Revista Colombiana de Estadística
Volumen 29 N° 1. pp. 35 a 56. Junio 2006

Estimación de información faltante, imputación y estadísticos de prueba en modelos mixtos a dos vías de clasificación

Estimation of Missing Data, Imputation and Test Statistics in Two-Way Classification Mixed Models

Diana Carolina Franco*, Oscar Orlando Melo†

Universidad Nacional de Colombia, Departamento de Estadística, Bogotá

Resumen

Se propone una metodología para la estimación de información faltante en modelos mixtos de medias de celda que permite la disminución de la correlación entre la información observada y la información estimada, basada en el método propuesto por Melo & Melo (2005). Éste se fundamenta en los métodos de estimación vía máxima verosimilitud, expuesto en Searle (1971), y de covariable, propuesto por Bartlett (1937). Después de realizar la imputación de la información, se plantea una manera de llevar a cabo el análisis de varianza en modelos sin interacción, mediante pruebas ponderadas para los efectos fijos y aleatorios involucrados en el modelo.

Palabras clave: Modelo de medias de celda, modelo mixto, información faltante, estimación e imputación, distribución de formas cuadráticas.

Abstract

We propose a methodology to estimate missing information in mixed cell means models. This methodology improves on that Melo & Melo (2005), which is based on the methods of maximum likelihood estimation and covariate proposed by Bartlett (1937), and reduces the correlation between the observed and estimated information. Once the imputation of the missing information is done, we suggest a way to perform the analysis of variance in models without interaction, by generating a weighted test for the fixed and random effects involved in the model.

Key words: Cell means model, Mixed model, Missing information, Estimation and imputation, Distribution of quadratic forms.

*Profesora temporal. E-mail: dcfrancos@unal.edu.co
†Profesor asistente. E-mail: oomelom@unal.edu.co

1. Introducción

En el contexto del diseño de experimentos, cuando se realizan las pruebas de campo, es posible que se pierdan unidades observacionales o experimentales. Esto ocurre a pesar de vigilar cuidadosamente el proceso, por razones fuera de control, ajenas a la voluntad del investigador. La ausencia de dicha información, en algunos casos puede poner en riesgo la validez del proceso.

En los diseños multifactoriales y, en especial, en el modelo de medias de celda, cuando hay celdas vacías no todas las medias son estimables y ciertas hipótesis sobre combinaciones lineales de los parámetros pierden sentido. Dicha falta de información destruye la ortogonalidad del diseño y se requiere identificar las funciones estimables, cuya naturaleza depende de si el modelo es conectado o no. En cuanto a este último aspecto se destacan los trabajos realizados por Murray & Smith (1985) y Dodge (1985) en modelos a n vías de clasificación.

El estudio de la información faltante se remonta al año de 1930, en el cual Allan & Wishart (1930) presentan dos métodos para estimar el valor de una sola observación faltante para un diseño en bloques al azar y un cuadrado latino. Luego, Yates (1933) propone imputar los valores faltantes y así llevar a cabo el análisis de varianza común, con las correcciones correspondientes en los grados de libertad del error. Son muchos los trabajos realizados al respecto desde esa época hasta la actualidad, los cuales en su mayoría son referenciados en una de las últimas y más importantes publicaciones de los principales aportes al análisis estadístico con información faltante, realizada por Little & Rubin (2002). Sin embargo, la mayoría de los estudios sobre información faltante realizados hasta la época, hacen referencia esencialmente al manejo de la ausencia de información cuando se está trabajando con modelos de efectos fijos, de tal manera que para modelos de efectos aleatorios o mixtos los desarrollos referentes al manejo de la ausencia de información no son muy amplios.

Haciendo una revisión de la literatura dedicada a los modelos de componentes de varianza y mixtos, se destacan los trabajos de Fisher (1935), Henderson (1952), Sheffé (1959) y Searle (1971), discutidos en un trabajo completamente dedicado a los modelos de componentes de varianza, realizado por Searle et al. (1992). Adicionalmente, cuando se dispone de la información completa son importantes los desarrollos presentados por Corbeil & Searle (1976) para la estimación de los componentes de varianza vía máxima verosimilitud restringida, y Lindstrom & Bates (1988), quienes implementan el método de Newton Raphson y el algoritmo *EM*. Por otra parte, en diseños con datos desbalanceados o que se encuentran afectados por el problema de la ausencia de información, se encuentran los trabajos de Thomsen (1975), Gallo & Khuri (1990), Harville & Carriquiry (1992), Barroso et al. (1998) y uno de los más recientes, realizado por Melo & Melo (2005), en el cual se propone una metodología para la estimación de la información faltante en modelos mixtos de medias de celda.

Los desarrollos presentados en la mayoría de estos trabajos muestran la ausencia de un estudio detallado relacionado con la construcción de los estadísticos de prueba para verificar hipótesis relativas a los efectos fijos y aleatorios involucrados

en los modelos mixtos cuando hay información faltante, al igual que la falta de una revisión exhaustiva de la distribución de las pruebas obtenidas.

Con el fin de suplir lo anterior, en el presente trabajo se propone una metodología para manejar el problema de la información faltante en modelos mixtos de medias de celda. Para ello, en la sección inmediatamente siguiente se presentan algunos conceptos básicos sobre modelos mixtos y una breve descripción del método de Melo & Melo (2005). En la tercera sección se expone la metodología propuesta para la estimación de información faltante en modelos mixtos de medias de celda y posteriormente se realiza un análisis de la distribución de los estadísticos de prueba construidos. En la última sección se presenta una aplicación de la metodología propuesta.

2. Conceptos preliminares

En esta parte se exponen algunos conceptos básicos, necesarios para el entendimiento adecuado de los desarrollos proporcionados en el presente artículo.

El modelo mixto de medias de celda presentado en Corbeil & Searle (1976) está dado por:

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

donde:

- $y_{(n\times 1)}$ es el vector de variables aleatorias,
- $W_{(n\times p)}$ es la matriz de incidencia de rango columna completo que asocia las observaciones con los respectivos componentes del vector $\mu_{(p\times 1)}$, el cual es un vector de valores medios desconocidos de los efectos fijos,
- $Z_{(n\times r)}$ es la matriz de incidencia que asocia las observaciones con los respectivos componentes del vector $\theta_{(r\times 1)}$, el cual es de constantes desconocidas de los efectos aleatorios, y $e_{(n\times 1)}$ es un vector de variables aleatorias no observables tal que $e \sim N(0_{(n\times 1)}, \sigma_e^2 I_n)$. En este caso σ_e^2 y las diferentes varianzas de los elementos de θ son los componentes de varianza del modelo.

2.1. Estimación de información faltante en modelos mixtos de medias de celda

Siguiendo los resultados presentados en Searle (1971), Barroso et al. (1998) y Melo & Melo (2005) exponen una metodología para la estimación de información faltante en modelos mixtos de medias de celda, vía el método de máxima verosimilitud y el método de covariable propuesto por Bartlett (1937). Dicha metodología es base fundamental del presente trabajo y por esta razón se describe brevemente a continuación.

Asumiendo que el modelo (1) tiene una estructura de variabilidad de la forma:

$$\begin{aligned} \theta_{(r\times 1)} &\sim N(0_{(r\times 1)}, D_{(r\times r)}), \\ e_{(n\times 1)} &\sim N(0_{(n\times 1)}, R_{(n\times n)}) \quad \text{y} \\ Cov(\theta, e) &= 0_{(r\times n)} \end{aligned} \tag{2}$$

donde D y R son las matrices de varianzas y covarianzas de rango completo asociadas a los factores aleatorios involucrados en el modelo.

