa $H_1 : \sigma_a^2 > 0$. Al rechazar H_0, podemos rechazar que todos los efectos son parecidos. Trataremos solamente del caso de modelos balanceados ($\forall i : n_i = m$)[1]. En este caso,

$$SC(A) \sim \left(\sigma_e^2 + \frac{n}{q}\sigma_a^2\right)\chi_{q-1}^2 \qquad \text{y} \qquad SC(E) \sim \sigma_e^2\chi_{n-q}^2$$

Bajo la hipótesis nula $\dfrac{CM(A)}{CM(E)} \sim F_{q-1,n-q}$, pero, bajo la hipótesis alternativa

$$\frac{CM(A)}{CM(E)} \sim \left(1 + \frac{n\sigma_a^2}{q\sigma_e^2}\right)F_{q-1,n-q}$$

Ejemplo 4.2 Consideremos los datos de la tabla 4.2, en donde se tiene un factor a tres modalidades y 10 observaciones por modalidad. Se presenta en la tabla 4.3 el análisis de la varianza de los datos y el gráfico asociado (Figura 4.2).

[1] Para el caso no balanceado consultar Searle, capítulos 10 y 11

Grupo	1	2	3	4	5	6	7	8	9	10
1	299	269	290	289	286	284	299	269	299	305
2	326	326	318	340	352	336	341	337	345	302
3	323	314	314	282	328	321	281	280	291	301

Cuadro 4.2: Datos

Fuente	SC	g.l.	CM	F
Factor A	9754	2	4877	20,25
Errores	6503	27	240,9	
Total	$1,626\Delta 10^4$	29		

Cuadro 4.3: Tabla ANOVA

El p-valor del test F es nulo y las estimaciones de las componentes de la varianza son:

$$\hat{\sigma}_a^2 = 463,61 \qquad \text{y} \qquad \hat{\sigma}_e^2 = 240,87$$

Rechazamos la hipótesis nula y concluimos que los efectos de los grupos son diferentes sobre la variable.

Ejemplo 4.3 Los 6 datos de la tabla 4.4 están repartidos en dos grupos de tres. Si bien las medias son parecidas y el p-valor igual a 0,75 asociado a F (tabla 4.5) mo permite rechazar la igualdad de los efectos, las distribuciones son bien diferentes como lo muestra un valor negativo (4.3):

$$\hat{\sigma}_a^2 = -15,33 \qquad \text{y} \qquad \hat{\sigma}_e^2 = 52.$$

Grupo	1	2	3
1	19	17	15
2	25	5	15

Cuadro 4.4: Dos grupos y 3 repeticiones

4.4. Clasificación a dos factores

4.4.1. Modelos a efectos fijos

Sean A y B los dos factores que toman q y r categorías respectivamente $(A_1, A_2, \ldots, A_q)$ para el factor A y $B_1, B_2, \ldots, B_r$ para el factor B) e y_{ijk} $(i = 1,2,\ldots,q;\ j = 1,2,\ldots,r;\ k = 1,2,\ldots,n_{ij})$ los valores obtenidos sobre la variable respuesta Y relativos a observaciones provenientes de diferentes categorías de los factores.

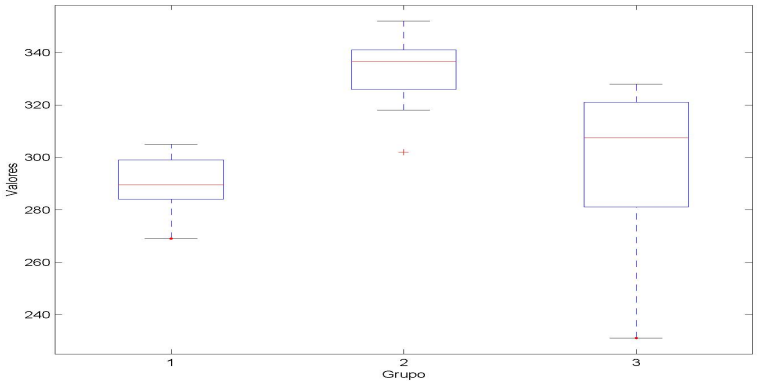

Figura 4.2: Valores por grupo

Fuente	SC	gl	CM	F
Factor A	6	1	6	0,1154
Errores	208	4	52	
Total	214	5		

Cuadro 4.5: ANOVA

Para estudiar si hay homogeneidad dentro de las celdas comparando las diferencias de un factor a otro o de una categoría a otra dentro de un mismo factor, el modelo más simple es un modelo lineal:

