Revista Colombiana de Estadística
Volumen 27 N° 1. Págs. 43 *a* 65. Junio 2004

Metodología para estimar celdas vacías con el modelo de medias de celdas en diseños conectados

Sandra E. Melo Martínez[*]
Luis A. López Pérez[**]
Oscar O. Melo Martínez[***]

Resumen

En este artículo se proponen dos métodos para estimar celdas vacías y para imputar datos faltantes en diseños conectados en un experimento multifactorial, a partir del modelo de medias de celdas modificado.

La primera propuesta tiene como soporte teórico el método de covarianza de Bartlett con base en el modelo de medias de celdas con restricción de efectos; la segunda se basa en la estimación por mínimos cuadrados con restricciones de no interacción. Además, se presenta una expresión general para estimar celdas con información faltante, con cualquiera de los métodos propuestos. Se demuestra que el estimador obtenido por el método de covarianza coincide con el de mínimos cuadrados con la restricción de no interacción entre los efectos. Se propone una metodología para determinar el número mínimo de celdas que se deben estimar para que el diseño se conecte, y además se muestra cuáles celdas son las que se deben imputar para lograr la conexión. Una vez hecha la imputación de la información faltante, se plantea el análisis de varianza con los ajustes respectivos en las sumas de cuadrados. Por último, se ilustra la aplicación de las propuestas con un ejemplo numérico a tres vías de clasificación, sin interacción, con efectos fijos.

Palabras claves: *Diseños conectados, modelos de medias de celda, funciones estimables, modelo con restricción, diseños desbalanceados.*

[*]Estadística. Universidad Nacional de Colombia. E-mail: semelom@unal.edu.co
[**]Profesor Asociado. Departamento de Estadística. Universidad Nacional de Colombia. E-mail: lalopezp@unal.edu.co
[***]Profesor Asociado. Departamento de Estadística. Universidad Nacional de Colombia. E-mail: oomelom@unal.edu.co

Abstract

In this paper we propose two methods to estimate empty cells and to impute missing data in connected designs. The first is supported on Bartlett's covariance method on the basis of cell means model with effect restriction; the second is based on the least squares estimation method with no interaction. A general expression is proposed to estimate cells with missing information. The equality of the estimators obtained by these methods is demonstrated and a lower bound to the number of cells needed to be estimated, so that the design is connected, is calculated. The adjusted anova is presented and finally, a 3-way no interaction fixed effects model is used to illustrate the two methods.

Keywords: *Connected designs, cell means models, estimable functions, restriction model, unbalanced designs.*

1. Introducción

En el estudio de diseños de experimentos se hace uso de modelos superparametrizados y modelos de medias de celdas. La estructura de los datos puede ser balanceada o desbalanceada, y aún más, de acuerdo con la naturaleza del problema y la complejidad del experimento, puede haber celdas vacías, cuya presencia, por lo general, no es planteada por el investigador y se debe más a causas ajenas al experimento que a las condiciones propias del material experimental.

Cuando hay celdas vacías, no todas sus medias asociadas son estimables y, en consecuencia, las hipótesis sobre combinaciones lineales de medias pueden no tener ningún interés práctico o no tener solución dentro de la teoría general de las hipótesis lineales. Sin embargo, si se imponen restricciones sobre las interacciones de efectos, es posible conectar el diseño y tener todas las hipótesis efectivas de interés práctico para la investigación.

Imponer restricciones para lograr que el diseño se conecte, debe ser una tarea conjunta entre el experimentador y el estadístico. En un arreglo factorial con tres factores, por ejemplo, puede suceder que la condición de no interacción entre los tres factores sea suficiente para que el diseño se conecte; si esto no se logra, se debe imponer restricciones de no interacción sobre algunos efectos dobles procurando así conectar el diseño para poder estimar algunas medias de celdas donde inicialmente no se tenía información.

La disponibilidad de paquetes estadísticos ha permitido manejar, dentro del marco del modelo lineal general, el análisis de varianza con modelos multifactoriales con presencia de celdas vacías. Sin embargo, se debe tener extremo

cuidado con el uso de aquellos paquetes que no hacen referencia explícita a las expresiones utilizadas para calcular las diferentes sumas de cuadrados, lo cual dificulta la identificación de las hipótesis asociadas a los factores e interacciones del arreglo.

En estudios experimentales multifactoriales, las celdas vacías generan problemas para identificar las funciones estimables. Desde la década de los 60 se ha venido trabajando en este tema, destacándose el trabajo de Weeks y Williams (1964) como marco para la identificación de funciones estimables en modelos a n-vías de clasificación. A pesar de los avances logrados en el tema, como se muestra en Dodge (1985), no hay consenso en cuanto a las metodologías que conllevan tanto a la identificación de funciones estimables, como a la estimación de celdas vacías.

Cuando hay información con celdas vacías, surge la preocupación por establecer con claridad el conjunto de medias de celdas estimables, y la base de las funciones estimables. La naturaleza de estas funciones depende de si el modelo puede conectarse, puesto que esta propiedad se relaciona con la cantidad de información que se pierde por las celdas vacías y la que queda disponible para el análisis.

El concepto de *conectés* fue introducido por Bose y Srivastava (1964) quienes la definieron para modelos a dos vías de clasificación, estableciendo una cadena entre bloques y tratamientos. Para el modelo de bloques, por ejemplo, dos celdas cualesquiera, (i, j) y (r, t), se dicen *conectadas* si se puede pasar de una a la otra a lo largo de una fila (o columna) de celdas observadas; el camino puede ser descrito como:

$$(i, j) \rightarrow (i, v) \rightarrow (u, v) \rightarrow (u, t) \rightarrow (r, t),$$

donde, todas las celdas en el camino contienen información. Si esta condición se tiene para todas las parejas de celdas observadas, el diseño se dice *conectado*.

Para el caso de información con celdas vacías, Searle (1971) y Graybill (1976) describen la *conectés* para modelos a dos vías de clasificación, donde es posible llevar a cabo la estimación de las medias correspondientes a celdas vacías, y además proponen una metodología simple para determinar la *conectés*. Searle (1987) presenta para el caso de dos factores la *conectés geométrica*, la cual suministra un método directo para determinar si los datos de un modelo a dos vías de clasificación son conectados. En el caso de la *conectés* a n vías de clasificación Weeks y Williams (1964), Murray y Smith (1985), Birkes (1976) y Dodge (1985) describen diferentes métodos para encontrar si un conjunto de datos obtenidos a partir de arreglos factoriales está conectado; además ilustran la forma de encontrar la base generadora de las funciones estimables, cuando

se tienen arreglos con celdas vacías.