Por lo tanto, la matriz de varianzas y covarianzas asociada al vector y está dada por:

$$Var(y) = V_{(n\times n)} = ZDZ' + R \tag{3}$$

Si el modelo mixto posee un único factor aleatorio, la estructura de variabilidad asumida para el modelo (1) es:

$$\begin{aligned} \theta_{(r\times 1)} &\sim N(0_{(r\times 1)}, \sigma^2_\theta I_r), \\ e_{(n\times 1)} &\sim N(0_{(n\times 1)}, \sigma^2_e I_n) \quad \text{y} \\ Cov(\theta, e) &= 0_{(r\times n)} \end{aligned} \tag{4}$$

de donde:

$$Var_p(y) = V_{p(n\times n)} = \sigma^2_\theta ZZ' + \sigma^2_e I_n \tag{5}$$

Por otra parte, la función de densidad conjunta de y y θ está dada por:

$$f(y,\theta) = \frac{1}{(2\pi)^{\frac{n+r}{2}}(|R||D|)^{\frac{1}{2}}} \exp\big[-\tfrac{1}{2}[(y - W\mu - Z\theta)'R^{-1}(y - W\mu - Z\theta) + \theta' D^{-1}\theta]\big] \tag{6}$$

A fin de encontrar los máximos de la función anterior, se calculan las derivadas parciales de $L = \ln f(y,\theta)$, de lo cual se obtiene el siguiente sistema de ecuaciones normales:

$$\begin{bmatrix} W'R^{-1}W & W'R^{-1}Z \\ Z'R^{-1}W & Z'R^{-1}Z + D^{-1} \end{bmatrix} \begin{bmatrix} \mu \\ \theta \end{bmatrix} = \begin{bmatrix} W'R^{-1}y \\ Z'R^{-1}y \end{bmatrix} \tag{7}$$

Resolviendo el sistema anterior se obtienen los siguientes estimadores de máxima verosimilitud para μ y θ:

$$\dot{\mu} = (W'V^{-1}W)^{-1}W'V^{-1}y \tag{8}$$

$$\dot{\theta} = DZ'V^{-1}(I_n - P_W)y \tag{9}$$

donde $P_{W(n\times n)} = W(W'V^{-1}W)^{-1}W'V^{-1}$.

2.1.1. Método de covariable para modelos mixtos de medias de celda

Si hay $m > 1$ observaciones faltantes en un diseño experimental, se puede incluir una covariable por cada observación faltante, teniendo un modelo de covariables a partir de la matriz diseño. En este caso, el modelo mixto de medias de

celda se puede escribir como:

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} \mu + \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \theta + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \tag{10}$$

donde el vector de variables aleatorias $y_{(n\times 1)}$ es particionado en dos vectores, uno con la información de los datos observados $y_{1((n-m)\times 1)}$ y otro que representa los datos no observados $y_{2(m\times 1)}$, el cual puede ser un vector de ceros o cualquier vector de valores iniciales supuestos. Además, $W_{(n\times p)}$ es particionada en dos submatrices cuyas filas están asociadas a la información observada $W_{1((n-m)\times p)}$ y a la información faltante $W_{2(m\times p)}$, y, $Z_{(n\times r)}$ es particionada en dos submatrices cuyas filas están asociadas a la información observada $Z_{1((n-m)\times r)}$ y a la información faltante $Z_{2(m\times r)}$.

Debido a que se consideran m valores faltantes, tomando $y_2 = 0_{(m\times 1)}$ en (10) e introduciendo una matriz de indicadores $U_{(n\times m)}$ en la cual los elementos de cada una de las m columnas toman el valor de cero si el dato es observado y el valor de -1 en caso contrario, y, $\gamma_{(m\times 1)}$ es el vector asociado con los valores faltantes en el diseño, el modelo con covariables se puede escribir como:

$$y^* = \begin{bmatrix} y_1 \\ 0 \end{bmatrix} = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} \mu + \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \theta + \begin{bmatrix} 0 \\ -I_m \end{bmatrix} \gamma + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

En general, el modelo mixto de medias de celda con covariables queda de la forma:

$$y^* = W\mu + U\gamma + Z\theta + e \tag{11}$$

donde U es la matriz de indicadores de los valores faltantes en el diseño.

Se asume que el modelo (11) tiene una estructura de variabilidad como la dada en (2), por lo tanto, la estructura conocida de la matriz de varianzas y covarianzas asociada al vector de variables aleatorias relacionadas con los datos observados y_1 está dada por:

$$\mathrm{Var}(y_1) = [V_1]_{((n-m)\times(n-m))} = Z_1 D Z_1' + R_1 \tag{12}$$

Al desarrollar las derivadas parciales de la función de densidad conjunta de y^* y θ, $L^* = \ln f(y^*, \theta)$, se encuentran los estimadores de los parámetros del modelo (11), los cuales están dados por Melo & Melo (2005):

$$\tilde{\mu} = (W_1' V_1^{-1} W_1)^{-1} W_1' V_1^{-1} y_1 \tag{13}$$

$$\tilde{\theta} = D Z_1' V_1^{-1} (I_{(n-m)} - P_{W_1}) y_1 \tag{14}$$

donde $P_{W_1((n-m)\times(n-m))} = W_1 (W_1' V_1^{-1} W_1)^{-1} W_1' V_1^{-1}$, y el estimador de los componentes del vector asociado a la información faltante está dado por:

$$\hat{\gamma} = \hat{y}_2 = \left[W_2 (W_1' V_1^{-1} W_1)^{-1} W_1' V_1^{-1} + Z_2 D Z_1' V_1^{-1} (I_{(n-m)} - P_{W_1}) \right] y_1 \tag{15}$$

Se puede comprobar que $\hat{\gamma} \sim N[\gamma, \mathrm{Var}(\hat{\gamma})]$ con:

$$\mathrm{Var}(\hat{\gamma}) = \left[W_2 (W_1' V_1^{-1} W_1)^{-1} W_2' + Z_2 D Z_1' V_1^{-1} (I - P_{W1}) Z_1 D Z_2' \right]_{(m\times m)}$$

3. Estimación e imputación de la información faltante con el modelo mixto de medias de celda

El estimador de las observaciones faltantes $\hat{y}_2$ propuesto por Melo & Melo (2005) es una combinación lineal del vector de información observada y_1 y, por lo tanto, la covarianza entre estos dos vectores es diferente de cero:

$$\mathrm{Cov}(y_1, \hat{y}_2) = \left[W_1(W_1'V_1^{-1}W_1)^{-1}W_2' + (I - P_{W1})Z_1DZ_2'\right]_{((n-m)\times m)}$$

Este hecho hace necesaria la existencia de una metodología mediante la cual se disminuya la dependencia entre el vector de la información estimada y el de la información observada. Para ello se propone adicionar una componente aleatoria $e_{2(m\times 1)}$ al vector de observaciones estimadas $\hat{y}_2$, tal que:

$$e_2 \sim N(0_{(m\times 1)}, \sigma^2_{e(1)}I_m)$$

donde $\sigma^2_{e(1)}$ es la varianza de los errores obtenidos al ajustar un modelo mixto con la información observada.

