UNIVERSIDAD DE ANTIOQUIA
1803

Revista Colombiana de Ciencias Pecuarias

Tipos de suma de cuadrados en el análisis de la varianza

Types of sum of squares for analysis of variance

Luis F Restrepo B[1*], Estad, Esp estad bioma.

[1] Grupo de Investigación Grica, Facultad de Ciencias Agrarias, Universidad de Antioquia, AA 1226, Medellín, Colombia.
lusitano@agronica.udea.edu.co

(Recibido: 31 mayo, 2006; aceptado: 26 abril, 2007).

Resumen

La suma de cuadrados se emplea con el fin de efectuar una descomposición de la variabilidad total atribuible a la variable respuesta Y, en los diferentes componentes o factores controlados o manipulados por el investigador x, y la adición del error experimental, que constituye la fuente de variación que aglutina a todos los componentes no controlados dentro del modelo de clasificación experimental.

Palabras clave*: anova, estimación, hipótesis, sumas de cuadrados.*

Summary

The sum of squares is used in order to carry out a decomposition of the entire variability attributable to a response variable Y in the different components or controlled or manipulated factors by the an investigator X, and the addition of the experimental error that constitutes the source of variation that agglutinates all the not controlled components inside the model of experimental classification.

Key words: *anava, estimation, hypothesis, sum of squared.*

Introducción

El objetivo en todo diseño experimental es minimizar la suma de cuadrados del error, con el fin de poder maximizar el rechazo de la hipótesis nula y así establecer divergencia en el efecto de los tratamientos (véase Figura 1).

* Autor para el envío de la correspondencia y la solicitud de separatas. Facultad de Ciencias Agrarias, Universidad de Antioquia, AA 1226, Medellín, Colombia. E-mail: lusitano@agronica.udea.edu.co

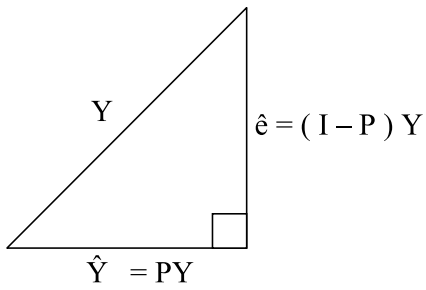


Figura 1. Distribución angular.

$$\hat{Y} = X(X'X)^{-1} \; X'Y = PY$$

$$P = X\,(X'X)^{-1} \; X'$$

El cual constituye el proyector ortogonal de y sobre el espacio generado por las columnas de X, C (x)

$$Y = \hat{Y} + \hat{e} \quad \text{donde,} \hat{Y} \; \varepsilon \; C\,(x),\; y \; \varepsilon \; R^n,\; \hat{e} \; \varepsilon \; C^I(x)$$

Y: variable dependiente, ê: error experimental, Ŷ: matriz diseño**.**

$I - P$ **,** es un proyector ortogonal de y sobre el complemento ortogonal del espacio columna de **X,** C (x)

Dado el concepto de ortogonalidad y aplicando el Teorema de Pitágoras, se genera la descomposición ortogonal clásica del análisis de la varianza (8):

$$\|Y\|^2 = \|\hat{Y}\|^2 + \|\hat{e}\|^2$$

$$\|Y\|^2 = Y'Y = \sum_{i=1}^{n} Y_i^2 = \text{Suma de cuadrados total}$$

$$\|\hat{Y}\|^2 = \hat{Y}\,'\,\hat{Y} = Y'PY = \text{Suma de cuadrados de parámetros}$$

$$\|\hat{e}\|^2 = \hat{e}\,'\,\hat{e} = Y'(I-P)Y = \text{Suma de cuadrados residual}$$

Tipos de sumas de cuadrados

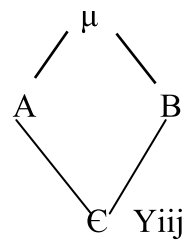
Suma de cuadrados tipo I

Se genera por medio del método de la ordenación a priori, en el cual se efectúan ordenaciones paramétricas de interés en forma a priori, obteniéndose un cuadrado único y un análisis que contenga todas las ordenaciones generadas por el modelo de clasificación experimental (10).