En este artículo inicialmente se estudia el método de *conectés* descrito en Murray y Smith (1985), cuando se tiene el modelo de medias de celdas presentado en Hocking (1985). La técnica de conexión se basa en el modelo de medias de celdas modificado, imponiendo restricciones de no interacción entre efectos, con lo cual se reduce su dimensionalidad. En el modelo con restricciones, se examina el rango de la matriz diseño; si es de rango completo, el diseño es conectado; en el caso contrario, se debe determinar el número de celdas por estimar para conectar el diseño y realizar luego los análisis respectivos sobre los diferentes factores de interés. En este artículo se proponen dos métodos estimar celdas vacías, sujetos a la restricción de no interacción.

2. Conceptos básicos de modelos lineales

En este apartado se muestran algunas ideas básicas sobre modelos lineales superparametrizados, como marco teórico inicial para el desarrollo de esta propuesta.

El modelo lineal presentado en Graybill (1961, 1976), Searle (1971, 1987), Hocking (1985), está dado por:

$$Y = X\theta + e, \tag{1}$$

donde: $Y_{n\times l}$ es un vector de variables aleatorias, $X_{n\times p}$ es una matriz conocida (matriz de diseño) de rango $k \leq \text{mín}\{n, p\}$, $\theta_{p\times 1}$ es un vector de parámetros desconocidos y $e_{n\times 1}$ es un vector de variables aleatorias no observables. Se supone que $e \sim N(0, \sigma^2 I)$.

El modelo de medias de celdas estudiado, entre otros, por Speed (1978), Hocking (1985, 1996), Searle (1987) y Murray y Smith (1985), está dado por:

$$Y = W\mu + e, \tag{2}$$

sujeto a

$$G\mu = g, \tag{3}$$

donde, Y y e son vectores descritos en (1), W es una matriz de incidencia de orden $n \times p$ que asocia las observaciones con sus valores medios e indica el número de observaciones con media $\mu_{ij\cdots s}$, μ es un vector de p medias poblacionales desconocidas y G una matriz $s \times p$, de constantes desconocidas de rango s, la cual representa las relaciones lineales conocidas sobre las medias de celdas, especificando contrastes sin interacción.

Si no hay celdas vacías, W tiene rango columna completo. Si las hay, entonces W tiene una columna igual a cero por cada celda vacía y μ mantiene su estructura; es decir, mantiene el mismo número de parámetros como si todas las celdas hubiesen sido observadas. En la siguiente sección se presenta el método de Murray y Smith, soporte para el desarrollo de los métodos propuestos.

3. Método de Murray y Smith

El investigador, por lo general, está interesado en saber si los parámetros pueden ser estimados con la información disponible. En el caso de celdas vacías, la estimación de parámetros y la formulación de hipótesis sobre los parámetros, depende de que el diseño se pueda conectar, al imponer restricciones específicas sobre éste. Si los datos del experimento son conectados con respecto al modelo original, entonces pueden ser estimados y además, el investigador puede plantear todas las hipótesis de interés práctico para la investigación. El método basado en el modelo de medias de celdas modificado presenta adicionalmente una prueba simple para determinar si el diseño es conectado.

El modelo de medias de celda modificado es más útil que el modelo (2) y (3) cuando hay celdas vacías. Se obtiene sustituyendo el conjunto de contrastes $G\mu = 0$, directamente dentro del modelo, con lo cual se reduce su dimensionalidad. Para ello se hace un reordenamiento de las columnas de G como $G = [G_1 \mid G_2]$, donde G_2 es una matriz $s \times s$ de rango s y G_1 una matriz de orden $s \times (p - s)$. Las filas de μ se particionan en correspondencia con la partición de G: $\mu' = [\mu_1' \mid \mu_2']$. Entonces, el contraste $G\mu = 0$ implica que

$$G\mu = G_1\mu_1 + G_2\mu_2 = 0. \tag{4}$$

Como G_2 es de rango completo, existe una solución única para μ_2 en términos de μ_1, dada por:

$$\mu_2 = -G_2^{-1}G_1\mu_1. \tag{5}$$

G_2 se elige arbitariamente, pero debe ser una matriz no singular. En muchas situaciones experimentales es fácil encontrar el conjunto de medias para μ_1 y μ_2, lo cual lleva a que la matriz G_2 sea no singular y por lo tanto invertible.

La partición de μ es independiente de los datos obtenidos; en particular, es independiente del número de celdas vacías (m) y de su localización. La partición sólo depende de G y de las relaciones lineales entre las medias involucradas; así que esta partición está basada en cómo se concibió el experimento y no en cómo se realizó.

La partición (4) lleva a un reordenamiento de la matriz de incidencia W, en concordancia con las particiones de G y de μ, es decir, $W = [W_1 \mid W_2]$; así, al sustituir (5) en (2), se llega al modelo de celdas modificado, más útil que (2) y (3) cuando hay celdas vacías. Por otro lado, si se sustituye (4) directamente dentro del modelo, se llega al modelo particionado:

$$Y = [W_1 \mid W_2] \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + e \tag{6}$$

$$= [W_1 \mid W_2] \begin{pmatrix} I \\ -G_2^{-1} G_1 \end{pmatrix} \mu_1 + e \tag{7}$$

$$= V\mu_1 + e, \tag{8}$$

donde $V = W_1 - W_2 G_2^{-1} G_1$, es de tamaño $n \times (p - s)$. En (8) si el rango de V es $(p - s)$, entonces $V'V$ es no singular y se puede aplicar el método usual de mínimos cuadrados para modelos de rango completo sin restricción para la solución de μ_1, y entonces el $MELI(\mu_1)$ es

$$\widehat{\mu}_1 = (V'V)^{-1} V'Y, \tag{9}$$

y el $MELI(\mu_2)$ es

$$\widehat{\mu}_2 = -G_2^{-1} G_1 \widehat{\mu}_1. \tag{10}$$

Cuando no hay celdas vacías, el rango de V es $(p - s)$, las soluciones (9) y (10) siguen siendo los únicos $MELI$ de μ_1 y μ_2 y la solución de parámetros es entonces $\widehat{\mu} = [\widehat{\mu}_1' \mid \widehat{\mu}_2']'$. Si el rango de V es menor que $(p - s)$, escencialmente se tiene el mismo problema del modelo superparametrizado cuando la matriz X no es de rango completo. En Murray y Smith (1985) y Hocking (1985) se muestra que cuando hay celdas vacías, el modelo descrito por (2), (3) y (8) es *conectado* si V tiene rango columna completo.

