Revista Colombiana de Estadística

Volumen 24 (2001) N^{o} 2, páginas 91 *a* 110

HIPÓTESIS Y SUMAS DE CUADRADOS TIPO III Y IV UN ENFOQUE A TRAVÉS DEL MODELO DE MEDIAS DE CELDA

Santana C., Juan Camilo[*]
López P., Luis Alberto[**]

Resumen

En este artículo se presenta una propuesta para obtener las funciones estimables y las sumas de cuadrados tipo III y IV determinadas por el procedimiento GLM del SAS en modelos de medias de celda conectados. En el caso de tener información con celdas vacías, se construyen los contrastes e hipótesis efectivas y se ilustran los resultados obtenidos con un modelo de clasificación a dos vías con interacción.
***Palabras claves**: Análisis de varianza, Celdas vacías, Datos desbalanceados, Funciones estimables, Modelos de medias de celda condicionados y no condicionados, Reparametrización.*

Abstract

This paper proposes a method to obtain the estimable functions and type III and IV sums of squares given by *SAS PROC GLM* in connected

[*]Estadístico Universidad Nacional de Colombia E-mail:juan_camilosantana@hotmail.com
[**]Profesor Asociado Departamento de Estadística. Universidad Nacional de Colombia E-mail:alopez@matematicas.unal.edu.co Departamento de Estadística

cell means models. In the case of empty cells, contrasts and hypotheses are constructed and the results are illustrated for a two way classification model with interaction.

***Key words**: Analysis of Variance, Empty Cells, Unbalanced data, Estimable Functions, Conditioned and Unconditioned Cell Means Models, Reparameterization.*

1. Introducción

Siempre que se trabajan diseños con o sin restricciones en la aleatorización, se desea información completa y balanceada; pero la realidad induce a circunstancias donde no se puede obtener toda la información deseada, ya sea por costos, por condiciones del material experimental o por cualquier otro motivo, con consecuencias como el desbalance en el experimento y la pérdida de la ortogonalidad del diseño. Los investigadores que manejan este tipo de información, se enfrentan a problemas no triviales, por el desconocimiento de la forma como se obtienen las diferentes sumas de cuadrados asociadas al análisis de varianza; esto conlleva a que se cometan errores en el planteamiento y la interpretación de sus hipótesis estadísticas asociadas a los parámetros de interés.

Las funciones estimables tipo I, II, III y IV dentro de la lógica de las salidas del SAS consiste en L-valores ligados cada uno con los parámetros del modelo cuando éste tiene una estructura superparametrizada. Una forma amplia de obtener estas funciones así como las correspondientes sumas de cuadrados puede estudiarse en Searle (1980).

Cuando se presentan estructuras de datos desbalanceados y hay presencia de celdas vacías hay dificultad en la construcción e interpretación de las hipótesis lineales en los modelos superparametrizados, sin embargo, cuando se trabaja con modelos de medias de celdas se tiene más claridad en la identificación e interpretación de estas hipótesis lineales.

La construcción de las funciones estimables tipo III y IV, y sus sumas de cuadrados en los modelos de medias de celda, no ha sido ampliamente difundida en la literatura estadística, y las referencias encontradas como Jennings and Ward (1982), Freund (1980), Hocking (1985, 1996) entre otros, no muestran con claridad la metodología de construcción de estas funciones estimables y sus respectivas sumas de cuadrados, dificultando con ello su difusión y aplicación entre los investigadores. Por ello en este artículo se busca mostrar en forma sencilla la construcción de dichas funciones en modelos de medias de celdas.

2. Marco teórico

El marco teórico de los modelos de medias de celdas fue propuesto por Speed et al. (1978), quienes presentan la siguiente definición:

Definición 1. Sea

$$y = W\mu + e \tag{1}$$

sujeto a

$$G\mu = g \tag{2}$$

donde y es un vector de N observaciones, μ un vector de p (p= $ran(W)$) medias de celda poblacionales, W una matriz de ceros y unos de orden $N \times p$, que vincula a las observaciones con su respectiva media poblacional y $e \sim N(0, \sigma^2 I)$, un vector de errores. La matriz G de orden $r \times p$, representa restricciones lineales conocidas respecto a las medias de celda, aunque no siempre se pueden imponer. Usualmente, estas restricciones van a reflejar los supuestos sobre las interacciones que se imponen en el modelo.

Dada la estructura de G, se satisface que el $ran(G) = r$. Es posible reordenar las columnas de esta matriz, lo cual conlleva a una partición de la forma $G = \left[G_1 \mid G_2\right]$, donde G_2 es de orden $r \times r$ con $ran(G_2)$= r y G_1 de orden $r \times x(p-r)$. Una forma sencilla de cómo debe hacerse la partición puede verse en López (1999) alterna a la propuesta de Murray and Smith (1985).