La suma de cuadrados tipo I se usa para probar hipótesis sobre medias ponderadas, ajustadas o no (1, 19).

La suma de cuadrados tipo I es igual a la suma de cuadrados tipo II, III y IV para diseños balanceados; es decir aquellos diseños donde cada tratamiento tiene igual número de replicaciones (2, 3).

Si se tiene un diseño de estructura:



$$Yiij = \mu + A_i + B_j + AB_{ij} + Є \; scij$$

μ: efecto promedio.
A y B factores de interés.
Є: error experimental.

Se desprende:

A	R(α/μ)
B	R(β/ μ, α)
AB	R(α β/ μ, α, β)

La suma de cuadrados tipo I se emplea en diseños donde hay interacción o no de factores o donde existen factores anidados (18).

Suma de cuadrados tipo ll

Esta se puede generar a partir de la suma de cuadrados tipo I, en la cual se escogen únicamente las hipótesis asociadas con medias ponderadas ajustadas. La suma tipo II es de la forma: R(α / μ, β) y R(β/ μ, α)

En la suma de cuadrados tipo I se tiene:

$R(\mu) + R(\alpha/ \mu) + R(\beta/ \mu, \alpha) + R(\gamma/ \mu, \alpha, \beta) = R(\mu) + R(\beta/ \mu) + R(\alpha/ \mu, \beta) + R(\gamma / \mu, \alpha, \beta) = R(\mu, \alpha, \beta, \gamma)$

En la suma de cuadrados tipo II en general no siempre ocurre lo anterior. Las sumas de cuadrados tipo II son provenientes de una partición ortogonal de la suma de cuadrados de los parámetros, así (10).

$$R(\mu) + R(\alpha/ \mu, \beta) + R(\beta/ \mu, \alpha) + R(\gamma/ \mu, \alpha, \beta) \neq R(\mu, \alpha, \beta, \gamma)$$

	$R(\ )$	
A	$R(\alpha /\mu, \beta)$	suma de cuadrados de A.
B	$R(\beta /\mu, \alpha)$	suma de cuadrados de B.
AB	$R(\gamma /\mu, \alpha, \beta)$	suma de cuadrados de la interacción AB.

La SC Tipo II =SC Tipo III = SC Tipo IV Si no existe interacción en el modelo.

La suma de cuadrados Tipo II puede ser descrita en general

R (@ factor / con todos los otros factores apropiados)

Por ejemplo, la suma de cuadrados para el factor **A**, es la suma de cuadrados ajustada por todos los otros factores y las interacciones, excepto las interacciones donde está el factor **A** y los factores anidados dentro de **A**.

La suma de cuadrados de A es: $R(\alpha /\mu, \beta)$ y no $R(\alpha / \mu, \beta, \alpha_x \beta)$

Para un modelo con dos factores cruzados **A**, **B**; y un factor **C** anidado dentro de A$\longrightarrow$ C: A

La suma de cuadrados Tipo II para A**,** es:

$R(\alpha/ \mu, \beta)$ y no $R(\alpha/ \mu, \beta, \gamma: \alpha)$

Suponga que hay tres factores A, B y C donde sólo hay interacción entre los factores A y B.

La suma de cuadrados de A es: $R(\alpha/ \mu, \beta, \gamma)$

La de B es: $R(\beta / \mu, \alpha, \gamma)$

Y la de C es: $R(\gamma / \mu, \alpha, \beta, \alpha_x \beta)$

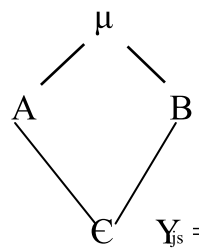
Finalmente, la de la interacción de A * B es:

$$R(\alpha_x \beta / \mu, \alpha, \beta, \gamma)$$

Suma de cuadrados tipo lll

Las sumas de cuadrados tipo III se pueden obtener entre otros, a través de los métodos de cuadrados de medias ponderadas (20), o por el método de los mínimos cuadrados completos (15), o por medio de la inversa de una fracción, de la inversa de Searle (16) Las hipótesis son sobre medias no ponderadas.

Con base en la estructura:



$Y_{js} = \mu + A_i + B_j + AB_{ij} + Є_{s(ij)}$

Se tiene por el método de Yates.