$$y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ijk} \qquad (i = 1,2,\ldots,q;\ j = 1,2,\ldots,r;\ k = 1,2,\ldots,n_{ij})$$

en donde μ es el modelo promedio, los τ_i los efectos del factor A, los β_j son los efectos de factor B y los ε_{ijk} son los errores asociados al modelo.

Un modelo más general

$$y_{ijk} = \mu + \tau_i + \beta_j + \delta_{ij} + \varepsilon_{ijk} \qquad (i = 1,2,\ldots,q;\ j = 1,2,\ldots,r;\ k = 1,2,\ldots,n_{ij})$$

permite efectos δ_{ij} que miden las interacciones entre los dos factores A y B.

Este modelo es de rango incompleto; el rango en igual a qr. Pero se ibservará que la función $\mu_{ij} = \mu + \tau_i + \beta_j + \delta_{ij}$ sea estimable. Introduciremos $q + r + 1$ restricciones para estimar a los $q + r + qr + 1$ parámetros. Sean las notaciones:

$$n_{i\bullet} = \sum_{j=1}^{r} n_{ij}, \qquad n_{\bullet,j} = \sum_{i=1}^{q} n_{ij}, \qquad \bar{y}_{ij} = \frac{1}{n} \sum_{k=1}^{n_{ij}} y_{ijk}$$

$$\bar{y}_{i\bullet} = \frac{1}{n_{i\bullet}} \sum_{j=1}^{r} n_{ij} \bar{y}_{ij}, \qquad \bar{y}_{\bullet j} = \frac{1}{n_{\bullet j}} \sum_{i=1}^{q} n_{ij} \bar{y}_{ij}$$

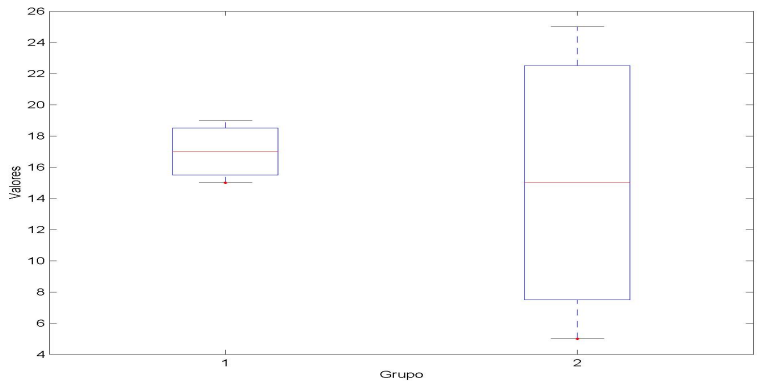

Figura 4.3: Valores por grupo

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{q}\sum_{j=1}^{r} n_{ij}\bar{y}_{ij} = \frac{1}{n}\sum_{i=1}^{q} n_{i\bullet}\bar{y}_{i\bullet} = \frac{1}{n}\sum_{j=1}^{r} n_{\bullet j}\bar{y}_{\bullet j}$$

Sean las restricciones:

$$\sum_{i=1}^{q} n_{i\bullet}\tau_i = 0, \qquad \sum_{j=1}^{r} n_{\bullet j}\beta_j = 0, \qquad \forall j : \sum_{i=1}^{q} n_{ij}\delta_{ij} = 0, \qquad \forall i : \sum_{j=1}^{r} n_{ij}\delta_{ij} = 0.$$