La partición de G, obliga a una partición tanto de μ como de W; es decir, $\mu^t = \left[\mu_1^t \mid \mu_2^t\right]$ y $W = [W_1 \mid W_2]$. Para este conjunto de matrices particionadas y teniendo en cuenta (2), se sigue que

$$G_1\mu_1 + G_2\mu_2 = g \tag{3}$$

luego

$$\mu_2 = G_2^{-1}(g - G_1\mu_1) \tag{4}$$

sustituyendo (4) en (1) se obtiene

$$y = W_1\mu_1 + W_2G_2^{-1}g - W_2G_2^{-1}G_1\mu_1$$

o de forma equivalente

$$y - W_2G_2^{-1}g = (W_1 - W_2G_2^{-1}G_1)\mu_1 \tag{5}$$

haciendo

$$y^* = y - W_2 G_2^{-1} g$$

y

$$V = (W_1 - W_2 G_2^{-1} G_1)$$

se llega al modelo

$$y^* = V\mu_1 + e \tag{6}$$

La solución por mínimos cuadrados, o por máxima verosimilitud, para μ_1 en (6) es igual a:

$$\hat{\mu}_1 = \left(V^t V\right)^- V^t y^* \tag{7}$$

Sustituyendo esta solución en (4), se tiene como estimación

$$\hat{\mu}_2 = G_2^{-1}(g - G_1\hat{\mu}_1) \tag{8}$$

de tal forma que la solución para μ en el modelo (1) es

$$\hat{\mu}^t = \left[\hat{\mu}_1^t | \hat{\mu}_2^t\right]$$

con

$$Var(\hat{\mu}_1) = \left(V^t V\right)^- \sigma^2 \tag{9}$$

$$Var(\hat{\mu}_2) = G_2^{-1} G_1 \left(V^t V\right)^- G_1^t \left(G_2^{-1}\right)^t \sigma^2 \tag{10}$$

y

$$Cov(\hat{\mu}_1, \hat{\mu}_2) = -G_2^{-1} G_1 \left(V^t V\right)^- \sigma^2 \tag{11}$$

con

$$\hat{\sigma}^2 = \frac{\left(y^* - V\hat{\mu}_1\right)^t \left(y^* - V\hat{\mu}_1\right)}{(N-p)}$$

En Murray & Smith (1985), se desarrolla un procedimiento general para la construcción de la matriz G dada en (2). Se supone que si se tienen k factores $(F_1, \cdots, F_k)$ bajo estudio

$$G = \Delta_{F_1} \otimes \Delta_{F_2} \otimes \cdots \otimes \Delta_{F_k}$$

donde $\Delta_{F_i}, i = 1, \cdots, k$ hace referencia a la presencia de un factor de interés y $\otimes$ al producto directo, y de acuerdo con Hocking (1996), estos Δ_{F_i} pueden ser obtenidos por la expresión:

$$\Delta_{F_i} = \left(I_{l_i-1} \mid -J_{l_i-1}\right) \quad i = 1, \cdots, k$$

donde I_{l_i-1} es una matriz identidad de orden igual al número de niveles del factor l_i, menos uno, mientras que J_{l_i-1} es un vector columna de unos de dimensión $l_i - 1$.

3. Modelo de medias de celda reparametrizado

El modelo de medias de celda reparametrizado propuesto por Bryce et al. (1980), constituye una herramienta relativamente simple para la estimación y prueba de hipótesis en modelos de efectos fijos y estructura desbalanceada de datos.

En esencia el modelo de medias de celda reparametrizado parte del modelo (1) no condicionado, de tal forma que si existe una matriz M no singular, entonces (1) puede escribirse como

$$y = WM^{-1}M\mu + e = Z\delta + e \tag{12}$$

En (12), la matriz $Z = WM^{-1}$ es de rango completo, lo cual facilita la construcción de hipótesis así como la estimación de parámetros, puesto que se satisface que cualquier elemento de $\delta = M\mu$ va a dar información de combinaciones lineales de los elementos de las medias de las celdas.

4. Estimabilidad y prueba de hipótesis en el modelo de medias de celda

En el modelo (1) se tiene como objetivo teórico estimar el conjunto de parámetros (μ) o alguna combinación lineal de ellos, $L\mu$, a partir de una combinación lineal de los componentes del vector y que tenga como valor esperado $L\mu$. Por lo tanto, $L\mu$ es estimable si y sólo si existe una combinación lineal de los componentes del vector y, cuyo valor esperado sea $L\mu$ (ver Rao, 1945).

Las ecuaciones normales para el modelo (1) a s vías de clasificación sin la restricción (2), son $\big(W^tW\big)\hat{\mu} = W^ty$, donde $W^tW = D\big\{\eta_{i,j\ldots s}\big\}$ es una matriz diagonal con elementos iguales al número de observaciones por celda $\eta_{i,j\ldots s}$. Las ecuaciones normales llevan a que el mejor estimador lineal insesgado (M.E.L.I) de $\mu_{i,j\ldots s}$ sea $\hat{\mu}_{i,j\ldots s} = \bar{y}_{i,j\ldots s}$, con varianza $Var(\hat{\mu}_{i,j\ldots s}) = \sigma^2/\eta_{i,j\ldots s}$, siendo $\hat{\sigma}^2$ el estimador de la varianza residual, la cual es obtenida por

$$\hat{\sigma}^2 = \frac{y^t\big(I - W\big(W^tW\big)^{-}W^t\big)y}{(N-p)}$$

Según Searle (1987), la hipótesis lineal general de combinaciones de medias de celdas se expresa como:

$$H : L^t\mu = g \tag{13}$$

y bajo la hipótesis nula, se satisface que el cociente:

$$F = \frac{SCQ}{\hat{\sigma}^2 r_L} = \frac{(L^t\hat{\mu} - g)^t\Big[L^t\big(W^tW\big)^{-}L\Big]^{-1}(L^t\hat{\mu} - g)}{\hat{\sigma}^2 r_L} \tag{14}$$

donde

$$SCQ = \big(L^t\hat{\mu} - g\big)^t\Big[L^t\big(W^tW\big)^{-}L\Big]^{-1}\big(L^t\hat{\mu} - g\big) \tag{15}$$

siendo r_L el rango de la matriz L, y además las matrices asociadas a las formas cuadráticas, SCQ y $\hat{\sigma}^2$, son independientes; este resultado implica que el cociente (14) se distribuya como una F con r_L y $(N-p)$ grados de libertad.

La SCQ en (15) puede obtenerse en forma más sencilla a partir de la definición del proyector ortogonal (veáse Iemma et al. (1999)), es decir

$$SCQ = y^tK(K^tK)^{-1}K^ty = y^tP_Hy \tag{16}$$

donde $P_H = K(K^tK)^{-1}K^t$, y K puede obtenerse a partir de la definición de estimabilidad propuesta por Rao (1945).

5. Contrastes efectivos

Hocking et al. (1980), proponen un método para determinar las hipótesis a ser examinadas cuando hay celdas vacías. El procedimiento está basado en la

premisa, que el investigador tiene en mente hipótesis las cuales son apropiadas si todas las celdas están llenas, pero no se pueden evaluar debido a la presencia de las celdas vacías. La idea a partir de esta situación es poder examinar hipótesis equivalentes a la hipótesis deseada (hipótesis efectiva) y obtener conclusiones acerca de la misma.

En Hocking (1996) se presenta un desarrollo teórico bastante claro acerca de los conceptos de contrastes e hipótesis efectivas; algunos de los resultados se resumen a continuación.

Definición 2. Los contrastes $G_{oo}\mu_o$ se dicen contrastes efectivos, si $G\mu = 0$ implica que $G_{oo}\mu_o = 0$, y G_{oo} es de rango máximo.

A partir del modelo (1) sujeto a (2) y en presencia de m celdas vacías, se define el vector de medias de celda poblacional como

$$\mu = \begin{bmatrix} \mu_o \\ \mu_m \end{bmatrix}$$

donde μ_o y μ_m denotan el vector de medias de celda asociado con las celdas observadas y faltantes respectivamente. La matriz de frecuencias de celda es entonces escrita como $W = \begin{bmatrix} W_o | W_m \end{bmatrix}$, donde W_m es una matriz de ceros, así mismo, las restricciones impuestas al modelo son particionadas como

$$G = \begin{bmatrix} G_o |\ G_m \end{bmatrix} \tag{17}$$

de tal forma que la construcción de G_{oo} se obtiene haciendo operaciones entre filas en la ecuación (17), hasta obtener la siguiente expresión

$$G = \begin{bmatrix} G_{oo} & 0 \\ G_{mo} & G_{mm} \end{bmatrix} \text{ y } g = \begin{bmatrix} g_o \\ g_m \end{bmatrix} \tag{18}$$

donde G_{mm} es de orden $t \times m$ de rango t; con esta partición se construyen las restricciones e hipótesis efectivas, de gran interés en la obtención de las funciones estimables tipo III y IV que se desarrollan a continuación.

6. Funciones estimables y sumas de cuadrados tipo III y IV

6.1. Funciones estimables tipo III

Para la mayoría de diseños desbalanceados generalmente es posible examinar el mismo conjunto de hipótesis (funciones estimables) que se probarían en diseños balanceados. Para aquellos diseños los cuales no fueron inicialmente pensados como balanceados, y para los cuales hubo pérdida parcial de observaciones, generalmente no hay razón para alterar las hipótesis que se realizarían en diseños balanceados, es decir, en diseños con información perdida, las funciones estimables pueden parecerse a las empleadas en el caso balanceado.

Definición 3. Un conjunto de funciones estimables ($L^{'}s$), asociadas a cada uno de los factores del modelo, son funciones estimables tipo III si y sólo si cada L es una hipótesis de rango máximo ortogonal a todos los $L^{'}s$ de los factores que contienen al factor en cuestión.

Definición 4. Si $F1$ y $F2$ son dos factores cualesquiera, se dice que $F1$ está contenido en $F2$ si :

- Ambos factores involucran el mismo número de variables continuas y si el número es positivo entonces los nombres de las variables coinciden.
- Si $F2$ tiene más variables que $F1$, y si $F1$ tiene variables, entonces todas las variables de $F1$ están contenidas en $F2$.

Esta propiedad de contenencia es propia de las funciones estimables tipo II, III y IV.

Se puede por tanto obtener las funciones estimables tipo III a partir de las funciones tipo II, haciendo que cada L de orden inferior sea ortogonal a los L de todos los factores que contengan al factor de interés. Adicionalmente, si un factor no está contenido en otro factor, las funciones estimables tipo II y tipo III son iguales.