A	$R(\hat{\alpha} / \hat{\mu}, \hat{\beta}, \hat{\gamma})$
B	$R(\hat{\beta} / \hat{\mu}, \hat{\alpha}, \hat{\gamma})$
AB	$R(\hat{\gamma} / \hat{\mu}, \hat{\alpha}, \hat{\beta}$

La suma de cuadrados tipo III sirve para modelos restringidos. Por ejemplo, de la forma: $R(\hat{\alpha} / \hat{\mu}, \hat{\beta}, \hat{\phi})$

Donde: $R(\hat{\mu}, \hat{\alpha}, \hat{\beta}, \hat{\phi}) \neq R(\alpha/\mu\ \beta, \gamma) = 0$

Cabe anotar que la suma de cuadrados de este tipo también se puede emplear ante la presencia de datos completos en todas las celdas (5). Siendo idéntica a la suma de cuadrados de promedios ponderados.

Suma de cuadrados tipo lV

Esta es similar a la suma de cuadrados tipo III, si no existen celdas vacías. Si al menos una celda está vacía, la SC tipo III ≠ SC tipo IV y se asocian a diferentes hipótesis.

Ante la presencia de celdas vacías la suma de cuadrados tipo IV puede no ser única ya que depende de la posición y del número de celdas vacías.

La suma de cuadrados tipo IV está referida a hipótesis sobre contrastes entre medias poblacionales de celdas ubicadas en la misma columna o fila. Siempre se inician las comparaciones por la última fila o columna.

Las sumas de cuadrados tipo I, II, III son empleadas para estimar un modelo y diferentes subconjuntos teóricos derivados del modelo (12). Mientras la suma de cuadrados tipo IV se utiliza para probar hipótesis que son determinadas por el procedimiento GLM (modelo lineal general) del paquete estadístico SAS, donde las hipótesis seleccionadas dependen del patrón de las celdas, y del estadístico F acorde a la prueba estadística del conjunto de hipótesis existentes.

La suma de cuadrados tipo IV puede diferenciar una secuencia de filas de otras (13), para el mismo conjunto de datos, donde se recuerda que no es única dicha suma de cuadrados para las filas o columnas existentes.

Suponga que se tiene:

FILA1	*M11*	*M13*	*M14*
FILA2	*M21*	*M22*	
FILA3	*M32*	*M33*	*M34*

Las hipótesis para el conjunto de filas en el proceso GLM del SAS pueden ser:

$$H1: \begin{cases} M33 + M34 - (M13 + M14) = \mathbf{0} \\ M32 - M22 = 0 \end{cases}$$

H1: hipótesis de contraste., y se relacionan con la suma tipo IV, la cual se puede calcular a partir de **Q**

$Q = \beta^{0'} K (K'GK)^{-1}K' \beta^0 = Y'XG'K(K'GK)^{-1}K'GX'Y$

Asociada en general a la hipótesis:

$H : K' \beta = 0$

Cuando

$\gamma_k = \gamma_x - 1$, $\mu = 0$ y $K' 1 = 0$ $Q= R(\beta)-R(\mu)$

Cuando $\gamma_k = \gamma_x$, M =0 correspondiente a la hipótesis

$H: X \beta= 0$ $Q= R(\beta)$

$Q=SSR$

Si $\gamma_x = \gamma_k$ $M = 0$

La simetría de X'X implica que la matriz R es de rango completo γ_x tal que, X'X =RR' donde $(R'R)^{-1}$ existe.

$Q_0 = Y' X R (R' R)^{-2} R' K \{ K' R (R' R)^{-2} R' K \}^{-1}$ $K'R(R'R)^{-2}R'X'Y = Y'XR(R'R)^{-1}L' (LL')^{-1}L (R'R)^{-1}$ $RX'Y$ donde $L = K'R(R'R)^{-1}$

K' tiene rango completo γ_x fila y R tiene rango completo columna γ_x ademas $(L'L)^{-1}$ existe $ILI \neq 0$ tambien L^{-1} existe.

$Q_0 = Y'XR(R'R)^{-2}RX'Y = Y' XGX'Y = SSR$

La suma de cuadrados tipo IV no necesariamente tiene en cuenta todos los datos. Las secuencias de filas pueden ser leídas en forma diferente conduciendo a diferentes sumas de cuadrado tipo IV (8, 10). La suma de cuadrados tipo IV depende de la secuencia en que esté la fila con los efectos promedios (10).