Dicho vector de errores $e_{2(m\times 1)}$ es generado aleatoriamente y adicionalmente debe satisfacer el supuesto de incorrelación con el vector de factores aleatorios $\theta_{(r\times 1)}$ y el vector de errores $e_{1((n-m)\times 1)}$, de tal forma que: $\mathrm{Cov}(\theta, e_2) = 0_{(r\times m)}$ y $\mathrm{Cov}(e_1, e_2) = 0_{((n-m)\times m)}$.

Así, la expresión propuesta para la estimación de la información faltante está dada por:

$$\hat{\gamma}^* = \hat{y}_2^* = \left[W_2(W_1'V_1^{-1}W_1)^{-1}W_1'V_1^{-1} + Z_2DZ_1'V_1^{-1}(I - P_{W1})\right]y_1 + e_2 \qquad (16)$$

Al adicionar dicha componente aleatoria, el vector de las observaciones estimadas ya no es una combinación lineal de la información observada, y así se disminuye la correlación entre los subvectores del vector de información y. Lo anterior se puede ver de forma más clara al hallar las respectivas matrices de correlaciones asociadas con las matrices de varianzas y covarianzas entre la información observada y la estimada (Franco 2005).

Se puede comprobar que $E(\hat{\gamma}^*) = \gamma$; además:

$$\mathrm{Var}(\hat{y}_2^*) = \left[W_2(W_1'V_1^{-1}W_1)^{-1}W_2' + Z_2DZ_1'V_1^{-1}(I - P_{W1})Z_1DZ_2' + \sigma^2_eI_m\right]$$

y la covarianza entre el vector de información observada y estimada está dada por:

$$\mathrm{Cov}(y_1, \hat{y}_2^*) = \left[W_1(W_1'V_1^{-1}W_1)^{-1}W_2' + (I - P_{W1})Z_1DZ_2'\right]$$

3.1. Estimación de parámetros y análisis de varianza en modelos mixtos de medias de celda

Para estimar los parámetros involucrados en el modelo y construir los estadísticos de prueba empleados en el análisis de varianza, se emplea el método de máxima

verosimilitud para el modelo mixto de medias de celda especificado en (1), con una estructura de variabilidad como la expuesta en (4).

Teniendo en cuenta la idea anterior, se exponen a continuación los principales desarrollos obtenidos para los modelos mixtos con un único factor aleatorio.

3.1.1. Estimación de parámetros en modelos mixtos con un único factor aleatorio

El modelo mixto de medias de celda presentado en (1), tiene una función de densidad conjunta de y y θ, como la expuesta en (6). En el caso de modelos mixtos con un único factor aleatorio sin interacción, este modelo tiene una estructura de variabilidad como la expuesta en (4), de tal forma que la estructura de variabilidad para y corresponde a la presentada en (5) y la función de densidad conjunta de y y θ es:

$$f(y,\theta) = \frac{1}{(2\pi)^{\frac{n+r}{2}}(\sigma_e^{2n}\sigma_\theta^{2r})^{\frac{1}{2}}} \exp\Big[-\frac{1}{2\sigma_e^2}\big\{(y-W\mu-Z\theta)'(y-W\mu-Z\theta)+\lambda\theta'\theta\big\}\Big] \quad (17)$$

donde $\lambda = \dfrac{\sigma_e^2}{\sigma_\theta^2}$.

Las ecuaciones normales presentadas en (7) se transforman en este caso en:

$$\begin{bmatrix} W'W & W'Z \\ Z'W & Z'Z+\lambda I_r \end{bmatrix}\begin{bmatrix}\mu\\ \theta\end{bmatrix} = \begin{bmatrix} W'y \\ Z'y\end{bmatrix} \quad (18)$$

Resolviendo el sistema anterior y empleando las propiedades de la inversa de la suma entre matrices, se obtienen los siguientes estimadores máximo verosímiles para μ y θ respectivamente:

$$\hat{\mu} = (W'V_p^{-1}W)^{-1}W'V_p^{-1}y \quad (19)$$

$$\hat{\theta} = D_pZ'V_p^{-1}[I_n - P_{Wp}]y \quad (20)$$

con $P_{Wp\,(n\times n)} = W(W'V_p^{-1}W)^{-1}W'V_p^{-1} = W(W'TW)^{-1}W'T$ y

$V_p^{-1} = \dfrac{I_n - ZP^{-1}Z'}{\sigma_e^2} = \dfrac{T}{\sigma_e^2}$, donde $P = Z'Z+\lambda I_r$ y $T = I_n - ZP^{-1}Z'$.

Sin embargo, los parámetros μ y θ no son los únicos a estimar en el modelo de interés. De hecho, para poder emplear los estimadores obtenidos previamente, es necesario conocer V_p o, en su defecto, conocer una estimación de los componentes de varianza que la generan.

Con el fin de hallar los estimadores de σ_e^2 y σ_θ^2, se debe tener en cuenta que W y Z son matrices de rango columna completo, entonces $ran\big[W\ Z\big] = ran(W) + ran(Z) - 1 = p + r - 1$.

Si se considera el *Método de Ajuste de Constantes*, también conocido como Método III de Henderson (1952), se tiene la siguiente descomposición de las sumas de cuadrados involucradas en la estimación de los componentes de varianza:

$$\begin{aligned} R(\mu^0,\theta^0) &= y'\begin{bmatrix} W & Z\end{bmatrix}\begin{bmatrix} W'W & W'Z \\ Z'W & Z'Z\end{bmatrix}^{-1}\begin{bmatrix} W' \\ Z'\end{bmatrix} y \\ R(\mu^0) &= y'W(W'W)^{-1}W'y \\ R(\theta^0) &= y'Z(Z'Z)^{-1}Z'y \\ R(\mu^0/\theta^0) &= \hat{\mu}^{0'}W'[I_n - Z(Z'Z)^{-1}Z']y \\ R(\theta^0/\mu^0) &= R(\mu^0,\theta^0) - R(\mu^0) \end{aligned}$$

donde $\hat{\mu}^0 = Q_0^{-1}W'[I_n - Z(Z'Z)^{-1}Z']y$, con $Q_0 = W'[I_n - Z(Z'Z)^{-1}Z']W$.

Y como se muestra en Searle et al. (1992), los estimadores insesgados para los componentes de varianza están dados por:

$$\tilde{\sigma}^2_{e(0)} = \frac{y'y - R(\mu^0,\theta^0)}{n - ran[W\ Z]} \tag{21}$$

y

$$\tilde{\sigma}^2_{\theta(0)} = \frac{R(\theta^0/\mu^0) - \tilde{\sigma}^2_{e(0)}(ran[W\ \ Z] - ran(W))}{tr[Z'Z - Z'W(W'W)^{-1}W'Z]} \tag{22}$$

donde tr es el operador traza de una matriz.

Las sumas de cuadrados citadas previamente son las empleadas tradicionalmente e implementadas en diferentes paquetes estadísticos, como en el procedimiento Proc Mixed de SAS. Sin embargo, el error de emplearlas radica en que hacen referencia en realidad a un modelo de la forma:

$$y = W\mu^0 + Z\theta^0 + e \tag{23}$$

donde μ^0 y θ^0 son vectores de parámetros asociados a efectos fijos; de esta forma sus ecuaciones normales son respectivamente:

$$\begin{bmatrix} W'W & W'Z \\ Z'W & Z'Z\end{bmatrix}\begin{bmatrix}\mu^0 \\ \theta^0\end{bmatrix} = \begin{bmatrix} W'y \\ Z'y\end{bmatrix} \tag{24}$$

Por lo anterior, los estimadores dados en (21) y (22) no son apropiados en modelos donde θ^0 es un vector de efectos aleatorios.