En Melo (2000) se obtienen las hipótesis tipo II a través del modelo de medias de celda modificado, descrito en las ecuaciones (3) a (11) adaptando un procedimiento propuesto por Goodnight (1978), el cual suministraba las mismas hipótesis cuando se trabaja con el modelo superparametrizado.

Para obtener las funciones estimables tipo III, se partió de la siguiente expresión desarrollada en Melo (2000):

$$H_1 = \left(V_2^t M V_2\right)^{-} \left(V_2^t M W\right) \tag{19}$$

donde

$$M = I - V_1 \left(V_1^t V_1\right)^{-} V_1^t$$

y V_j, $j = 1, 2$ corresponde a la matriz dada bajo el modelo de medias de celda modificado, con la restricción (2) apropiada.

Si se desea encontrar las hipótesis tipo III para un factor F_1, se define en primera instancia a V_1 como una matriz cuyos factores asociados no contienen a F_1 y a factores asociados a F_1, mientras que V_2 es una matriz que contiene al factor F_1.

Cada vector fila linealmente independiente obtenido a partir de (19), se busca que sea ortogonal al subespacio generado por los vectores asociados a los factores de orden superior que contienen a F_1, obteniendo de esta manera las hipótesis tipo III del factor en cuestión.

Los factores de orden superior pueden ser calculados por medio de contrastes efectivos en caso que se presenten celdas vacías. Si los factores de orden superior no existen bajo el modelo de medias de celda (por efecto de una alta dispersión de las celdas llenas), las hipótesis tipo III serán calculadas bajo una redefinición de las matrices V_1 y V_2 anteriormente descritas.

La matriz V_1 se construye con todos los factores, menos el de interés, y en la matriz V_2 se incluye al factor de interés y a todos los demás definidos en V_1. Esta forma de construir las matrices, preserva la filosofía del cálculo de las hipótesis tipo III, que busca reducir la suma de cuadrados cuando un factor es ajustado por todos los demás.

El procedimiento para obtener las hipótesis tipo III sin interacción, puede ser muy costoso en tiempo computacional, cuando en el modelo de medias de celdas modificado hay mas de dos factores de clasificación. Es más recomendable en ese caso construir las hipótesis tipo III con el modelo de medias reparametrizado.

En el modelo de medias de celda reparametrizado que se muestra en (12), la matriz Z se obtiene a partir de la matriz M^{-1} y W, pudiéndose construir

$$\mathbf{A} = \left[Z\left(Z^t Z\right)^{-} Z^t - Z_1 \left(Z_1^t Z_1\right)^{-} Z_1^t \right] \tag{20}$$

con la cual se obtienen las hipótesis tipo III para un factor de interés. En (20), la matriz Z contiene todos los factores que intervienen en el análisis, en tanto que la matriz Z_1 va a tener información de todos los factores menos el de interés. La matriz $\mathbf{A}$ es simétrica e idempotente, por lo tanto se puede expresar como $\mathbf{A} = \mathbf{R}^t\mathbf{R}$, y con esta descomposición, las hipótesis tipo III para un factor dado serán de la forma

$$H_2 : \mathbf{R}W\mu = 0 \tag{21}$$

6.2. Sumas de Cuadrados tipo III.

Para un modelo de clasificación cruzada a dos vías con A, B y AB, factores principales e interacción respectivamente, la suma de cuadrados tipo III asociada a un factor es calculada como una reducción en la suma de cuadrados, cuando el factor es ajustado por todos los demás, incluyendo las interacciones. En términos de la notación $R(.)$ propuesta en Searle (1987), se expresa como:

$$\begin{aligned} R(A \mid B, AB) &= R(A, B, AB) - R(B, AB) \\ R(B \mid A, AB) &= R(A, B, AB) - R(A, AB) \\ R(AB \mid A, B) &= R(A, B, AB) - R(A, B) \end{aligned}$$

Para modelos de clasificación con interacción, se calculan las sumas de cuadrados a partir de la expresión SCQ en (15). Para los modelos sin interacción, se propone calcular las sumas de cuadrados a partir de (22) empleando proyectores ortogonales como se describió en (16), de tal forma que

$$SCQ_1 = y^t \left[V_2 \big(V_2^t V_2\big)^{-} V_2^t - V_1 \big(V_1^t V_1\big)^{-} V_1^t \right] y \tag{22}$$

Cuando se trabaja con el modelo reparametrizado, la suma de cuadrados tipo III, para un factor específico, se define como:

$$SCQ_2 = y^t \mathbf{A} y \tag{23}$$

donde $\mathbf{A}$ está dada en (20). Las hipótesis y las sumas de cuadrados tipo III para los modelos sin interacción, bajo el modelo de medias de celda modificado y el reparametrizado son iguales.

6.3. Funciones estimables y sumas de cuadrados tipo IV

Las funciones estimables tipo IV no tienen el propósito de explicar alguna suma de cuadrados en función de un orden predeterminado, como sí lo son las tipo I, II, y III. Estas funciones se obtienen de subconjuntos no únicos de celdas llenas. Este hecho hace que algunos paquetes estadísticos arrojen en la salida de las sumas de cuadrados tipo IV, un comentario respecto a la no unicidad de estas sumas de cuadrados.