Conclusiones

En general se pueden concluir las siguientes relaciones entre las sumas de cuadrados por el procedimiento GLM del SAS:

1. Si las muestras son balanceadas.
   SCI = SCII = SCIII = SCIV (8)

2. Si todas las celdas están ocupadas (14).
   SCIII = SCIV

3. Si el modelo no contiene interacción.
   SCII = SCIII = SCIV

4. Las sumas de cuadrados pueden estar asociadas a diferentes hipótesis.

5. El investigador deberá tener un conocimiento amplio de estadística, a fin de poder distinguir bien las hipótesis; o estar muy bien asesorado por un profesional con amplio conocimiento en el tema estadístico experimental. Debe existir una perfecta interacción entre el investigador y el estadístico en todas las fases del proceso experimental, y sobre todo al momento de plantear las hipótesis de interés práctico y las estrategias que serán adoptadas para probarlas (10). Los métodos estadísticos pueden servir para simplificar la elección, entre otras, de los tipos de suma de cuadrados más indicados para probar las hipótesis del verdadero interés para el investigador.

La suma de cuadrados tipo I es apropiada para diseños balanceados ortogonales (4, 11). También es empleada en diseños no ortogonales (6), tal como el citado por tal proceso se obtienen particulares anidamientos ajustados para algunos efectos, pero no para otros. Se pueden asignar algunos efectos *a priori* y ubicarlos en el modelo.

La suma de cuadrados tipo III es altamente recomendada ante la presencia de no ortogonalidad. En el anova se usa promedio muestral armónico para ajustar el total de la celda.

La suma de cuadrados tipo I depende de las hipótesis del orden en que el efecto esta especificado (7, 17).

La suma de cuadrados tipo II es apropiada para modelos construidos, y es naturalmente seleccionada en modelos de regresión (8, 9).

Las sumas tipo III y IV tienen las mismas hipótesis para datos balanceados o no, y trabajan con promedios marginales, donde algunos promedios marginales no son definidos. Es generalmente obvio cuando se comparan muchos efectos.

La suma tipo III no depende del orden del efecto, o de niveles. Sin embargo los contrastes ortogonales empleados son complejos para interpretar. Algunas interacciones son cero.

La sumas tipo I y II dependen del conteo en la celda.

La suma tipo IV se emplea para analizar subconjuntos de niveles de factores elegidos automáticamente (8).

La suma tipo III no depende del orden de los efectos. Los contrastes son complejos y se asumen algunas interacciones como cero.

Ejemplo

Se efectuó un experimento con tilapía donde se evaluó la ganancia de peso expresada en gramos. Se tenían dos factores alimenticios A y B cada uno con dos dosificaciones. El interés es obtener las sumas de cuadrados bajo dos situaciones (diseño balanceado y no balanceado). Cabe anotar que el diseño experimental empleado fue completamente aleatorizado en arreglo factorial 2*2 efecto fijo.

Se puede apreciar que todas las sumas de cuadrados coinciden.

Diseño balanceado

Obs	A	B	Ganancia de peso
1	0	1	300
2	0	1	280
3	0	1	275
4	0	0	255
5	0	0	267
6	0	0	245
7	1	0	233
8	1	0	247
9	1	0	250
10	1	1	320
11	1	1	315
12	1	1	245
13	0	1	287
14	0	0	240
15	1	0	276
16	1	1	243

Dependent Variable: GP

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	3	4014.25000	1338.08333	2.24	0.135
Error	12	7161.50000	596.79167		
Corrected Total	15	11175.75000			

Source	DF	Type I SS	Mean Square	F Value	Pr> F
A	1	25.000000	25.000000	0.04	0.8410
B	1	3969.000000	3969.000000	6.65	0.0241
A*B	1	20.250000	20.250000	0.03	0.8569

Source	DF	Type II SS	Mean Square	F Value	Pr > F
A	1	25.000000	25.000000	0.04	0.8413
B	1	3969.000000	3969.000000	6.65	0.0241
A*B	1	20.250000	20.250000	0.03	0.8569