Sin embargo, nótese la gran similitud entre las ecuaciones normales (24) y las dadas en (18) para los modelos mixtos de interés con un único factor aleatorio. Como puede verse, estos sistemas de ecuaciones son iguales excepto porque se tiene $P = Z'Z + \lambda I_r$ en lugar de $Z'Z$. Aprovechando esta similitud y realizando un procedimiento análogo al presentado en Searle (1971), se generan unas sumas de cuadrados similares a las asociadas al sistema (7) reemplazando $Z'Z$ por P y con ellas se obtienen unos estimadores similares a los presentados en (21) y (22).

De esta forma, los estimadores para los componentes de varianza están dados por:

$$\tilde{\sigma}_e^2 = \frac{y'y - R(\mu,\theta)}{n - ran[W\ Z]} \tag{25}$$

y

$$\tilde{\sigma}_\theta^2 = \frac{R(\theta/\mu) - \tilde{\sigma}_e^2(ran[WZ] - ran(W))}{tr[P - Z'W(W'W)^{-1}W'Z]} \tag{26}$$

donde $R(\mu,\theta) = y'[I_n - ZP^{-1}Z']W\hat{\mu} + y'ZP^{-1}Z'y$, $R(\theta/\mu) = R(\mu,\theta) - R(\mu)$ con $R(\mu) = y'W(W'W)^{-1}W'y$. Además $\hat{\mu} = Q^{-1}W'[I_n - ZP^{-1}Z']y$, con $Q = W'[I_n - ZP^{-1}Z']W = W'TW$.

Sin embargo, aunque estos estimadores tienen sentido, dejan de ser insesgados cuando se sustituye $Z'Z$ por P.

Uno de los caminos para volver estos estimadores insesgados consiste en hallar las esperanzas de las sumas de cuadrados involucradas y luego tomar como estimador del respectivo componente de varianza el cuadrado medio cuyo valor esperado es igual al componente de interés. Basados en dichas esperanzas, los estimadores insesgados de los componentes de varianza para el caso de interés (Franco 2005) son:

$$\hat{\sigma}_e^2 = \frac{(y - W\hat{\mu} - Z\hat{\theta})'(y - W\hat{\mu} - Z\hat{\theta}) + \lambda\hat{\theta}'\hat{\theta}}{n - ran(W)} \tag{27}$$

$$\hat{\sigma}_\theta^2 = \frac{y'[I_n - W(W'W)^{-1}W']y - (y - W\hat{\mu} - Z\hat{\theta})'(y - W\hat{\mu} - Z\hat{\theta}) - \lambda\hat{\theta}'\hat{\theta}}{tr[Z'Z - Z'W(W'W)^{-1}W'Z]} \tag{28}$$

3.1.2. Prueba de hipótesis general

Para el modelo mixto expuesto en (1), la hipótesis general es:

$$\begin{cases} H_0 : M\beta = 0; & \sigma_e^2 > 0, \quad \sigma_\theta^2 > 0 \\ H_a : M\beta \neq 0; & \sigma_e^2 > 0, \quad \sigma_\theta^2 > 0 \end{cases} \tag{29}$$

donde:

$$M_{(n\times(p+r))} = \begin{bmatrix} W_{(n\times p)} & Z_{(n\times r)} \end{bmatrix} \qquad \beta_{((p+r)\times 1)} = \begin{bmatrix} \mu_{(p\times 1)} \\ \theta_{(r\times 1)} \end{bmatrix}$$

De tal manera que el modelo (1) se escribe en términos de un modelo de la forma:

$$y = M\beta + e \tag{30}$$

Para desarrollar la prueba, se parte de la función de verosimilitud de y y β, y a partir de ésta se hace la prueba de hipótesis utilizando la razón de verosimilitud generalizada dada por:

$$\varphi^{\frac{2}{n}} = \frac{1}{\left(\frac{\tilde{\sigma}^2_{e(0)H_0}}{\tilde{\sigma}^2_{e(0)}}\right)} \tag{31}$$

donde $\tilde{\sigma}^2_{e(0)}$ está dado por (21) y $\tilde{\sigma}_{e(0)H_0}$ es este mismo asumiendo H_0 cierta.

Por otra parte, si H_0 es cierta en (29), se tiene que:

$$(n - ran[WZ])\tilde{\sigma}^2_{e(0)H_0} = y'y$$

y por consiguiente de (21):

$$(n - ran[W\ Z])\tilde{\sigma}^2_{e(0)} = (n - ran[W\ Z])\tilde{\sigma}^2_{e(0)H_0} - R(\mu^0, \theta^0)$$

Reemplazando los dos anteriores resultados en (31), se obtiene:

$$\varphi^{\frac{2}{n}} = \frac{1}{\left(1 + \frac{R(\mu^0, \theta^0)}{(n - ran[W\ Z])\tilde{\sigma}^2_{e(0)}}\right)} \tag{32}$$

Sin embargo, el estimador insesgado de σ^2_e dado en (21) está asociado a un modelo como el presentado en (23), en el cual θ^0 es considerado un vector de efectos fijos más que de efectos aleatorios. De acuerdo con todo lo discutido previamente respecto a este modelo y la relación que tiene con el modelo mixto de interés, para que la razón obtenida sea útil en el modelo en estudio, lo apropiado es hacer uso del estimador insesgado de σ^2_e dado en (27). En este caso,

$$(n - ran(W))\hat{\sigma}^2_e = (n - ran(W))\hat{\sigma}^2_{e(H_0)} - R(\mu, \theta)$$

y reemplazando de nuevo en (31), se llega a:

$$\varphi^{\frac{2}{n}} = \frac{1}{\left(1 + \frac{R(\mu, \theta)}{(n - ran(W))\hat{\sigma}^2_e}\right)} \tag{33}$$

Ésta es la razón adecuada para el modelo mixto de interés dado en (1), para la cual se cumple que:

$$\varphi^{\frac{2}{n}} \text{ es pequeña si } \frac{R(\mu, \theta)}{(n - ran(W))\hat{\sigma}^2_e} \text{ es grande}$$

y es una función monótona que permite utilizar a:

$$\frac{R(\mu, \theta)}{(n - ran(W))\hat{\sigma}^2_e}$$

para llevar a cabo la prueba de hipótesis general.

Al considerar la hipótesis general de la forma presentada en (29), se evalúa conjuntamente el aporte tanto de la parte fija como de la aleatoria al modelo, pero no se tiene en cuenta que en el contexto del modelo mixto la componente aleatoria no debe ser tratada de la misma manera que la componente fija. Esto se debe a que para la parte aleatoria no se tiene interés en obtener conclusiones aplicadas únicamente a los niveles de los factores considerados, sino en generalizarlas a la población de la cual los tratamientos se seleccionaron y, por esta razón, el interés se centra en probar hipótesis referentes a la variabilidad de los efectos de los tratamientos.

