MA34B Sección 01 - Análisis de Varianza (ANOVA)

17 de junio 2004

Profesor cátedra: Rodrigo Abt B.
Auxiliar: Ismael Vergara

Supongamos que tenemos un experimento donde se registran observaciones de una variable Y en q categorías distintas, en que A_j representa una categoría, como por ejemplo, una marca de detergente. Así, A_1 podría ser *Omo*, A_2 sería *Drive*, etcétera. Las observaciones Y_{ij} representan un concepto de interés, por ejemplo *índice de blancura para ropa*. Luego Y_{ij} representaría el índice de blancura de la ropa i para el detergente j. Los resultados se podrían tabular como sigue:

A_1	A_2	A_3	$\ldots$	A_q
Y_{11}	Y_{12}	Y_{13}	$\ldots$	Y_{1q}
Y_{21}	Y_{22}	Y_{23}	$\ldots$	Y_{2q}
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
$Y_{n_1 1}$	$Y_{n_2 2}$	$Y_{n_3 3}$	$\ldots$	$Y_{n_q q}$

La idea es poder determinar si existen diferencias entre los grupos A_j para los valores observados Y. Podríamos modelar lo anterior a través un modelo lineal de la siguiente manera:

$$Y_{ij} = \mu_j + \varepsilon_{ij} \qquad j = 1, \ldots, q \qquad i = 1, \ldots, n_j$$

En que μ_j representa el efecto del grupo j, y ε_{ij} es un error aleatorio normal de media 0 y varianza σ^2. Es fácilmente verificable que el estimador de mínimos cuadrados para $\hat{\mu}_j$ es $\bar{Y}_j$, siendo esta última expresión el promedio de las observaciones Y en la categoría j, esto es:

$$\bar{Y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}$$

Sin embargo, si se desea utilizar un modelo con constante, se tiene:

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \qquad j = 1, \ldots, q \qquad i = 1, \ldots, n_j$$

Matricialmente:

$$Y = X\beta + \varepsilon$$

Donde:

$$\begin{pmatrix} Y_{11} \\ Y_{21} \\ \vdots \\ Y_{n_1 1} \\ Y_{12} \\ Y_{22} \\ \vdots \\ Y_{n_2 2} \\ \vdots \\ Y_{n_q q} \end{pmatrix} = \begin{pmatrix} \begin{matrix} 1 \\ 1 \\ \vdots \\ 1 \end{matrix} & n_1 \left\{ \begin{matrix} 1 \\ 1 \\ \vdots \\ 1 \end{matrix} \right. & & 0 & \\ \begin{matrix} 1 \\ 1 \\ \vdots \\ 1 \end{matrix} & 0 & n_2 \left\{ \begin{matrix} 1 \\ 1 \\ \vdots \\ 1 \end{matrix} \right. & 0 & \\ \vdots & & & & \\ \begin{matrix} 1 \\ 1 \\ \vdots \\ 1 \end{matrix} & & 0 & & n_q \left\{ \begin{matrix} 1 \\ 1 \\ \vdots \\ 1 \end{matrix} \right. \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_q \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{n_1 1} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{n_2 2} \\ \vdots \\ \varepsilon_{n_q q} \end{pmatrix}$$

Se puede observar que la primera columna de la matriz X es linealmente dependiente de las demás, es más, es igual a la suma del resto de las columnas. Esto significa que la matriz $X^t X$ NO es invertible.

Para remediar esta situación se procede a minimizar la suma de los cuadrados de los errores sujeto a una restricción adicional. Como μ representa el efecto de una media total, es natural pensar que la suma ponderada de los efectos marginales de cada grupo debería compensarse para no alterar la media total. Luego, se propone imponer la restricción:

$$\sum_{j=1}^{q} n_j \alpha_j = 0$$

Con esta restricción, se resuelve:

$$\min Q = (\sum_{j=1}^{q} \sum_{i=1}^{n_j} (Y_{ij} - \mu - \alpha_j)^2) - 2\lambda(\sum_{j=1}^{q} n_j \alpha_j)$$

Cuyas soluciones son: $\hat{\mu} = \bar{Y}, \hat{\alpha}_j = \bar{Y}_j - \bar{Y}$, lo cual responde bien a lo que dice la intuición. El estimador $\hat{\mu}$ corresponde entonces al efecto total de la media de las observaciones, mientras que los $\hat{\alpha}_j$ representan el efecto MARGINAL de cada grupo j. Debe notarse que la suma de ambos estimadores es igual $\bar{Y}_j$, que es lo mismo obtenido en el modelo sin constante.

Dado el modelo con la constante, es posible hacer el test de que todos los coeficientes α_j son cero, quedando solo μ, lo que se pueder llevar a cabo a través de un test F:

- La variación aportada por el modelo (o variación ENTRE grupos) se puede escribir como:

  $$B = SSE = \sum_{j=1}^{q} \sum_{i=1}^{n_j} (\hat{Y}_{ij} - \bar{Y})^2 = \sum_{j=1}^{q} \sum_{i=1}^{n_j} (\bar{Y}_j - \bar{Y})^2 = \sum_{j=1}^{q} n_j (\bar{Y}_j - \bar{Y})^2$$

  donde B/σ^2 sigue una distribución χ^2_{q-1}.