Es conveniente tener en cuenta los siguientes puntos cuando se hace uso del modelo de medias de celda modificado:

i. G puede obtenerse de infinitas formas, para un conjunto específico de contrastes.

ii. Según Murray (1986), puede ser complicado encontrar la submatriz G_2 cuando el número de contrastes es grande, lo cual complica también la partición de la matriz G.

3.1. Estimabilidad y conectés

El concepto de estimabilidad en el modelo de medias de celdas es simple y se puede sintetizar en la siguiente pregunta: *¿hay suficiente información disponible para estimar funciones lineales de las medias de celdas?* La respuesta a esta pregunta es muy sencilla si todas las celdas son observadas; sin embargo, cuando hay celdas vacías y se encuentran relaciones apropiadas entre las medias de las celdas teniendo en cuenta la restricción $G\mu = g$, entonces el problema importante se centra en la estimación. El método relaciona la estimabilidad de las medias de celda con el concepto de conectés y posteriormente se desarrolla una prueba simple para la estimabilidad de μ cuando el diseño se ha conectado, basado en el modelo (8). La siguiente definición ilustra el concepto de conectés según Hocking (1985, 1996).

Definición 3.1. Un experimento conformado por un conjunto de datos y un modelo de medias de celdas asociado, (modelo M) como en (8), se dice *conectado* si μ es linealmente estimable, de forma única.

Se puede notar en las definiciones dadas por Searle (1971, 1987) y Weeks y William (1964) que la conectés se restringe a modelos de efectos principales, mientras que la definición 3.1 se aplica a cualquier modelo de medias de celda con o sin restricción (Murray y Smith, 1985).

Un conjunto particular de datos puede ser conectado por un modelo y no por otro. Por ejemplo, en el modelo de clasificación con tres factores, se tienen varios niveles de conectés dependiendo de la restricción necesaria para alcanzar la estimabilidad de μ. La condición de no interacción entre los factores puede ser suficiente para obtener la estimación de todos los μ_{ijk} cuando hay celdas vacías en un diseño a tres vías de clasificación. Si ésta no es suficiente, entonces la imposición de restricciones sobre una o más de las interacciones de dos factores puede conducir a la estimabilidad.

Un nuevo criterio para establecer la conectés, basado en el modelo de medias de celdas modificado (8) y la definición 3.1, es objeto del siguiente teorema, propuesto y demostrado por Murray y Smith (1985):

Teorema 3.1. Para el modelo (2) y (3) el experimento es *conectado* si y sólo si $V = W_1 - W_2 G_2^{-1} G_1$ tiene rango columna completo, es decir, el rango de V es $p - s < n$.

Si V tiene rango columna completo, entonces el experimento es conectado, μ es linealmente estimable en su totalidad y el análisis original puede ser llevado a cabo como se planeó. Si V no tiene rango columna completo, entonces el

experimento no es conectado, μ no es linealmente estimable y el análisis original no puede realizarse. En este caso el modelo de medias de celdas (2) y (3) no es de rango completo.

3.2. Expresión matricial para la restricción de no interacción

En Murray y Smith (1985) se presenta una expresión matricial conveniente para la restricción de interacción en el modelo de dos vías de clasificación:

$$D_a \otimes D_b, \tag{11}$$

donde D_i está definido por:

$$D_i = (I_{i-1} \mid -J_{i-1}). \tag{12}$$

I_{i-1} es la matriz identidad de dimensión $i-1$ y J_{i-1}, el vector columna de unos, de longitud $i-1$. El uso de estas matrices D_i es conveniente para describir las restricciones que involucran interacciones de orden dos o superior.

Esta idea puede ser generalizada para n-vías de clasificación con factores con m_i niveles, $i = 1, 2, \ldots, n$.

Sea

$$M_i = \begin{cases} D_i & \text{si la interacción involucra el factor } i, \\ J_i' & \text{en caso contrario,} \end{cases} \tag{13}$$

entonces la matriz de restricciones (3) para una interacción dada es:

$$G = M_1 \otimes M_2 \otimes \cdots \otimes M_n. \tag{14}$$

4. Métodos propuestos para estimar las celdas vacías

En esta sección se presentan dos métodos diferentes para estimar las celdas vacías teniendo como base, en primer lugar, el modelo de covariable con restricción a través del modelo de medias de celdas. El segundo método, permite la estimación de datos faltantes cuando se imponen restricciones de no interacción al modelo, pero la estimación se hace con mínimos cuadrados ponderados. En los dos casos, para que el método sea aplicado, se requiere que el diseño sea conectado; cuando no lo es, se deben tener en cuenta algunas consideraciones adicionales.

4.1. Método de estimación por mínimos cuadrados con restricción

Para implementar el método, se considera el modelo presentado en (2) suponiendo que hay m celdas vacias; luego, se realiza una partición del vector $Y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$, donde y_1 se arregla de tal forma que contenga la información observada y y_2 sea el vector asociado con la información faltante. En forma similar se hace la partición de la matriz W, de tal forma que las matrices de diseño W_1 y W_2 se asocien con la información observada y faltante respectivamente. Se puede escribir ahora (2) como:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix} \mu + e, \tag{15}$$

sujeto a la restricción (4).

Se hace luego una partición adicional de las submatrices W_1 y W_2, que corresponda con la partición de la matriz de restricciones G. Así, (15) se escribe como:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + e,$$

donde $W_{11}, W_{12}, W_{21}, W_{22}$ son de orden $(n-m)\times(p-s)$, $(n-m)\times s$, $m\times(p-s)$, $m \times s$ respectivamente. Esta partición lleva a los siguientes modelos:

$$y_1 = W_{11}\mu_1 + W_{12}\mu_2 + e_1 \tag{16}$$

$$y_2 = W_{21}\mu_1 + W_{22}\mu_2 + e_2, \tag{17}$$

sustituyendo (5) en (16) se obtiene:

$$y_1 = W_{11}\mu_1 - W_{12}G_2^{-1}G_1\mu_1 + e_1,$$

la cual también puede expresarse como:

$$y_1 = V_1\mu_1 + e_1, \tag{18}$$

donde, $V_1 = W_{11} - W_{12}G_2^{-1}G_1$ es una matriz de orden $(n-m)\times(p-s)$, V_1 es la submatriz de indicadores con las restricciones de no interacción asociada a los datos observados.