De lo anterior, la hipótesis general apropiada para el modelo mixto de interés es:

$$\begin{cases} H_0 : W\mu = 0 \quad \text{y} \quad \sigma_\theta^2 = 0; \qquad \sigma_e^2 > 0 \\ H_a : \text{Al menos uno de los dos, } W\mu \text{ o } \sigma_\theta^2, \text{ es diferente de cero;} \quad \sigma_e^2 > 0 \end{cases}$$

Al construir el estadístico de prueba asociado a esta hipótesis, se encuentra que el estadístico de prueba obtenido empleando estos resultados es equivalente al expuesto en (33) (Franco 2005).

3.1.3. Prueba de hipótesis marginal para los efectos fijos involucrados en el modelo

Para el modelo mixto de interés expuesto en (1), la hipótesis para los efectos fijos está dada por:

$$\begin{cases} H_0 : W\mu = 0; \quad \sigma_e^2 > 0, \quad \sigma_\theta^2 > 0 \\ H_a : W\mu \neq 0; \quad \sigma_e^2 > 0, \quad \sigma_\theta^2 > 0 \end{cases} \tag{34}$$

Para desarrollar la prueba, se parte de la función de verosimilitud asociada a la función de densidad conjunta de y y θ expresada en (17).

La prueba de hipótesis se hace realizando un razonamiento similar al presentado anteriormente. En este caso, se hace uso del estimador máximo verosímil de σ_e^2 dado en (27) y se llega a la expresión:

$$\vartheta^{\frac{2}{n}} = \frac{1}{\left(1 + \dfrac{R(\mu/\theta)}{(n - ran(W))\hat{\sigma}_e^2}\right)} \tag{35}$$

esta última expresión tiene las mismas características que la razón obtenida para llevar a cabo la prueba de la hipótesis general.

3.1.4. Prueba de hipótesis marginal para los efectos aleatorios involucrados en el modelo

Para el modelo mixto expuesto en (1), la hipótesis asociada a los efectos aleatorios está dada por:

$$\begin{cases} H_0 : Z\theta = 0; \qquad \sigma_e^2 > 0 \\ H_a : Z\theta \neq 0; \qquad \sigma_e^2 > 0 \end{cases} \tag{36}$$

Al igual que en la prueba para los efectos fijos, se parte de la función de verosimilitud asociada a la función de densidad conjunta de y y θ expresada en (17).

La prueba de hipótesis se hace utilizando la razón de verosimilitud generalizada y el estimador máximo verosímil de σ_e^2 dado en (27). Al hacer los reemplazos respectivos se tiene que:

$$\zeta^{\frac{2}{n}} = \frac{1}{\left(1 + \dfrac{R(\mu,\theta) - y'P_{Wp}y}{(n - ran(W))\hat{\sigma}_e^2}\right)} \tag{37}$$

De manera similar a lo expuesto en la prueba de hipótesis general, cuando se considera que la hipótesis marginal para los efectos aleatorios es como en (36), se evalúa el aporte de la parte aleatoria $Z\theta$ al modelo, como si esta fuese fija al igual que $W\mu$, razón por la cual no se tiene en cuenta que en el contexto del modelo mixto dado en (1), para la parte aleatoria del modelo el interés se centra en probar hipótesis referentes a la variabilidad de los efectos de los tratamientos.

De lo anterior, la hipótesis marginal para los efectos aleatorios apropiada para el modelo mixto de interés es de la forma:

$$\begin{cases} H_0 : \sigma_\theta^2 = 0; & \sigma_e^2 > 0 \\ H_a : \sigma_\theta^2 > 0; & \sigma_e^2 > 0 \end{cases}$$

En este caso, el estadístico de prueba asociado a esta hipótesis es equivalente a la involucrada en la razón de verosimilitud presentada en (37) (Franco 2005).

3.1.5. Análisis de varianza

En la tabla 1 se resume el análisis de varianza basado en las sumas de cuadrados tipo I, para el modelo mixto (1) después de imputar los m datos estimados; donde:

$$\begin{aligned} SCM &= R(\mu,\theta), \\ SCMF &= R(\mu^*) = y'P_{Wp}y, \\ SCMA &= R(\theta/\mu^*) = R(\theta) - y'ZP^{-1}Z'P_{Wp}y, \\ SCE &= y'y - R(\mu,\theta), \quad \text{y} \\ SCT &= y'y \end{aligned}$$

Los grados de libertad asociados a cada una de las causas de variación en el modelo corresponden al rango de la matriz involucrada en la suma de cuadrados de cada uno de los factores de interés (Hocking 1996). Según Yates (1933), en diseños con información faltante, aún después de imputada la información en el arreglo, tanto los grados de libertad del error como del total deben corregirse por la cantidad m de datos ausentes en el arreglo. Dicha corrección se incluye en las tablas y los cuadrados medios asociados a cada uno de los componentes del modelo se obtienen como el cociente entre la suma de cuadrados y los respectivos grados de libertad.

El análisis de varianza basado en las sumas de cuadrados tipo III, para el modelo mixto dado en (1) después de imputar los m datos estimados, se resume en la tabla 2, en donde:

$$SCM^* = R(\mu, \theta) - y'ZP^{-1}Z'P_{Wp}y,$$
$$SCMF^* = R(\mu/\theta) = R(\mu, \theta) - R(\theta), \quad \text{y}$$
$$SCT^* = y'y - y'ZP^{-1}Z'P_{Wp}y$$

Tabla 1: Análisis de varianza basado en las sumas de cuadrados tipo I.

Causa de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	Estadístico F
Modelo	$p+r-1$	SCM	CMM	$FG = \frac{CMM}{CME}$
E. fijos	p	$SCMF$	$CMMF$	$FF = \frac{CMMF}{CME}$
E. aleatorios	$r-1$	$SCMA$	$CMMA$	$FA = \frac{CMMA}{CME}$
Error	$n-p-r+1-m$	SCE	CME	
Total	$n-m$	SCT		

Tabla 2: Análisis de varianza basado en las sumas de cuadrados tipo III.

Causa de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	Estadístico F
Modelo	$p+r-2$	SCM^*	CMM^*	$FG^* = \frac{CMM^*}{CME}$
E. fijos	$p-1$	$SCMF^*$	$CMMF^*$	$FF = \frac{CMMF^*}{CME}$
E. aleatorios	$r-1$	$SCMA$	$CMMA$	$FA = \frac{CMMA}{CME}$
Error	$n-p-r+1-m$	SCE	CME	
Total	$n-1-m$	SCT^*		

3.2. Distribución de las formas cuadráticas obtenidas

A continuación se realiza un resumen de los principales resultados obtenidos al analizar la distribución de las formas cuadráticas involucradas en los estadísticos de prueba obtenidos a través del empleo de la razón de verosimilitud generalizada, para llevar a cabo la prueba de hipótesis general y las pruebas de hipótesis marginales para los efectos fijos y aleatorios que hacen parte del modelo mixto especificado en (1) (Franco 2005).

3.2.1. Prueba de hipótesis general

Suponiendo que $y/Z\theta \sim N(M\beta, \sigma_e^2 I_n)$, las formas cuadráticas asociadas al modelo y al error, generadas a partir de la función de verosimilitud, la cual hace

referencia a un modelo como el presentado en (30), tienen las siguientes distribuciones (Franco 2005):

$$\frac{SCM^0}{\sigma_e^2} = \frac{y'(M(M'M)^{-1}M')y}{\sigma_e^2} \sim \chi^{2'}_{(p+r-1,\psi)}$$

con

$$\psi = \frac{\beta' M' M \beta}{2\sigma_e^2}, \quad \text{y} \quad \frac{SCE^0}{\sigma_e^2} = \frac{y'(I_n - M(M'M)^{-1}M')y}{\sigma_e^2} \sim \chi^2_{(n-(p+r-1))}$$

Además, se puede comprobar que SCM^0 y SCE^0 son independientes y por ello:

$$\frac{(n-(p+r-1))(SCM^0)}{(p+r-1)(SCE^0)} \sim F'_{(p+r-1,n-(p+r-1),\psi)} \tag{38}$$

este cociente se distribuye como una F central, si y sólo si $H_0 : M\beta = 0$ es cierta.