Veamos en primer lugar el caso del **modelo equilibrado**: $\forall(i,j) : n_{ij} = m$:

$$Q = \sum_{i,j,k}(y_{ijk} - \mu - \tau_i - \beta_j - \delta_{ij})^2 + 2\lambda_1\sum_i \tau_i + 2\lambda_2\sum_j \beta_j + 2\sum_i \gamma_i\left(\sum_j \delta_{ij}\right) + 2\sum_j \varphi_j\left(\sum_i \delta_{ij}\right)$$

$$\frac{\partial Q}{\partial \mu} = 0 \Rightarrow \hat{\mu} = \bar{y} \qquad \frac{\partial Q}{\partial \tau_i} = 0 \Rightarrow \hat{\tau}_i\bar{y}_{i\bullet} - \bar{y} \qquad \frac{\partial Q}{\partial \beta_j} = 0 \Rightarrow \hat{\beta}_j = \bar{y}_{\bullet j} - \bar{y}$$

$$\frac{\partial Q}{\partial \delta_{ij}} = 0 \Rightarrow \hat{\delta}_{ij} = \bar{y}_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y} \Rightarrow \hat{\mu}_{ij} = \hat{y}_{ij}$$

estimación que no depende de las restricciones utilizadas.

Sean las hipótesis nulas: $H_A : \forall i : \tau_i = 0$, $H_B : \forall j : \beta_j$ y $H_\delta : \forall(i,j) : \delta_{ij} = 0$ y la hipótesis alternativa H_1 : $E(y_{ijk}) = \mu + \tau_i + \beta_j + \delta_{ij}$. Los test F de Fisher son dados por:

$$F_A = \frac{\dfrac{mr}{q-1}\displaystyle\sum_{i=1}^{q}(\bar{y}_{i\bullet} - \bar{y})^2}{\dfrac{1}{n-qr}\displaystyle\sum_{i,j,k}(y_{ijk} - \bar{y}_{ij})^2}$$

para la hipótesis nula H_A contra la alternativa H_1.

$$F_B = \frac{\dfrac{mq}{r-1}\displaystyle\sum_{i=1}^{r}(\bar{y}_{i\bullet}-\bar{y})^2}{\dfrac{1}{n-qr}\displaystyle\sum_{i,j,k}(y_{ijk}-\bar{y}_{ij})^2},$$

para la hipótesis nula H_B contra la alternativa H_1.

$$F_\delta = \frac{\dfrac{m}{(q-1)(r-1)}\displaystyle\sum_{i=1}^{q}\sum_{j=1}^{r}(\bar{y}_{ij}-\bar{y}_{i\bullet}-\bar{y}_{\bullet j}+\bar{y})^2}{\dfrac{1}{n-qr}\displaystyle\sum_{i,j,k}(y_{ijk}-\bar{y}_{ij})^2},$$

para la hipótesis nula H_δ contra la alternativa H_1. Lo que se resuma en la tabla 4.6:

Fuente de validación	SC	g.l.	F
Factor A	$SC(A) = mr\displaystyle\sum_{i=1}^{q}(\bar{y}_{i\bullet}-\bar{y})^2$	$q-1$	$\dfrac{SC(A)/(q-1)}{SC(E)/(n-qr)}$
Factor B	$SC(B) = mq\displaystyle\sum_{j=1}^{r}(\bar{y}_{\bullet j}-\bar{y})^2$	$r-1$	$\dfrac{SC(B)/(r-1)}{SC(E)/(n-qr)}$
Interacciones	$SC(AB) = m\displaystyle\sum_{i=1}^{q}\sum_{j=1}^{r}(\bar{y}_{ij}-\bar{y}_{i\bullet}-\bar{y}_{\bullet j}+\bar{y})^2$	$(q-1)(r-1)$	$\dfrac{SC(AB)/((q-1)(r-1))}{SC(E)/(n-qr)}$
Errores	$SC(E) = \displaystyle\sum_{i,j,k}(y_{ijk}-\bar{y}_{ij})^2$	$n-qr$	
Total	$T = \displaystyle\sum_{i,j,k}(y_{ijk}-\bar{y})^2$	$n-1$	

Cuadro 4.6: Tabla ANOVA para un modelo equilibrado a dos factores con interacciones.