El estimador de mínimos cuadrados de μ_1 en el modelo (18) es:

$$\widehat{\mu}_1 = (V_1'V_1)^{-1}V_1'y_1. \tag{19}$$

Por otra parte, sustituyendo (5) en (17), se llega al modelo:

$$y_2 = W_{21}\mu_1 - W_{22}G_2^{-1}G_1\mu_1 + e_2,$$

que puede reducirse a:

$$y_2 = V_2\mu_1 + e_2, \tag{20}$$

donde $V_2 = W_{21} - W_{22}G_2^{-1}G_1$ es una matriz de orden $m \times (p-s)$, que corresponde a la submatriz de indicadores con las restricciones de no interacción correspondiente a los datos no observados. A partir del modelo (20) y del estimador para μ_1, obtenido en (19), se encuentra la expresión alternativa para la estimación de información faltante, cuando se imponen restricciones en el modelo:

$$\widehat{y}_2 = V_2\widehat{\mu}_1 = V_2(V_1'V_1)^{-1}V_1'y_1. \tag{21}$$

El estimador presentado en (21) depende del vector de datos observados y de las matrices V_2 y V_1.

Es importante determinar cuántas celdas deben ser estimadas para que el diseño se pueda conectar, pues la *conectés* del diseño garantiza la posibilidad de aplicar los resultados encontrados en (21). Para ello, se realiza una nueva partición del vector de datos faltantes y_2. De esta manera se determinan las celdas que no se pueden conectar. El conjunto de estas celdas indica las que deben ser estimadas para lograr la *conectés*. Así, una vez lograda la conexión, se puede utilizar el método de estimación por mínimos cuadrados sujeto a la restricción para estimar estas celdas vacías o el método de covariable sujeto a la restricción de no interacción, propuestos en este artículo.

Si se realiza la partición $y_2 = \begin{pmatrix} y_3 \\ y_4 \end{pmatrix}$, donde y_3 corresponde a las celdas que se pueden conectar y y_4 a las que no, entonces el modelo (20) se puede escribir como:

$$y_2 = \begin{pmatrix} y_3 \\ y_4 \end{pmatrix} = V_2\mu_1 + e_2. \tag{22}$$

Si ahora se particiona la matriz V_2 de forma que corresponda a la partición hecha para y_2, el modelo anterior se puede escribir como:

$$\begin{pmatrix} y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} V_3 \\ V_4 \end{pmatrix}\mu_1 + \begin{pmatrix} e_3 \\ e_4 \end{pmatrix}, \tag{23}$$

o en forma explícita los modelos:

i) $y_3 = V_3\mu_1 + e_3$. Este modelo se puede conectar y, en consecuencia, todas las celdas vacías son estimables.

ii) $y_4 = V_4\mu_1 + e_4$. Con este modelo se debe determinar el conjunto de celdas no conectadas para estimarlas y lograr la *conectés*. La estimación de μ_1 se encuentra con $\widehat{\mu}_1^* = (V_1'V_1)^- \; V_1' \; y_1$, donde $(V_1'V_1)^-$ es una inversa generalizada. La solución para las celdas desconectadas es entonces:

$$\widehat{y}_4 = V_4(V_1'V_1)^- \; V_1' \; y_1. \tag{24}$$

En un diseño desconectado, las estimaciones negativas o cero para los datos faltantes, obtenidas por el proceso de estimación descrito, son las que generan la *desconectés* del diseño. En el caso de diseños conectados, el método propuesto es una buena alternativa para la estimación de los datos faltantes.

4.2. Metodología para determinar el número de celdas por estimar para conectar el diseño

Para determinar el número de celdas vacías que deben ser estimadas para conectar el diseño, es necesario encontrar en (24) el número filas que convierte la matriz $V_1'V_1$ en una de rango completo, ya que siendo singular, el diseño es desconectado. La diferencia entre el número de filas de la matriz $V_1'V_1$ y su rango determina el número de celdas que es necesario estimar para tener un diseño conectado. Además, es posible desarrollar una expresión algebraica que permite determinar las filas de V_2 en el modelo (20) que causan la desconexión del diseño.

Como V_2 es una submatriz de indicadores con las restricciones de no interacción, asociada a los datos no observados, la matriz $V^* = V_2(V_1'V_1)^- \; (V_1'V_1)$ contiene las filas que corresponden a celdas conectadas y no conectadas. Al establecer una correspondencia de filas entre la matriz V_2 y V^*, aquellas cuyos elementos coinciden corresponden a las filas de las celdas que se conectan, en tanto que las filas que no coinciden, corresponden a las filas de las celdas no conectadas.

Por otra parte, $V_3(V_1'V_1)^- \; (V_1'V_1) = V_3$, pues las filas de V_3 son combinaciones lineales de las filas de V_1, ya que en esta matriz se encuentran las celdas que se pueden conectar.

Las filas de la matriz V_3 son precisamente las que corresponden a las celdas conectadas, mientras que el resto son las que generan la desconectés del diseño y por esta razón deben ir en la matriz V_4. Las filas de V_2 que no coincidan con sus correspondientes en la matriz V^* están determinando las celdas se deben ser estimadas para conectar el diseño.

4.3. Método de estimación por covariable sujeto a restricción

En esta sección se presenta el segundo método propuesto, denominado *método de estimación por covariable, sujeto a la restricción de no interacción entre efectos.* Con él se llega a una expresión para la estimación de celdas vacías que coincide con la obtenida por mínimos cuadrados, como se muestra en la sección 4.4.

Se considera el siguiente modelo de medias de celdas con covariable:

$$Y^* = \begin{pmatrix} y_1 \\ 0 \end{pmatrix} = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix} \mu + Z\gamma + e, \tag{25}$$

sujeto a la restricción (4) con $\widehat{\mu}_2$ dado por (10).

En este caso la matriz W se particiona como en la sección 4.1, de tal forma que corresponda con la partición hecha de la matriz de restricciones G, y como por el método de covariable $y_2 = 0$, entonces se tiene el sistema:

$$\begin{pmatrix} y_1 \\ 0 \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} O \\ -I \end{pmatrix} \gamma + e, \tag{26}$$

donde W_{11}, W_{12}, W_{21} y W_{22} son matrices de orden $(n-m)\times(p-s)$, $(n-m)\times s$, $m \times (p-s)$ y $m \times s$ respectivamente.