La no unicidad de estas hipótesis fue comentado inicialmente por Freund (1980), quien lo atribuyó a un reordenamiento de los datos, el cual consideraba, influía en las funciones estimables y las sumas de cuadrados. Jennings & Ward (1982) discuten los resultados de Freund y sugieren que la escogencia del subconjunto de datos que deberá usarse es arbitraria, pero limitada por el patrón de celdas llenas.

Definición 5. Para un factor F_1, las hipótesis tipo IV se calculan como contrastes simples de la diferencia entre medias de celda, respecto a la posición del factor en la tabla de contingencia :

1. Si el factor F_1 está ubicado en las filas, se calculan como las medias de celda que están en la misma columna, comenzando por la última fila.

2. Si el factor F_1 está ubicado en las columnas, se calculan como las medias de celda que están en la misma fila, comenzando por la última columna.

En la definición anterior el término última fila o columna, no debe entenderse literalmente puesto que para ciertas tablas de contingencia, la última fila (o columna) puede tener información faltante, de modo que para satisfacer la definición anterior, se debe tomar la fila (o columna) anterior. En presencia de celdas vacías se preserva la filosofía descrita en la definición 5, pero se debe tener cuidado en la forma como es aplicada.

En modelos a k vías de clasificación, se calcula las hipótesis tipo IV para un factor cualquiera, generando tablas de contingencia sobre las modalidades de los otros factores y realizando la lectura como en la definición 5. Por ejemplo, para un factor F_1, y otro seleccionado arbitrariamente, F_2, se generan tablas de contingencia sobre las modalidades de los demás factores $F_3, \ldots, F_k$, generándose

$$\Omega = \prod_{i=3}^{k} l_i \tag{24}$$

tablas de contingencia, donde $l_i = 3, \ldots, k$ corresponde al número de modalidades del i-ésimo factor. Independientemente del factor que acompañe a F_1 en la construcción de la tabla de contingencia, los contrastes obtenidos serán los mismos; además como las tablas generadas son disyuntas, entonces los contrastes generados son linealmente independientes. Construidas todas las tablas asociadas al factor, se reagrupan todos los contrastes asociados al factor en una matriz notada como H_3, y se crea una nueva matriz, C, que suma para cada fila los coeficientes de las medias de celda, con el fin de generar las modalidades asociadas al factor de interés, obteniendo la matriz C aumentada.

$$\Big[C|H_3\Big]$$

En esta matriz, si hay filas iguales, se mantiene solo una y, posteriormente con las filas resultantes, se hace una reducción entre filas de tal forma que C se transforme en $C^* = \Delta_{mod}$, donde mod corresponde al número de modalidades del factor en cuestión siendo, $\Delta_{mod} = \Big[I_{mod-1}| - J_{mod-1}\Big]$ y H_3 se transforma en H_3^*. En esta última matriz van a estar los coeficientes lineales asociados con el factor F_1.

Con este método, el número de contrastes linealmente independientes para el factor F_1 será $(l_1 - 1)$, desde que haya información en cada fila o columna que relacione todas las modalidades del factor F_1, entre sí; si hay filas o columnas sin datos, entonces el número de contrastes será menor que $(l_1 - 1)$ y en tal caso $C^* = \Delta_{mod}$ no puede ser obtenido completamente, debido a la alta dispersión de las celdas llenas en el modelo. Cuando esto sucede, debe tenerse cuidado con la aplicación de la definición 5, como se muestra en el trabajo de Hudson & Searle (1982). Para las interacciones ocurre algo semejante ya que en presencia de m celdas vacías el número de contrastes linealmente independientes se reduce.

Por ejemplo, los contrastes para la interacción entre los factores F_1 y F_2, pueden ser obtenidos con el procedimiento anterior, calculando sobre cada una de las Ω tablas de contingencia, contrastes efectivos y agrupándolos en forma semejante a la descrita para los factores principales. Para las interacciones de orden mayor que dos se debe emplear el método de contrastes efectivos discutido al inicio de esta sección.

Las hipótesis tipo IV construidas a partir del método anterior son aplicables en modelos con interacción, pero pueden ser empleadas igualmente en modelos sin interacción, ya que cualquier otro contraste tipo IV puede ser construido en función del arreglo factorial. Sin embargo, los obtenidos a través de la metodología propuesta son construidos de manera lógica y proveen una visión más clara de la naturaleza de las hipótesis que va a probar el experimentador, con relación a un determinado arreglo factorial. Cuando se quiera construir

contrastes en modelos sin interacción, para probar hipótesis respecto a un factor, se debe determinar si el modelo es conectado, pues cuando esto sucede, se pueden plantear hipótesis semejantes a las que se construyen con experimentos balanceados como puede verse en Murray & Smith (1985).

6.4. Sumas de cuadrados tipo IV.

Para calcular las sumas de cuadrados tipo IV en modelos con interacción se parte de la expresión general de las hipótesis lineales presentada en (15). Para los modelos sin interacción se puede usar el modelo reparametrizado o el modelo de medias de celda modificado.