El estadístico de prueba obtenido tiene una distribución F; sin embargo, esta prueba está asociada a un modelo como el presentado en (23), en el cual θ^0 es considerado un vector de efectos fijos más que de efectos aleatorios. Aún así, este estadístico de prueba es el empleado tradicionalmente e implementado en diferentes paquetes estadísticos como SAS, el cual, al contrario de lo que se creía, realmente no tiene en cuenta las características de los modelos mixtos en el procedimiento Proc Mixed.

Por lo anterior, para el modelo mixto de interés (1), el estadístico apropiado para probar la hipótesis general dada en (29) es realmente el que involucra las sumas de cuadrados SCM y SCE.

Las formas cuadráticas asociadas al modelo y al error, generadas a partir de la función de verosimilitud, dada en (17), son:

$$\frac{SCM}{\sigma_e^2} = \frac{R(\mu,\theta)}{\sigma_e^2} = \frac{y'(ZP^{-1}Z' + TW(W'TW)^{-1}W'T)y}{\sigma_e^2} \tag{39}$$

$$\frac{SCE}{\sigma_e^2} = \frac{y'y - R(\mu,\theta)}{\sigma_e^2} = \frac{y'[T - TW(W'TW)^{-1}W'T]y}{\sigma_e^2} \tag{40}$$

El uso de una función de verosimilitud como la expuesta en (17), implica que en el modelo mixto (1), $y \sim N(W\mu, V_p)$, con V_p dada en (5).

Como se muestra en Franco (2005), la matriz asociada a la suma de cuadrados del modelo posmultiplicada por V_p no es idempotente a menos que $P = Z'Z$ y que la estructura de variabilidad de y tenga la forma particular $V_p = \sigma_e^2 I_n$; en otras palabras, sólo se cumple si se tiene un modelo de efectos fijos, como el expuesto en (23). Por el contrario, la matriz asociada a la suma de cuadrados del error posmultiplicada por V_p es idempotente, de tal forma que:

$$\frac{SCM}{\sigma_e^2} \not\sim \chi^{2'}_{(p+r-1,\psi)} \qquad \text{y} \qquad \frac{SCE}{\sigma_e^2} \sim \chi^2_{(n-(p+r-1))}$$

Por otro lado, se puede comprobar que SCM y SCE no son independientes (Franco 2005). De lo anterior:

$$\frac{(n-(p+r-1))(SCM)}{(p+r-1)(SCE)}$$

no se distribuye exactamente como una $F'_{(p+r-1,n-(p+r-1),\psi)}$. Sin embargo, por la estrecha relación existente entre esta prueba y la presentada en (38), se asume que aproximadamente tiene esta distribución y que se comporta de manera aproximada como una F central, si y sólo si $H_0 : M\beta = 0$ es cierta.

Aunque este último resultado no parece ser satisfactorio, esta última prueba, aunque no tenga una distribución F exacta, es más apropiada que la empleada tradicionalmente, ya que tiene en cuenta la naturaleza del modelo mixto de interés presentado en (1), tal y como se ha venido exponiendo a lo largo de esta sección.

3.2.2. Prueba de hipótesis marginal para los efectos fijos involucrados en el modelo

Para la hipótesis marginal asociada a los efectos fijos dada en (34), la forma cuadrática de la parte fija del modelo, generada a partir de la función de verosimilitud apropiada, expuesta en (17), está dada por:

$$\frac{SCMF^*}{\sigma_e^2} = \frac{y'(TW(W'TW)^{-1}W'T)y}{\sigma_e^2} \sim \chi^{2'}_{(p-1,\phi)}$$

con $\qquad \phi = \dfrac{\mu' W' W \mu}{2\sigma_e^2}$

Adicionalmente, como la forma cuadrática asociada al error también tiene distribución χ^2 y se verifica que $SCMF^*$ y SCE son independientes, entonces:

$$\frac{(n-(p+r-1))(SCMF^*)}{(p-1)(SCE)} \sim F'_{(p-1,n-(p+r-1),\phi)}$$

este cociente se distribuye como una F central, si y sólo si $H_0 : W\mu = 0$ es cierta.

3.2.3. Prueba de hipótesis marginal para los efectos aleatorios involucrados en el modelo

Para la hipótesis marginal relativa a los efectos aleatorios dada en (36), la forma cuadrática asociada a la parte aleatoria del modelo, generada a partir de la función de verosimilitud expuesta en (17), está dada por:

$$\frac{SCMA}{\sigma_e^2} = \frac{R(\mu,\theta) - y' P_{Wp} y}{\sigma_e^2} = \frac{y'[(I_n - T)(I_n - P_{Wp})]y}{\sigma_e^2} \tag{41}$$

y la forma cuadrática asociada al error corresponde a la expuesta en (40). Si se estuviese trabajando con un modelo de efectos fijos como el expuesto en (23),

entonces:

$$P = Z'Z, \quad T = I_n - Z(Z'Z)^{-1}Z' = TT, \quad \text{y} \quad V_p = \sigma_e^2 I_n$$

de donde para la forma cuadrática dada en (41) se tendría la idempotencia y la independencia de la forma cuadrática asociada al error. Sin embargo, es claro que en este caso no se está trabajando con un modelo que esté involucrando únicamente efectos fijos, como el expuesto en (23) y, por lo tanto, como se muestra en Franco (2005), esto no se cumple a menos que $P = Z'Z$ y que la estructura de variabilidad de y tenga la forma particular $V_p = \sigma_e^2 I_n$.

Teniendo en cuenta que la única forma cuadrática que se distribuye χ^2 es la asociada al error, y que entre ésta y la de la parte aleatoria del modelo no hay independencia, entonces:

$$\frac{(n-(p+r-1))(SCMA)}{(r-1)(SCE)} \tag{42}$$

no se distribuye exactamente como una $F'_{(r-1,n-(p+r-1),\phi_1)}$, con $\phi_1 = \dfrac{\theta' Z' Z \theta}{2\sigma_e^2}$.

Al igual que el estadístico obtenido para la prueba de hipótesis general, en este caso no se tiene una distribución F exacta. Sin embargo, el estadístico dado en (42) resulta ser más apropiado que el utilizado comúnmente.

3.3. Metodología para el manejo de la información faltante en modelos mixtos

1. Hallar unas estimaciones iniciales de los componentes de varianza involucrados en el modelo, σ_θ^2 y σ_e^2, empleando la información con la que se cuenta originalmente en el arreglo a través del Método III de Henderson (1952).

2. Encontrar las estimaciones iniciales de los vectores de efectos fijos y aleatorios involucrados en el modelo, μ y θ, de acuerdo con las expresiones dadas en (13) y (14), respectivamente.

3. Realizar la estimación de la información faltante, empleando la expresión dada en (16) e imputarla en el diseño.

4. Encontrar unas nuevas estimaciones de los componentes de varianza involucrados en el modelo, σ_θ^2 y σ_e^2, empleando la información completa en el arreglo a través del Método III de Henderson (1952).