- A su vez, la variación aportada por los residuos (o variación INTRA grupos) se puede escribir como:

  $$W = SSR = \sum_{j=1}^{q} \sum_{i=1}^{n_j} \hat{\varepsilon}_{ij}^2 = \sum_{j=1}^{q} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2$$

  donde W/σ^2 sigue una distribución χ^2_{n-q}.

Luego, el estadístico F queda definido como:

$$F = \frac{\frac{B}{q-1}}{\frac{W}{n-q}} \sim F_{q-1,n-q}$$

Luego se rechaza $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_q = 0$ (no hay diferencia entre grupos) si $F > F_{q-1,n-q}(\delta)$ donde δ es el nivel de significación.

Ejemplo práctico:

Supongamos que queremos ver si existen diferencias de rendimiento (medido en kilómetros por litros) entre automóviles americanos, japoneses y europeos. Los mediciones se reportan en la siguiente tabla:

Americanos	Japonenes	Europeos
18.0	20.1	19.3
17.6	15.6	17.4
15.4	16.1	15.2
19.1	18.3	15.5
16.9	19.5	16.1

Sea A_1=Americanos,A_2=Japoneses y A_3=Europeos. Veamos las cantidades involucradas en el cálculo del estadístico F:

- Medias por grupos: $\frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}$.

$$\bar{Y}_1 = 17{,}4 \quad \bar{Y}_2 = 17{,}92 \quad \bar{Y}_3 = 16{,}7 \quad \bar{Y} = 17{,}34$$

- Variación ENTRE grupos (B): $\sum_{j=1}^{q} n_j(\bar{Y}_j - \bar{Y})^2$.

$$B = 5\cdot(17{,}4 - 17{,}34)^2 + 5\cdot(17{,}92 - 17{,}34)^2 + 5\cdot(16{,}7 - 17{,}34)^2 = 3{,}748$$

- Variación INTRA grupos (W): $\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2$.

  NOTA: Se puede usar esta fórmula, o el hecho de que cuando hay constante en el modelo: $SST = SSE + SSR$, donde SST es la varianza de Y multiplicada por el número de observaciones[1]:

[1] Usualmente en ANOVA se designa por la letra T

$$SST = T = \sum_{j=1}^{q}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y})^2 = nS_Y^2 = 38{,}676$$

De donde se puede despejar W, como $SST - B = 38{,}676 - 3{,}748 = 34{,}928$.

Normalmente los resultados obtenidos se resumen en la denominada ***Tabla ANOVA***:

Fuente	**gl**	**SSC**	**CM**	**F**	**P-Valor**
Entre grupos	$q-1$	B	$B/q-1$	$\frac{B/q-1}{W/n-q}$	-
Intra grupos	$n-q$	W	$W/n-q$		
Total	$n-1$	T			

Donde, **Fuente** es el origen de la variabilidad, **gl** son los grados de libertad asociados a dicha fuente de variabilidad, **SSC** es el valor de la fuente de variabilidad (B,W o T), **CM** es el cuadrado medio, es decir, corresponde a SSC/gl, **F** es el cuociente entre los cuadrados medios calculados, y **P-Valor** es el valor P asociado al estadístico **F** recién calculado.

En nuestro caso, se tiene:

Fuente	**gl**	**SSC**	**CM**	**F**	**P-Valor**
Entre grupos	$3-1$	3,748	$3{,}748/2 = 1{,}874$	$1{,}874/2{,}911 = 0{,}6438$	0,5425
Intra grupos	$15-3$	34,928	$34{,}928/12 = 2{,}911$		
Total	$15-1$	36,676			

NOTA: El valor P en general no se calcula directamente, y se deriva de los paquetes estadísticos, salvo que el valor del estadístico **B** sea muy grande comparado con el de las tablas, en cuyo caso se asume como 0.

Finalmente como $F = 0{,}6438 < F_{2,12}(5\,\%) = 3{,}885 \Rightarrow$ No se rechaza H_0, esto es, NO hay diferencias de rendimiento entre los distintos grupos de autos.

IMPORTANTE:

- Se cumple el balance de la suma de los grados de libertad: $gl(B) + ql(W) = gl(T)$
- Se cumple el balance de la suma de las fuentes de variabilidad: $B+W = T$
- Dadas las relaciones entre las distintas fuentes, es posible resolver la tabla ANOVA solo conociendo las medias de cada grupo($\bar{Y}_j$), desviaciones estándar de cada grupo ($\sqrt{S_j^2}$, donde S_j^2 es la varianza de la columna j), desviación estándar total (S_Y^2), y números de observaciones de cada categoría (n_j), ya que $W = n_1 S_1^2 + n_2 S_2^2 + \ldots + n_q S_q^2$. De donde, dado que $T = nS_Y^2$, se puede obtener $B = T - W$, y así completar la tabla correspondiente.