Con los modelos reparametrizados la ecuación (15) se calcula como:

$$SCQ_3 = \left(L^t\hat{\mu}\right)^t\left[\left(L^tM^{-1}\right)\left(Z^tZ\right)^-\left(L^tM^{-1}\right)^t\right]^{-1}\left(L^t\hat{\mu}\right) \tag{25}$$

donde M^{-1} no relaciona la interacción entre los factores y $\hat{\mu} = M^{-1}\hat{\delta}$, con $\hat{\delta}$, obtenido a partir de la ecuación (12).

Cuando se trabaja con el modelo de medias de celda modificado con la restricción (2) sobre las interacciones, se obtiene el modelo (6) a partir del cual se encuentran las estimaciones de las medias de celda (7) y (8), lo mismo que sus varianzas y covarianzas (9) a (11), con las cuales se construye la matriz de varianzas y covarianzas llamada $\hat{\Sigma}$. En el cálculo de la suma de cuadrados se requiere el reordenamiento de los coeficientes en la matriz de contrastes L^t, en forma semejante a como se hace la partición de μ, teniendo así la siguiente expresión para la suma de cuadrados tipo IV

$$SCQ_4 = \left(L^t\hat{\mu}\right)^t\left[L^t\ \hat{\Sigma}\ L\right]^{-1}\left(L^t\hat{\mu}\right) \tag{26}$$

7. Ejemplo numérico

En esta sección se ilustran los resultados teóricos desarrollados anteriormente cuando los datos se ajustan con el modelo de medias de celda. El conjunto de datos propuesto es ficticio y sólo se busca ilustrar la manera de aplicar estos desarrollos.

Los datos fueron caracterizados por el modelo $y_{ijk} = \mu_{ij} + e_{ijk}$; $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, b$ y $k = 1, 2, \ldots, n_{ij}$; $n_{ij} \geqslant 0$

donde y_{ijk}, corresponde a la k - ésima observación asociada al nivel i-ésimo del factor A en el nivel j-ésimo del factor B; μ_{ij} la media de celda poblacional asociada la nivel i-ésimo del factor A en el nivel j-ésimo del factor B y e_{ijk} es una componente de errores aleatorios el cual se supone sigue una distribución normal. Si además $a = 3$ y $b = 4$, según el siguiente arreglo

Tabla 1. Estructura de datos para un arreglo Factorial 3 x 4 con celdas vacías

i/j	**1**	**2**	**3**	**4**
1	1	2	0	0
2	1	0	1	2
3	0	1	1	1

Las filas corresponden a las modalidades del factor A, y las columnas a las del factor B. La información consignada en cada entrada de la tabla 1 corresponde a la cantidad de observaciones (n_{ij}). Para esta Tabla se ilustra el cálculo de las sumas de cuadrados y las funciones estimables tipo III y IV asociadas al factor A.

Si además suponemos los siguientes datos como las respuestas de interés, entonces la matriz W es

$$y = \begin{bmatrix} 3 & 3 & 4 & 4 & 5 & 5 & 6 & 6 & 7 & 7 \end{bmatrix}^t$$

$$W = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Para obtener las hipótesis y sumas de cuadrados tipo III, con el modelo de medias de celda modificado, inicialmente se considera la ecuación (19). Las matrices V_1 y V_2 para el factor A se construyen teniendo en cuenta las siguientes restricciones:

$$G_2 = \left[\begin{array}{c} G_B = G_{21} \\ G_{AB} = G_{22} \end{array} \right]_{9\times 12} \text{ y } \; G_1 = \left[\begin{array}{c} G_A = G_{11} \\ G_{AB} = G_{12} \end{array} \right]_{8\times 12}$$

Para que en las matrices G_2 y G_1 se satisfaga que G_{22} y G_{12} sean no singulares, se deben construir teniendo en cuenta que en G_{22} se agrupen las medias de las celdas $\{11, 12, 13, 21, 22, 23, 31, 32, 33\}$ y en la matriz G_{12} se agrupen las medias de celda $\{11, 12, 13, 14, 21, 22, 23, 24\}$; las matrices G_{21} y G_{11} se construyen respectivamente con los complementos de cada conjunto de medias de celda agrupados por G_{22} y G_{12}. A partir de estas restricciones, se calculan las matrices V_2 y V_1, siguiendo los pasos (3) a (6).

Una vez construidas las matrices G_2 y G_1, se construye H_1 de orden 3×12, con rango $ran(H_1) = 2$, a partir de la ecuación (19).