5. Hallar unas nuevas estimaciones de los vectores de efectos fijos y aleatorios involucrados en el modelo, μ y θ, utilizando la información completa en el diseño, de acuerdo con las expresiones (19) y (20), respectivamente.

6. Realizar los análisis de varianza con la información completa expuestos en las tablas 1 y 2, involucrando las últimas estimaciones de los componentes de varianza y de los vectores de efectos fijos y aleatorios, y la respectiva corrección en los grados de libertad del error y del total, de acuerdo con la cantidad de datos faltantes en el arreglo.

4. Caso de estudio

Tomando la adaptación realizada por Hocking (1996) del diseño factorial trabajado por Milliken & Johnson (1984), se tiene un diseño factorial 2×6, en el cual se considera la productividad de diferentes máquinas en una fábrica al ser maniobradas por diferentes operarios. El objetivo es comparar la productividad de dos máquinas usadas en la fabricación de cierto producto, para lo cual seis empleados fueron seleccionados aleatoriamente y cada uno de ellos maniobró cada una de las máquinas en tres ocasiones. En este diseño se asume que las máquinas son fijas mientras que los operarios son aleatorios. El arreglo factorial 2×6 sin interacción, descrito previamente, se caracteriza por el siguiente modelo:

$$y_{ijk} = m_i + p_j + e_{ijk} \tag{43}$$

con: $i = 1, 2$, $j = 1, 2, 3, 4, 5, 6$ y $k = 1, 2, 3$, y donde y_{ijk} es la productividad de la k-ésima réplica correspondiente a la i-ésima máquina maniobrada por el j-ésimo operario; m_i es la productividad promedio de la i-ésima máquina, p_j es el efecto aleatorio del j-ésimo operario, tal que $p_j \sim N(0, \sigma_p^2)$; y e_{ijk} es la componente del error aleatorio, tal que $e_{ijk} \sim N(0, \sigma_e^2)$.

Matricialmente, el modelo expresado en (43) se puede escribir como el modelo mixto de medias de celda, expresado en (1). En este caso, $y_{(36\times1)}$ es el vector de productividades, $W_{(36\times2)}$ es la matriz de incidencia de rango columna completo que asocia las observaciones con los respectivos componentes del vector $\mu_{(2\times1)}$ de valores medios desconocidos de los efectos fijos de las máquinas, $Z_{(36\times6)}$ es la matriz de incidencia que asocia las observaciones con los respectivos componentes del vector $\theta_{(6\times1)}$ de constantes desconocidas de los efectos aleatorios de los operarios y $e_{(36\times1)}$ es el vector de errores aleatorios.

Para este modelo se asume una estructura de variabilidad como la expuesta en (4), es decir:

$$\begin{aligned} \theta_{(6\times1)} &\sim N(0_{(6\times1)}, \sigma_\theta^2 I_{(6\times6)}), \\ e_{(36\times1)} &\sim N(0_{(36\times1)}, \sigma_e^2 I_{(36\times36)}) \quad \text{y} \\ Cov(\theta, e) &= 0_{(6\times36)} \end{aligned}$$

De lo anterior, se tiene una estructura conocida de la matriz de varianzas y covarianzas asociada al vector de variables aleatorias y, presentada en (5).

Para el arreglo con la información completa, los estimadores máximo verosímiles de los parámetros involucrados en el modelo dados en (8) y (9) son:

$$\dot{\mu} = \begin{bmatrix} 52.36 \\ 60.32 \end{bmatrix} \qquad \dot{\theta} = \begin{bmatrix} 1.58 \\ -0.39 \\ 6.90 \\ 0.83 \\ 1.69 \\ -10.62 \end{bmatrix} \tag{44}$$

Para hacer uso de la metodología propuesta, se considera que el diseño factorial de interés tiene $m = 10$ datos faltantes. En la tabla 3 se muestra el arreglo factorial 2×6 con la información faltante.

Tabla 3: Arreglo de la estructura factorial 2×6, con $m = 10$ datos faltantes.

Máquina	Operario	Réplica		
		1	2	3
I	1	52.0	x	x
	2	51.8	52.8	x
	3	60.0	x	x
	4	51.1	52.3	x
	5	50.9	51.8	51.4
	6	46.4	44.8	49.2
II	1	64.0	x	x
	2	59.7	60.6	59.0
	3	68.6	65.8	x
	4	63.2	62.8	62.2
	5	64.8	65.0	x
	6	43.7	44.2	43.0

En la tabla 4 se presentan los resultados de las estimaciones de los datos faltantes, asumiendo que $e_{2(6\times1)} \sim N(0_{(6\times1)}, \hat{\sigma}^2_{e(1)} I_{(6\times6)} = 13.5 I_{(6\times6)})$. Al hacer una revisión de las estimaciones obtenidas se puede ver que en general éstas son buenas al compararlas con los datos originales, aunque algunos datos son subestimados y otros sobreestimados. Se puede observar, además, que las estimaciones de los datos faltantes empleando los estimadores propuestos son cercanas a las resultantes al usar la metodología de estimación presentada en Melo & Melo (2005), cuyos resultados coinciden con los generados al hacer uso del algoritmo *EM* de Análisis de Varianza presentado en Hocking (1996), tal y como se muestra en Melo & Melo (2005). En este caso, los parámetros estimados considerando únicamente la información disponible en el arreglo son:

$$\tilde{\mu} = \begin{bmatrix} 52.65 \\ 59.77 \end{bmatrix} \qquad \dot{\theta}^0 = \begin{bmatrix} 1.52 \\ -0.25 \\ 6.63 \\ 1.30 \\ 1.19 \\ -10.39 \end{bmatrix}$$

Y las estimaciones iniciales de los componentes de varianza son: $\hat{\sigma}^2_\theta = 38.58$ y $\hat{\sigma}^2_e = 13.45$.

Las estimaciones de los componentes de varianza obtenidas después de estimar e imputar la información faltante a través del método propuesto en Melo & Melo (2005) son:

$\hat{\sigma}^2_\theta = 33.8125$ y $\hat{\sigma}^2_e = 8.8481$. Las obtenidas al hacer uso del método propuesto en el presente trabajo son respectivamente: $\hat{\sigma}^2_\theta = 33.9955$ y $\hat{\sigma}^2_e = 9.3358$.

Tabla 4: Valores estimados para la estructura factorial 2×6, con $m = 10$ datos faltantes.

y_2	$\hat{y}_2$	e_2	$\hat{y}_2^*$
52.1	54.2	-2.76	51.4
52.1	54.2	-1.10	53.1
52.3	52.4	-0.71	51.7
59.9	59.3	0.44	59.7
59.9	59.3	0.80	60.1
51.7	54.0	-1.33	52.7
63.9	61.3	0.27	61.6
63.9	61.3	1.31	62.6
67.2	66.4	-0.44	66.0
64.8	61.0	2.12	63.1

En la tabla 5 se presentan las convenciones empleadas para la identificación de los análisis de varianza obtenidos a través del empleo de las diferentes metodologías de estimación de información faltante con el modelo mixto de medias de celda.

Tabla 5: Convenciones empleadas para la identificación de los análisis de varianza.