De (26) y sustituyendo por la expresión obtenida para μ_2 en (5) se tiene que:

$$\begin{pmatrix} y_1 \\ 0 \end{pmatrix} = \begin{pmatrix} W_{11}\mu_1 & -W_{12}G_2^{-1}G_1\mu_1 \\ W_{21}\mu_1 & -W_{22}G_2^{-1}G_1\mu_1 \end{pmatrix} + \begin{pmatrix} O \\ -I \end{pmatrix} \gamma + e. \tag{27}$$

De (27) se obtiene el modelo presentado a continuación:

$$Y^* = \begin{pmatrix} y_1 \\ 0 \end{pmatrix} = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} \mu_1 + Z\gamma + e, \tag{28}$$

donde, $V_1 = W_{11} - W_{12}G_2^{-1}G_1$ matriz de orden $(n-m)\times(p-s)$, $V_2 = W_{21} - W_{22}G_2^{-1}G_1$ una matriz de orden $m \times (p-s)$ y $Z = \begin{pmatrix} O \\ -I \end{pmatrix}$, con $I = I_m$.

Reescribiendo (28) se obtiene:

$$Y^* = V\mu_1 + Z\gamma + e = \begin{pmatrix} V & Z \end{pmatrix} \begin{pmatrix} \mu_1 \\ \gamma \end{pmatrix} + e. \tag{29}$$

Las ecuaciones normales para el modelo (29) se obtienen premultiplicando por $\begin{pmatrix} V' \\ Z' \end{pmatrix}$, donde se satisface que $\begin{pmatrix} V' \\ Z' \end{pmatrix} e = 0$, obteniendo así:

$$\begin{pmatrix} V' \\ Z' \end{pmatrix} Y^* = \begin{pmatrix} V' \\ Z' \end{pmatrix} \begin{pmatrix} V & Z \end{pmatrix} \begin{pmatrix} \mu_1 \\ \gamma \end{pmatrix} \begin{pmatrix} V'Y^* \\ Z'Y^* \end{pmatrix} = \begin{pmatrix} V'V & V'Z \\ Z'V & Z'Z \end{pmatrix} \begin{pmatrix} \mu_1 \\ \gamma \end{pmatrix},$$

lo cual lleva a los siguientes modelos:

$$V'Y^* = V'V\mu_1 + V'Z\gamma \tag{30}$$

$$Z'Y^* = Z'V\mu_1 + Z'Z\gamma, \tag{31}$$

los cuales, teniendo en cuenta (29) satisfacen que:

(1) $V'Y^* = \begin{pmatrix} V_1' & V_2' \end{pmatrix} \begin{pmatrix} y_1 \\ 0 \end{pmatrix} = V_1'y_1$

(2) $Z'Y^* = \begin{pmatrix} O & -I_m \end{pmatrix} \begin{pmatrix} y_1 \\ 0 \end{pmatrix} = 0$

(3) $V'Z = \begin{pmatrix} V_1' & V_2' \end{pmatrix} \begin{pmatrix} O \\ -I \end{pmatrix} = -V_2'$

(4) $Z'Z = \begin{pmatrix} O & -I \end{pmatrix} \begin{pmatrix} O \\ -I \end{pmatrix} = I_m.$

Entonces (30) y (31) pueden ser escritos de la siguiente manera:

$$V'Y^* = V_1'y_1 = V'V\mu_1 - V_2'\gamma$$
$$Z'Y^* = 0 = -V_2\mu_1 + I_m\gamma.$$

Por otro lado, premultiplicando (30) por $V(V'V)^{-1}$ se obtiene la siguiente expresión:

$$V(V'V)^{-1}V'Y^* = V(V'V)^{-1}V'V\mu_1 + V(V'V)^{-1}V'Z\gamma,$$

de la anterior ecuación se sabe que $P_V = V(V'V)^{-1}V'$ es el proyector ortogonal; por lo tanto, la ecuación anterior se escribe también como:

$$P_VY^* = V\mu_1 + P_VZ\gamma,$$

agrupando términos se obtiene:

$$P_V(Y^* - Z\gamma) = V\mu_1, \tag{32}$$

si se premultiplica (29) por Z', y si $Z'e = 0$, se obtiene la expresión:

$$Z'Y^* = Z'V\mu_1 + Z'Z\gamma. \tag{33}$$

Reemplazando (32) en (33) se tiene que:

$$Z'Y^* = Z'P_V(Y^* - Z\gamma) + Z'Z\gamma,$$

y reagrupando términos se encuentra el sistema de ecuaciones normales:

$$Z'(I - P_V)Y^* = Z'(I - P_V)Z\gamma. \tag{34}$$

Ahora, sea $I - P_V = I - V(V'V)^{-1}V' = \begin{pmatrix} S & M \\ M' & L \end{pmatrix}$, donde S, M y L son matrices de orden $(n-m)\times(n-m)$, $(n-m)\times m$ y $m\times m$ respectivamente, con n el tamaño de Y^*. Como $I - P_V$ es idempotente, se cumple que:

$$\begin{pmatrix} S & M \\ M' & L \end{pmatrix}\begin{pmatrix} S & M \\ M' & L \end{pmatrix} = \begin{pmatrix} S & M \\ M' & L \end{pmatrix},$$

entonces $M'M + L^2 = L$, luego:

$$Z'(I-P_V)Y^* = (O'-I')\begin{pmatrix} S & M \\ M' & L \end{pmatrix}\begin{pmatrix} y_1 \\ 0 \end{pmatrix} = (-M'-L)\begin{pmatrix} y_1 \\ 0 \end{pmatrix} = -M'y_1 \tag{35}$$

$$Z'(I-P_V)Z = (O'-I')\begin{pmatrix} S & M \\ M' & L \end{pmatrix}\begin{pmatrix} O \\ -I \end{pmatrix} = (-M'-L)\begin{pmatrix} O \\ -I \end{pmatrix} = L. \tag{36}$$

Sustituyendo (35) y (36) en (34), se llega a:

$$L\widehat{\gamma} = -M'y_1,$$

y finalmente a

$$\widehat{\gamma} = -L^{-1}M'y_1. \tag{37}$$

La expresión anterior permite estimar la información faltante en las celdas, cuando los datos se ajustan a un modelo de medias de celdas y se imponen restricciones de no interacción. Es importante que se tenga en cuenta que si el diseño no se puede conectar, la matriz L es singular, por lo tanto la expresión (37) no es pertinente.