Para construir las funciones estimables tipo III asociadas al factor A es necesario obtener las interacciones efectivas. Para ello, se construyen los contrastes efectivos:

$$G_{AB} = \begin{array}{c} \begin{array}{cccccccccccc} \mu_{11} & \mu_{12} & \mu_{13} & \mu_{14} & \mu_{21} & \mu_{22} & \mu_{23} & \mu_{24} & \mu_{31} & \mu_{32} & \mu_{33} & \mu_{34} \end{array} \\ \left[\begin{array}{cccccccccccc} 1 & -1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 \end{array} \right] \end{array}$$

Se puede observar que estos dos vectores son linealmente independientes y generan el subespacio de las interacciones. Posteriormente se toma cada uno de los vectores fila de la matriz H_1 y se ortogonalizan sobre este subespacio, obteniendo la matriz que corresponde con los contrastes tipo III para el factor A, es decir,

$$L_A^t = \begin{array}{c} \begin{array}{cccccccccccc} \mu_{11} & \mu_{12} & \mu_{13} & \mu_{14} & \mu_{21} & \mu_{22} & \mu_{23} & \mu_{24} & \mu_{31} & \mu_{32} & \mu_{33} & \mu_{34} \end{array} \\ \left[\begin{array}{cccccccccccc} 0{,}4 & 0{,}6 & 0 & 0 & -0{,}4 & 0 & 0{,}2 & 0{,}2 & 0 & -0{,}6 & -0{,}2 & -0{,}2 \\ -0{,}2 & 0{,}2 & 0 & 0 & 0{,}2 & 0 & 0{,}4 & 0{,}4 & 0 & -0{,}2 & -0{,}4 & -0{,}4 \end{array} \right] \end{array}$$

La suma de cuadrados tipo III, asociada al factor A para este arreglo de datos fue

$$SCQ_A = \left(L_A^t \hat{\mu}\right)^t \left[L_A^t \left(W^t W\right)^- L_A\right]^{-1} \left(L_A^t \hat{\mu}\right) = 8{,}147541$$

Con celdas vacías las hipótesis tipo III pueden no ser de mucha utilidad para el investigador; esto se puede ver en los contrastes obtenidos tanto en el modelo superparametrizado como en el modelo de medias de celda.

Por otro lado, cuando se trabaja con el modelo de medias de celda reparametrizado, para la construcción de las hipótesis y sumas de cuadrados tipo III, se obtiene la matriz

$$M^{-1} = \left[J_a \otimes J_b \mid \Delta_a^t \otimes J_b \mid J_a \otimes \Delta_b^t \mid \Delta_a^t \otimes \Delta_b^t\right]$$

con $a = 3$ y $b = 4$; las columnas de la matriz $Z = WM^{-1}$, que corresponden a este modelo son:

$$Z = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & -1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & -1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & -1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & -1 & -1 & -1 & 1 & -1 \\ 1 & 0 & 1 & -1 & -1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 0 & 1 & 0 & 1 & 0 \\ 1 & -1 & -1 & 0 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 \end{bmatrix}$$

Con la matriz Z anterior se calcula $\mathbf{A}$ definida en (20) y se prueba la hipótesis (21), para cada uno de los factores e interacciones. La matriz de contrastes y la suma de cuadrados tipo III obtenida a partir de este método es la misma que la de la sección anterior.

Así por ejemplo, para el factor A, la matriz $\mathbf{A} = \mathbf{R}^t\mathbf{R}$, es:

$$\mathbf{R} = \begin{bmatrix} 0{,}57 & 0{,}20 & 0{,}20 & -0{,}57 & -0{,}17 & -0{,}08 & -0{,}08 & -0{,}40 & 0{,}17 & 0{,}17 \\ 0 & 0{,}22 & 0{,}22 & 0 & 0{,}44 & 0{,}22 & 0{,}22 & -0{,}44 & -0{,}44 & -0{,}44 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Finalmente los contrastes asociados al factor A con el modelo de medias de

celda reparametrizado, y las respectivas sumas de cuadrados coinciden exactamente con las obtenidas con el modelo de medias de celdas modificado.

En la construcción de las hipótesis y sumas de cuadrados tipo IV, se tuvo en cuenta la definición 5; y en concreto para el factor A se obtuvieron los siguientes contrastes iniciales:

$$\left\{ \begin{array}{c} \mu_{24} - \mu_{34} \\ \mu_{23} - \mu_{33} \\ \mu_{12} - \mu_{32} \end{array} \right\}$$

a partir de los cuales se obtienen los contrastes tipo IV para el factor A:

$$L_A^t = \begin{array}{c} \begin{array}{cccccccccccc} \mu_{11} & \mu_{12} & \mu_{13} & \mu_{14} & \mu_{21} & \mu_{22} & \mu_{23} & \mu_{24} & \mu_{31} & \mu_{32} & \mu_{33} & \mu_{34} \end{array} \\ \left[\begin{array}{cccccccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0{,}5 & 0{,}5 & 0 & 0 & -0{,}5 & -0{,}5 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \end{array} \right] \end{array}$$

y la suma de cuadrados tipo IV para el factor A, obtenida a partir de la expresión (25) da:

$$SCQ_A = 7{,}6666667$$

Los resultados de esta suma de cuadrados se resumen en la Tabla 2.

Tabla 2. Resumen del análisis tipo III y IV para el arreglo factorial $\mathbf{3 \times 4}$

Factor	**g.l.**	$\mathbf{R}_{\mathrm{III}}(.\mid.)$	$\mathbf{SS_{III}}$	$\mathbf{SS_{IV}}$
A	2	$R(A \mid B, AB)$	8.147541	7.666666
B	3	$R(B \mid A, AB)$	2.112449	2.088235
AB	2	$R(AB \mid A, B)$	0.081967	0.081967

8. Ejemplo sobre conectés en el modelo de medias de celda

Se empleará la tabla 1 para mostrar cómo la conectés permite al investigador probar las hipótesis de interés, como si estuviera analizando un modelo con todas las celdas llenas.