Convención	Análisis de varianza
$CTD(I)$ y $CTD(III)$	*Con todos los datos originales observados, tipos I y III, respectivamente.*
$CDI(I)$ y $CDI(III)$	*Con datos imputados a través de la técnica presentada en* Melo & Melo (2005), *tipos I y III, respectivamente.*
$CDI^*(I)$ y $CDI^*(III)$	*Con datos imputados a través del método propuesto en este trabajo, tipos I y III, respectivamente.*

Para los análisis de varianza con los datos completos, los grados de libertad están dados por: $p = 2$, $r - 1 = 5$, $n - (p + r - 1) = 36 - 7 = 29$, $n = 36$.

Resulta interesante ver que en el contexto de los modelos mixtos, los análisis de varianza tradicionales e implementados en paquetes estadísticos como SAS, son generados a partir de estimaciones de los parámetros μ y θ que son poco apropiadas al omitir la naturaleza de los modelos de interés. Por ejemplo, para el diseño con datos completos, las sumas de cuadrados son generadas con base en las siguientes estimaciones de los parámetros:

$$\dot{\mu}^0 = \begin{bmatrix} 52.36 \\ 60.32 \end{bmatrix} \qquad \dot{\theta}^0 = \begin{bmatrix} 35.03 \\ -8.63 \\ 152.63 \\ 18.48 \\ 37.50 \\ -235.02 \end{bmatrix}$$

Como puede verse, para el vector de efectos aleatorios esta estimación difiere mucho de la presentada en (44), la cual es la apropiada, ya que como se muestra en

(9), tiene en cuenta que en el modelo de interés θ es un vector de efectos aleatorios y no de efectos fijos. Por el contrario, los análisis de varianza propuestos en este trabajo son generados a partir de estimaciones de los parámetros μ y θ que tienen en cuenta la naturaleza de los modelos de interés.

En la tabla 6 se presentan los resultados obtenidos en el análisis de varianza, utilizando las pruebas propuestas en este trabajo. Los grados de libertad están dados por: $n-(p+r-1)-m = 36-7-10 = 19$ y $n-m = 36-10 = 26$.

Tabla 6: Análisis de varianza propuesto para el diseño factorial 2×6, $(m = 10)$.

Metodología	$CTD(I)$	$CTD(III)$	$CDI(I)$	$CDI(III)$	$CDI^*(I)$	$CDI^*(III)$
SCM	116000	1776	115345	1611	115257	1683
$SCMF$	114848	625	114238	504	114140	566
$SCMA$	1151	1151	1107	1107	1118	1118
SCE	233	233	208	208	220	220
SCT	116233	2009	115554	1819	115477	1903
CMM	16571	296	16478	269	16465	281
$CMMF$	57424	625	57119	504	57070	566
$CMMA$	230	230	221	221	224	224
CME	8	8	11	11	12	12
FG	2062	37	1503	25	1425	24
FF	7147	78	5210	46	4938	49
FA	29	29	20	20	19	19

De la tabla de análisis de varianza se puede ver que los valores de los estadísticos de prueba obtenidos a través del empleo de la metodología propuesta son cercanos a los resultantes al usar el análisis de varianza tradicional cuando se han imputado los datos faltantes empleando los estimadores de Melo & Melo (2005), y son los que menos se alejan de los obtenidos al realizar el análisis con los datos completos. Esto se debe a que en las metodologías de estimación tradicionales se adicionan residuales nulos a los valores estimados, razón por la cual las estimaciones quedan totalmente determinadas por la información observada en el arreglo, al contrario de lo que sucede con la metodología propuesta, en la cual, para que las estimaciones cuenten con una variabilidad interna un poco más fiel a la realidad, se adiciona un error aleatorio al vector de valores estimados, con lo que además se logra una disminución de la correlación lineal entre los subvectores del vector de información.

5. Conclusiones

Se propone una metodología para la estimación de información faltante con el modelo mixto de medias de celda, la cual proporciona estimadores insesgados y permite estimar tanto los datos individuales en las celdas como los valores de

las celdas vacías. En la metodología propuesta se adiciona un error aleatorio al vector de valores estimados, con lo cual se disminuye la correlación lineal entre los vectores de información observada y estimada.

Los estadísticos de prueba construidos tradicionalmente e implementados en paquetes estadísticos como SAS, omiten las características de los modelos mixtos, y por ello resultan menos apropiados que los generados en el presente trabajo, los cuales tienen en cuenta la naturaleza de los modelos de interés.

Recibido: marzo de 2006
Aceptado: mayo de 2006

Referencias

Allan, F. & Wishart, J. (1930), 'A Method of Estimating the Yield of a Missing Plot in Field Experiments', *J. Agric. Sci.* **20**, 399–406.

Barroso, L. P., Bussab, W. O. & Knott, M. (1998), 'Best Linear Unbiased Predictor in the Mixed Model with Incomplete Data', *Communications in Statistics* **21**, 121–129.

Bartlett, M. S. (1937), 'Some Examples of Statistical Methods of Research in Agriculture and Applied Biology', *Journal of the Royal Statistical Society* **4**(2), 137–183.

Corbeil, R. R. & Searle, S. R. (1976), 'Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model', *Technometrics* **18**, 31–38.

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

Fisher, R. A. (1935), *The Design of Experiments*, Oliver & Boyd, Edinburgh.

Franco, D. C. (2005), Metodología para la estimación de información faltante, imputación y estadísticos de prueba en modelos mixtos a dos vías de clasificación, Tesis de maestría, Universidad Nacional de Colombia, Departamento de Estadística, Bogotá D. C.

Gallo, J. & Khuri, A. I. (1990), 'Exact Test for the Random and Fixed Effects in an Unbalanced Mixed Two Way Cross-Classification Model', *Biometrics* **46**, 1087–1095.

Harville, D. A. & Carriquiry, A. L. (1992), 'Classical an Bayesian Prediction as Applied to an Unbalanced Mixed Linear Model', *Biometrics* **48**, 987–1003.

Henderson, C. R. (1952), Estimation of Variance and Covariance Components, Cornell University. North Carolina Summer Statistics Conference.

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

Lindstrom, M. J. & Bates, D. M. (1988), 'Newton Raphson and EM Algorithms for Linear Mixed Effects Models for Repeated Measures Data', *Journal of the American Statistical Association* **83**, 1014–1022.

Little, R. & Rubin, D. (2002), *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.

Melo, O. O. & Melo, S. E. (2005), Metodología para la estimación de datos faltantes en modelos mixtos de medias de celdas, *in* 'Modelamiento Estadístico. Memorias del XV Simposio de Estadística', Universidad Nacional de Colombia. Departamento de Estadística, Bogotá.

Milliken, G. & Johnson, D. (1984), *Analysis of Messy Data*, Vol. I of *Designed Experiments*, Van Nostrand Reinhold, New York.

Murray, L. W. & Smith, D. W. (1985), 'Estimability, Testability, and Connectedness in the Cell Means Models', *Communications in Statistics* **14**, 1889–1917.

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

Searle, S. R., Casella, G. & McCulloch, C. (1992), *Variance Components*, John Wiley & Sons, New York.

Sheffé, H. (1959), *The Analysis of Variance*, Wiley.

Thomsen, I. (1975), 'Testing Hyphoteses in Unbalanced Variance Components Models for Two-Way Layouts', *Annals of Statistics* **3**, 257–265.

Yates, F. (1933), 'The Analysis of Replicate Experiments When the Field Results are Incomplete', *Emp. Journ. Exp. Agric.* (1).