4.4. Comparación de los métodos propuestos

El teorema 4.1 muestra la equivalencia entre los dos métodos de estimación propuestos en este trabajo para la estimación de información faltante cuando los datos se ajustan con modelos de medias de celdas sujetos a restricciones de no interacción.

Teorema 4.1. El estimador $\widehat{\gamma}$, obtenido por el método de covariables y el estimador $\widehat{y}_2$, obtenido por el método de los mínimos cuadrados, son equivalentes.

Demostración: Teniendo en cuenta que

$$\begin{aligned} I - P_V = I - V(V'V)^{-1}V' &= \begin{pmatrix} S & M \\ M' & L \end{pmatrix} \\ &= \begin{pmatrix} I_1 & O \\ O & I_2 \end{pmatrix} - \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} (V'V)^{-1} \begin{pmatrix} V_1' & V_2' \end{pmatrix} \\ &= \begin{pmatrix} I_1 & O \\ O & I_2 \end{pmatrix} - \begin{pmatrix} V_1(V'V)^{-1}V_1' & V_1(V'V)^{-1}V_2' \\ V_2(V'V)^{-1}V_1' & V_2(V'V)^{-1}V_2' \end{pmatrix} \\ &= \begin{pmatrix} I_1 - V_1(V'V)^{-1}V_1' & -V_1(V'V)^{-1}V_2' \\ -V_2(V'V)^{-1}V_1' & I_2 - V_2(V'V)^{-1}V_2' \end{pmatrix}, \end{aligned}$$

el estimador (37) es igual a:

$$\widehat{\gamma} = -(I_2 - V_2(V'V)^{-1}V_2')^{-1}(-V_2(V'V)^{-1}V_1'y_1), \tag{38}$$

y utilizando el resultado $(D - BA^{-1}U)^{-1}BA^{-1} = D^{-1}B(A - UD^{-1}B)^{-1}$, propuesto por Henderson y Searle (1981), $\widehat{\gamma}$ se reduce a:

$$\begin{aligned} \widehat{\gamma} &= I_2^{-1}V_2((V'V) - V_2'I_2^{-1}V_2)^{-1}V_1'y_1 \\ &= V_2(V'V - V_2'V_2)^{-1}V_1'y_1 \\ &= V_2\left(\begin{pmatrix} V_1' & V_2' \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} - V_2'V_2 \right)^{-1} V_1y_1 \\ &= V_2(V_1'V_1)^{-1}V_1'y_1 = V_2\widehat{\mu}_1 = \widehat{y}_2. \end{aligned}$$

Es decir, se ha comprobado que el estimador por covariable (37) es equivalente al de mínimos cuadrados expuesto en (21).

4.5. Descomposición de las sumas de cuadrados con la información imputada

En esta sección se lleva a cabo la descomposición de las sumas de cuadrados del análisis de varianza a partir de la metodología propuesta, cuando se incluyen

los valores estimados en las celdas donde inicialmente no se tenía información.

Se sabe que la suma de cuadrados total corregida por la media es:

$$SCT = Y'Y - \frac{1}{n}Y'11'Y.$$

La suma total de cuadrados con los datos imputados es:

$$\begin{aligned} SCT_{\text{CDI}} &= \begin{pmatrix} y_1' & \widehat{y}_2' \end{pmatrix} \begin{pmatrix} y_1 \\ \widehat{y}_2 \end{pmatrix} - \frac{1}{n} \begin{pmatrix} y_1' & \widehat{y}_2' \end{pmatrix} \begin{pmatrix} 1_1 \\ 1_2 \end{pmatrix} \begin{pmatrix} 1_1' & 1_2' \end{pmatrix} \begin{pmatrix} y_1 \\ \widehat{y}_2 \end{pmatrix} \\ &= y_1'y_1 + \widehat{y}_2'\widehat{y}_2 - \frac{1}{n}(y_1'1_1 + \widehat{y}_2'1_2)(1_1'y_1 + 1_2'y_2) \\ &= y_1'y_1 + \widehat{y}_2'\widehat{y}_2 - \frac{1}{n}\Big(y_1'1_1 1_1'y_1 + 2\widehat{y}_2'1_2 1_1'y_1 + \widehat{y}_2'1_2 1_2'\widehat{y}_2\Big) , \end{aligned} \tag{39}$$

donde el subíndice CDI significa "con datos imputados".

Cuando no se incluyen las estimaciones de las celdas en el modelo, la suma total de cuadrados se obtiene como:

$$SCT_{\text{SID}} = y_1'y_1 - \frac{1}{n_1}y_1'1_1 1_1'y_1.$$

donde el subíndice SID significa "sin imputar datos".

Comparando la expresión anterior con (39), se observa que la suma total de cuadrados imputando datos es mayor o igual que la suma total de cuadrados sin imputar.

Por otra parte, la suma de cuadrados del modelo (8) es $SCM = Y'V(V'V)^{-1}V'Y$, luego la suma de cuadrados del modelo al imputar los datos es:

$$\begin{aligned} SCM_{\text{CDI}} &= \begin{pmatrix} y_1' & \widehat{y}_2' \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} \left(\begin{pmatrix} V_1' & V_2' \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} V_1' & V_2' \end{pmatrix} \begin{pmatrix} y_1 \\ \widehat{y}_2 \end{pmatrix} \\ &= (y_1'V_1 + \widehat{y}_2'V_2)(V_1'V_1 + V_2'V_2)^{-1}(V_1'y_1 + V_2'\widehat{y}_2) \\ &= (y_1'V_1 + y_1'V_1(V_1'V_1)^{-1}V_2'V_2)(V_1'V_1 + V_2'V_2)^{-1}(V_1'y_1 + V_2'V_2(V_1'V_1)^{-1}V_1'y_1) \\ &= y_1'V_1(V_1'V_1)^{-1}V_1'y_1 + y_1'V_1(V_1'V_1)^{-1}(V_2'V_2)(V_1'V_1)^{-1}V_1'y_1 \\ &= y_1'V_1(V_1'V_1)^{-1}V_1'y_1 + \widehat{y}_2'\widehat{y}_2. \end{aligned} \tag{40}$$

La suma de cuadrados corregida por la media, imputando los datos, es:

$$SCM_{\text{CDICM}} = y_1'V_1(V_1'V_1)^{-1}V_1'y_1 + \widehat{y}_2'\widehat{y}_2 - \frac{1}{n}(y_1'1_1 1_1'y_1 + 2\widehat{y}_2'1_2 1_1'y_1 + \widehat{y}_2'1_2 1_2'\widehat{y}_2),$$

y la suma de cuadrados del modelo sin imputar los datos es:

$$SCM_{\text{SID}} = y_1'V_1(V_1'V_1)^{-1}V_1'y_1.$$

En forma semejante, la suma de cuadrados del modelo, corregida por la media, sin imputar los datos, es:

$$SCM_{\text{SDICM}} = y_1' V_1 (V_1' V_1)^{-1} V_1' y_1 - \frac{1}{n}(y_1' 1_1 1_1' y_1).$$

De (40) se observa que la suma de cuadrados del modelo al imputar los datos, es mayor o igual que la suma de cuadrados del modelo sin imputar.