La conectés en el modelo de clasificación a dos vías sin interacción.

El investigador puede suponer que la interacción entre los factores A y B puede no ser de interés en su estudio y por lo tanto no tenerla en cuenta para

el análisis. Para ello, puede considerar la matriz de contrastes que define la interacción como una matriz G semejante a la planteada en (2) y tomar $g = 0$ para este caso.

Esta restricción lleva a la estimación de todas las celdas como puede verse a continuación:

$$\hat{\mu}_C = M^{-1}\hat{\delta} = \begin{bmatrix} 3{,}049 \\ 3{,}475 \\ 4{,}262 \\ 4{,}540 \\ 3{,}950 \\ 4{,}377 \\ 5{,}163 \\ 5{,}442 \\ 5{,}622 \\ 6{,}049 \\ 6{,}836 \\ 7{,}114 \end{bmatrix}$$

El investigador podrá generar contrastes que le permitan concluir respecto a un factor principal; por ejemplo, si quiere probar la significancia del factor A puede tomar el contraste:

$$L_A^t = \begin{array}{c} \begin{matrix} \mu_{11} & \mu_{12} & \mu_{13} & \mu_{14} & \mu_{21} & \mu_{22} & \mu_{23} & \mu_{24} & \mu_{31} & \mu_{32} & \mu_{33} & \mu_{34} \end{matrix} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \end{bmatrix} \end{array}$$

y emplear (15) para obtener la suma de cuadrados, construir (14) y probar la significancia del factor A. La suma de cuadrados asociada al factor A es

$$SCQ_A = 8{,}0846995$$

la cual corresponde con la suma de cuadrados tipo IV para el factor A en el modelo sin interacción. Para el factor B ocurre algo semejante.

9. Conclusiones

Se han propuesto dos métodos para la obtención de la base de las funciones estimables y las sumas de cuadrados tipo III y un método para la obtención

de las funciones estimables y las sumas de cuadrados tipo IV, a través de los modelos de medias de celda desbalanceados con celdas vacías.

Las funciones estimables tipo III se obtuvieron para modelos de clasificación a dos vías con interacción en presencia de celdas vacías y para cualquier modelo a k-vías con interacción en ausencia de celdas vacías. Para modelos a k-vías de clasificación sin interacción siempre es posible obtener la base de las funciones estimables.

Las funciones estimables tipo IV se pueden obtener para cualquier modelo a k-vías con interacción o sin interacción en presencia o ausencia de celdas vacías.

Referencias

[1] BRYCE, G. R; SCOTT, D. T. & CARTER, M. W. *Estimation and hypothesis testing in liner models - A reparameterization approach to the cells means model.* En: Communications in Statistics. Vol 2. 1980. P 131 - 150.

[2] FREUND, R. J. *The case of missing Cells.* En: The American Statistician. Vol 34. 1980. P 94 - 98.

[3] GOODNIGHT, J. H. *The sweep operator : Its importance to statistical computing.* En: Proceedings of the Eleventh Interface of Statistics and Computer Science. Institute of Statistics, N. C. State University., Raleigh, N. C. 1978.

[4] HOCKING, R. R. *Methods and Applications of Linear Models.* John Wiley & Sons. N.Y. 1996.

[5] HOCKING, R. R; SPEED, F. M. & COLEMAN, A. T. *Hypotheses to be tested with unbalanced data.* En: Communications in Statistics. Vol 2. 1980. P 117 - 129.

[6] HUDSON, G. F. & SEARLE, S. R. *Hypothesis testing with type IV sums of squares of the computer routine SAS GLM.* En: Proceedings, 7th Annu. SAS User

[7] JENNINGS, E. & WARD, J.H. *Hypothesis Identification in the case of missing cell.* En: The American Statistician. Vol 36. 1982. P 25 - 27.

[8] ———. *Los Modelos de Medias de Celda, una herramienta fundamental en la Estadística Industrial.* Simposio de Estadística. Rionegro, Antioquia. 1999.

[9] MELO, C. E. *Hipótesis Efectivas en Modelos de medias de celda, construcción a través del método de Murray - Smith*. Bogotá. 2000. Trabajo de Especialización en Estadística. Universidad Nacional de Colombia. Departamento de Matemáticas y Estadística.

[10] MURRAY, L.W. & SMITH, D.W. *Estimability, Testability and Connectedness in the cell means model*. En: Communications in Statistics. Vol 14. 1985. P 1889 - 1917.

[11] RAO, C. R. *On the lineal combination of observations an the general theory of least Squares*. Sankhya. 1945. P 237 - 256.

[12] SEARLE, S. R. *Linear Models for Unbalanced Data*. John Wiley & Sons. N. Y. 1987.

[13] SEARLE, S. R. *Arbitrary Hypothesis in Linear Models with Unbalanced Data*. Communications in Statistics - Theory and Methods. A(9)2. P 181-200. 1980.