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	1	25.000000	25.000000	0.04	0.8413
B	1	3969.000000	3969.000000	6.65	0.0241
A*B	1	20.250000	20.250000	0.03	0.8569

Source	DF	Type IV SS	Mean Square	F Value	Pr > F
A	1	25.000000	25.000000	0.04	0.8413
B	1	3969.000000	3969.000000	6.65	0.0241
A*B	1	20.250000	20.250000	0.03	0.8569

Diseño desbalanceado

Obs	A	B	GP
1	0	1	300
2	0	1	280
3	0	1	275
4	0	0	255
5	0	0	267
6	0	0	245
7	1	0	233
8	1	0	247
9	1	0	250
10	1	1	320
11	1	1	315
12	1	1	245
13	0	1	287
14	0	0	240
15	1	0	276

Dependent Variable: GP

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5280.58333	1760.19444	3.68	0.0469
Error	11	5261.41667	478.31061		
Corrected Total	14	10542.00000			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
A	1	2.410714	2.410714	0.01	0.9447
B	1	5217.858516	5217.858516	10.91	0.0070
A*B	1	60.314103	60.314103	0.13	0.7292

Source	DF	Type II SS	Mean Square	F Value	Pr > F
A	1	45.001374	45.001374	0.09	0.7648
B	1	5217.858516	5217.858516	10.91	0.0070
A*B	1	60.314103	60.314103	0.13	0.7292

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	1	53.083333	53.083333	0.11	0.7453
B	1	5273.391026	5273.391026	11.03	0.0068
A*B	1	60.314103	60.314103	0.13	0.7292

Source	DF	Type IV SS	Mean Square	F Value	Pr > F
A	1	53.083333	53.083333	0.11	0.7453
B	1	5273.391026	5273.391026	11.03	0.0068
A*B	1	60.314103	60.314103	0.13	0.7292

Sólo coinciden la suma de cuadrados tipolll y tipo lV.

Referencias

1. Calzada BJ. Métodos estadísticos para la investigación. Diversidad de la Molina. Lima. 1970. 640p.
2. Cochran WG, Cox GM. Experimental designs. 2. ed. New York, John Wiley, 1977. 611p.
3. Cochran WG, Cox GM. Diseños experimentales. Trillas, México D. F. 1981. 615p.
4. Cordeiro GM. Modelos lineares generalizados. Unicamp, Campinas. 1986. 286p.
5 Dobson. An introduction to linear models, Chadpman-Hall 2º edition. New York. 1991. 221p.
6. Federer WT. Experimental design, the MaCmillan Company, New York .1955. 554p.
7 Hinkelman K, Kempthorne O. Design and analysis of experiments. John wiley, New York.1994. 512p.
8. John JA, Draper NR. An alternative family of transformations. Appl Stat 1980; 2:190-197.
9. Kempthorne O, Folks L. Probability statistics and data analysis . Ames, Iowa. Iowa State University Press. 1971. 555p.
10. Lemma AF. Hipoteses estatisticas com amostras desequilibradas, Fac Sci Agrom Gembloux, Bélgica. 1991.
11. Little TM, Hills FJ. Métodos estadísticos aplicados en agricultura, trillas México D. F, 1976. 270p.
12. Martínez GA. Diseños experimentales, México; Colegio de Postgrado de Chapingo. 1983. 1058p.
13. Montgomery DC. Design and analysis of experiments, John wiley. New York 1991. 649p.
14. Ostle B. Estadística aplicada. Limusa-Wiley S.A. México D .F. 1973. 629p.
15. Overall JE, Spiegel DK. Concerning least squares analysis of experimental data. Psych Bull 1969; 72:311-322.
16. Searle SR. Linear models for unbalanced data. Wiley J Roy Statl Soc New York. 1987; 83:911-912.
17. Senedecor GW. Métodos estadísticos aplicados a la investigación agrícola y biológica. Continental. México D. F.1989. 503p.
18. Sokal R, Rohlf J. Introducción a la bioestadística: Reverte S.A. New Cork. 1980. 363p.
19. Torrie JH. Bioestadística. Principios y procedimientos. McGraw-Hill. México D.F. 1985. 666p.
20. Yates F The analysis of multiple classifications with unequal numbers in the different classes. J Am Stat 1934; 7:121-140.