La suma de cuadrados del error, sin imputación de datos es:

$$\begin{aligned} SCE &= Y'(I - P_V)Y \\ &= Y'Y - Y'V(V'V)^{-1}V'Y \end{aligned}$$

y con imputación de datos es:

$$\begin{aligned} SCE_{\text{CDI}} &= \begin{pmatrix} y_1' & \widehat{y}_2' \end{pmatrix} \begin{pmatrix} y_1 \\ \widehat{y}_2 \end{pmatrix} - y_1' V_1 (V_1' V_1)^{-1} V_1' y_1 - \widehat{y}_2' \widehat{y}_2, \\ &= y_1' y_1 + \widehat{y}_2' \widehat{y}_2 - y_1' V_1 (V_1' V_1)^{-1} V_1' y_1 - \widehat{y}_2' \widehat{y}_2 \\ &= y_1' (I_1 - V_1 (V_1' V_1)^{-1} V_1') y_1. \end{aligned} \tag{41}$$

Así, la suma de cuadrados del error con imputación de datos es igual a la suma de cuadrados del error sin imputación.

La Tabla 1 resume el análisis de varianza con los datos imputados.

Tabla 1: Análisis de varianza para el modelo de medias de celdas, corregido por la media, imputando datos.

Causas de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	F
Modelo	$r-1$	SCM	$CMM = \dfrac{SCM}{r-1}$	$F(M) = \dfrac{CMM}{CME}$
Residual	$n-r-m$	SCE	$CME = \dfrac{SCE}{n-r-m}$	
Total	$n-m-1$	SCT		

Nota: $r = p - s$.

5. Ejemplo de aplicación

En esta sección se presenta, como ejemplo, un diseño a tres vías de clasificación con efectos fijos, sin interacción, para ilustrar el desarrollo de las metodologías propuestas en la sección 4.

Tabla 2: Arreglo de una estructura factorial $2 \times 3 \times 4$

Tasa de desgaste	Material 1 Profundidad de corte				Material 2 Profundidad de corte			
	0.15	0.20	0.30	0.40	0.15	0.20	0.30	0.40
0.20	74	79	89	102	63	73	77	101
	78	82	**94**	98	68	**74**	79	103
0.25	98	97	98	105	74	85	83	105
	91	**93**	105	102	**77**	81	87	104
0.30	114	115	122	133	100	105	111	118
	108	111	**117**	138	97	108	**107**	122

El ejemplo es tomado de Myers & Montgomery (1995). En el experimento se consideraron tres factores que influyen sobre la superficie terminal de una partícula metálica: *tasa de desgaste* (en pulgadas por minuto), *profundidad de corte* (en pulgadas) y *tipo de material*. Se tomaron dos observaciones para cada una de las 24 combinaciones, obteniendo el conjunto de datos que se muestra en la Tabla 2. Suponiendo una pérdida aleatoria de información, se procedió a retirar la información que aparece en negrilla en esta tabla.

5.1. Ilustración del método de covariable sujeto a restricción

Con la información de la Tabla 2 se ilustra el método presentado en la sección 4, a partir de la cual se llega a la imputación del conjunto de datos que aparecen en negrilla.

Para obtener estas estimaciones, se trabaja con el modelo de covariable bajo restricción, presentado en (29). Se construye la matriz de incidencia W de orden

48×24, que luego se particiona en dos submatrices W_1 de orden 42×24 y W_2 de orden 6×24, que identifican, por celda, el número de datos observados y el número de datos faltantes, respectivamente. Luego, se particiona como sigue:

$$\begin{pmatrix} W_1 \\ W_2 \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}.$$

Estas cuatro submatrices corresponden a la partición de la matriz de restricciones, G. Las matrices W_{21} y W_{22} son de orden 6×7 y 6×17 respectivamente. Con las submatrices W_{11} y W_{12} se obtiene la matriz $V_1 = W_{11} - W_{12}G_2^{-1}G_1$ de orden 42×7; similarmente se obtiene $V_2 = W_{21} - W_{22}G_2^{-1}G_1$ de orden 6×7:

$$V_2 = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 1 & -2 \\ 1 & 0 & 1 & 0 & 1 & 0 & -2 \\ 1 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 1 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Con V_1 y V_2 se construye la matriz $V = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix}$, de orden 48×7, a partir de la cual se obtiene el proyector ortogonal $P_V = V(V'V)^{-1}V'$ de orden 48×48; además,

$$V'V = \begin{pmatrix} 24 & 8 & 8 & 6 & 6 & 6 & -34 \\ 8 & 16 & 0 & 4 & 4 & 4 & -20 \\ 8 & 0 & 16 & 4 & 4 & 4 & -20 \\ 6 & 4 & 4 & 12 & 0 & 0 & -14 \\ 6 & 4 & 4 & 0 & 12 & 0 & -14 \\ 6 & 4 & 4 & 0 & 0 & 12 & -14 \\ -34 & -20 & -20 & -14 & -14 & -14 & 72 \end{pmatrix};$$

como esta matriz es no singular, el diseño es conectado.

Con el proyector ortogonal se obtiene la matriz $I - P_V = \begin{pmatrix} S & M \\ M' & L \end{pmatrix}$, con S, M y L, matrices de orden 42×42, 42×6 y 6×6 respectivamente.

Especificamente,

$$L = \begin{pmatrix} 17/20 & 0 & -1/12 & -1/48 & 1/24 & -1/24 \\ 0 & 17/20 & 0 & -1/24 & -1/48 & 1/24 \\ -1/12 & 0 & 17/20 & 1/24 & 1/24 & -1/10 \\ -1/48 & -1/24 & 1/24 & 17/20 & 0 & 0 \\ 1/24 & -1/48 & 1/24 & 0 & 17/20 & 0 \\ -1/24 & 1/24 & -1/10 & 0 & 0 & 17/20 \end{pmatrix}.$$

El vector de datos observados es: $y_1' = \big(74,\ 78,\ 79,\ 82,\ 89,\ 102,\ 98,\ \cdots, 122\big)$, de tamaño 1×42.

Utilizando (37), se construye el vector de estimaciones de los datos faltantes en las celdas de interés: $\widehat{\gamma}' = (88{,}6;\ 92{,}8;\ 120{,}1;\ 73{,}1;\ 78{,}1;\ 110{,}35)$. El valor real para los datos de estas celdas está dado por los resultados en negrilla de la Tabla 2. Como puede verse, las estimaciones obtenidas por el método propuesto son cercanas a los valores reales, aunque algunos datos son subestimados y otros sobreestimados.

5.2. Ilustración del método de mínimos cuadrados con restricción

La matriz de incidencia, W, coincide con la que se obtuvo para el modelo de covariables con restricción y, como en 5.1, se particiona en las cuatro submatrices ya mencionadas.

Se construyen dos modelos a partir de las matrices V_1 y V_2 obtenidas en (18) y (20). Se estima inicialmente el vector de medias de celdas μ_1 encontrado en (19); para esto se calcula

$$V_1'V_1 = \begin{pmatrix} 21 & 7 & 7 & 6 & 5 & 4 & -29 \\ 7 & 14 & 0 & 4 & 3 & 3 & -17 \\ 7 & 0 & 14 & 3 & 3 & 4 & -17 \\ 6 & 4 & 3 & 11 & 0 & 0 & -13 \\ 5 & 3 & 3 & 0 & 10 & 0 & -11 \\ 4 & 3 & 4 & 0 & 0 & 9 & -11 \\ -29 & -17 & -17 & -13 & -11 & -11 & 61 \end{pmatrix}.$$

Con esta matriz y aplicando (19), se encuentra que

$$\widehat{\mu}_1' = \begin{pmatrix} 134 & 92 & 102 & 100 & 105 & 110 & 124 \end{pmatrix}.$$

Por (21), la estimación de la información faltante del vector $\widehat{y}_2$ es:

$$\widehat{y}_2 = V_2\widehat{\mu}_1; \widehat{y}_2' = \begin{pmatrix} 88{,}6 & 92{,}8 & 120{,}1 & 73{,}1 & 78{,}1 & 110{,}35 \end{pmatrix}.$$

Este resultado coincide, como se demostró teóricamente, con el obtenido mediante el método de estimación presentado en la sección 5.1.

6. Conclusiones

Se presentaron dos nuevos métodos para imputar información faltante a partir del modelo de medias de celdas, sujetos a la restricción de no interacción entre los efectos. Además, se mostró que los dos métodos son equivalentes, pues los dos dan como resultado el mismo estimador.

Los métodos propuestos, además de estimar las celdas vacías, permiten estimar los datos individuales en las celdas, asi como corregir la diferentes sumas de cuadrados del análisis de varianza.

Como otro resultado importante, se detemina el número de celdas vacías que deben ser estimadas para conectar el diseño, y una vez determinado, se sabe cuáles son las celdas que deben ser estimadas. Esto resuelve el problema de la *desconectés* en el diseño y por lo tanto, todos los contrastes de efectos principales resultan estimables, lo cual facilita la prueba de las hipótesis linealmente independientes que se generen.

Agradecimientos

Este trabajo tuvo el apoyo económico de la División de Investigación de la Universidad Nacional de Colombia, Sede Bogotá, a través del proyecto *Metodología estadística en análisis de datos longitudinales y medidas repetidas.*

Bibliografía

Birkes, D., Dodge, Y. & Seely, J. (1976), 'Spanning sets for estimable contrasts in classification models', *Ann. Statist* **4**, 86–107.

Bose, R. C. & Srivastava, J.Ñ. (1964), 'Mathematical theory of factorial designs: I analysis, II construction', *Bull Int Stat Inst* pp. 780–794.

Dodge, Y. (1985), *Analysis of Experiments with Missing Data*, John Wiley & Sons.

Graybill, F. A. (1961), *An Introduction to Linear Statistical Models*, Vol. I, McGraw-Hill.

Graybill, F. A. (1976), *Theory and Application of the Linear Model*, North Scituate.

Hocking, R. R. (1985), *The Analysis of Linear Models*, Brooks/Cole.

Hocking, R. R. (1996), *Methods and Application of Linear Models*, John Wiley & Sons.

López, L. A. (1999), 'Los modelos de medias de celdas, una herramienta fundamental en la estadística industrial', *Simposio de Estadística* .

Melo, O. O. & Lozano, A. R. (1998), 'Funciones estimables en modelos de clasificación con datos desbalanceados a través del algoritmo de Cholesky'. Trabajo de grado.

Melo, O. O., Lozano, A. R. & López, L. A. (1999), 'Funciones estimables en modelos de clasificación con datos desbalanceados a través del algoritmo de Cholesky', *Revista Multiciencia* **3**(2), 131–147.

Melo, S. E. (2000), 'Comparación de métodos de conectés en modelos a n-vías de clasificación sin interacción'. Trabajo de grado.

Murray, L. W. (1986), 'Estimation of missing cells in randomized block and latin square designs', *The American Statistical Association* **40**(4), 289–293.

Murray, L. W. & Smith, D. W. (1985), 'Estimability, testability and connectedness in the cell mean model', *Communications in Statistic, Part A-Theory and Methods* **14**, 1889–1915.

Myers, R. H. & Montgomery, D. C. (1995), *Response Surface Methodology :Process and Product Optimization Using Designed Experiments*, John Wiley & Sons.

Rao, C. R. (1945), 'On the linear combination of observations and the general theory of least squares', *Sankhyâ* **7**, 237–256.

Rao, C. R. & Mitra, S. R. (1971), *Generalized Inverse of Matrices and its Applications*, John Wiley & Sons.

Searle, S. (1987), *Linear Models for Unbalanced Data*, John Wiley & Sons.

Searle, S. R. (1971), *Linear Models*, John Wiley & Sons.

Speed, F. M., Hocking, R. R. & Hackney, O. P. (1978), 'Methods of analysis of linear models with unbalanced data', *Journal of the American Statistical Association* **73**, 105–112.

Weeks, D. L. & Williams, D. R. (1964), 'A note on the determination of connectedness in a n-way cross classification', *Technometrics* **6**(3), 319–324.