Revista Colombiana de Estadística
Volumen 25 N° 2. Págs 127 *a* 143. Diciembre 2002

Medidas repetidas con datos faltantes: estimación de parámetros vía análisis de covarianza

LUZ MERY GONZALEZ.[*]
LUIS ALBERTO LÓPEZ.[**]

Resumen

En este artículo se lleva a cabo la estimación de parámetros y se obtienen diferentes sumas de cuadrados ajustadas para diseños balanceados, en medidas repetidas, con información incompleta a través de tres procedimientos: el método de análisis de varianza de Bartlett; un método multivariado con base en los datos completos y finalmente un método multivariado alterno usando toda la información disponible en el arreglo experimental. Con los tres procedimientos anteriores se llevan acabo aplicaciones numéricas.

Palabras Clave: *Análisis de covarianza, información faltante, datos longitudinales, medidas repetidas, mínimos cuadrados generalizados, análisis multivariado.*

Abstract

In this paper the estimation of parameters and the different adjusted sums of squares for balanced designs in repetead measures with incomplete information is done through three procedures: the Bartlett's method of covariance analysis; a multivariate method with complete data, and finally, an alternative multivariate method using the available information in the experimental arrangements. Numerical applications using the above procedures are done.

Keywords: *Covariance analysis, missing information, longitudinal data, repeated measures, generalized least squares, multivariate analysis.*

[*]Profesora Asistente, Departamento de Estadística; Universidad Nacional de Colombia; e-mail: lmgong@matematicas.unal.edu.co
[**]Profesor Asociado, Departamento de Estadística; Universidad Nacional de Colombia; e-mail: alopez@matematicas.unal.edu.co

1. Introducción

Las investigaciones con datos longitudinales involucran observaciones de un conjunto de unidades experimentales (humanos, lugares geográficos, animales, etc.) clasificados en diferentes subpoblaciones teniendo en cuenta uno o más factores (raza, lugar de origen, tipo de dieta, etc.) a lo largo de diversas condiciones de evaluación (tiempos, dosis, etc.). En este sentido, se pueden destacar los trabajos de Laird y Ware [10], Ware [16], Andrade y Singer [2], Liang y Zeger [11] y Andreoni [3] entre otros.

La diferencia entre un estudio longitudinal y uno de medidas consiste en que en el primero, los individuos participantes son seguidos por periodos extensos y en el segundo, las observaciones son recolectadas en periodos de tiempo relativamente cortos y, frecuentemente, bajo condiciones experimentales. Esta diferencia se puede ver más en detalle en Crowder y Hand [5].

Otra característica fundamental asociada a los estudios con medidas repetidas es la posibilidad de correlación no nula entre las observaciones realizadas en las mismas Unidades Experimentales.

Infortunadamente, en muchos casos no se pueden usar las técnicas clásicas de análisis porque se pierde observaciones o porque el diseño es desbalanceado por alguna razón, o porque hay covariables que varían en el tiempo. Una revisión de literatura sobre observaciones faltantes en datos multivariables puede encontrarse en Afifi y Elashoff [1], donde se resaltan los trabajos de Yates [18] en 1933, Bartlett [4] en 1937, Tocher [15] en 1952, Wilkinson [17] en 1958 y Dear [6] en 1959 entre otros, como los pioneros en estudiar métodos para la estimación de información faltante.

Algunos autores que han tratado este tema son Timm y Mieczkowski [14], Crowder y Hand [5] y Laird et al [9]; pero no han hecho propuestas de estimación basados implicitamente en el método de Bartlett, el cual es apropiado cuando se tiene poca información faltante.

2. Estimación de parámetros en medidas repetidas

En esta sección se llevan a cabo los desarrollos teóricos y se muestran aplicaciones de la técnica del análisis de covarianza, como método propuesto para la estimación de parámetros en diseños de medidas repetidas con información faltante. Inicialmente se implementa el método de Bartlett particionando en

forma adecuada el vector de respuestas, según contenga o no información faltante, luego se procede a la imputación de la información faltante en forma multivariada y posteriormente se muestra el procedimiento para la estimación de los parámetros así como, para la obtención de las sumas de cuadrados del modelo y del error corregida una vez hecha la imputación.

2.1. Método del análisis de covarianza en medidas repetidas

En esta subsección se implementa el método de Bartlett para la imputación de información faltante en modelos con medidas repetidas bajo el supuesto de perdida de información en forma aleatoria. Se supone que se observan n individuos bajo t condiciones de evaluación y que se presentan m_0 valores perdidos en n_0 de los n individuos iniciales ($n_0 \leq m_0$), pudiendo en este caso representar esa información con el modelo de covarianza (Véase [13]):

$$y = X\beta + Z\gamma + e, \tag{1}$$

siendo y el vector respuesta de orden $nt \times 1$ ya que n individuos fueron evaluados en t diferentes ocasiones, X la matriz diseño de orden $nt \times p$, β el vector de parámetros desconocidos de orden $p \times 1$, Z la matriz de covariables de orden $nt \times m_0$, γ el vector de coeficientes para las covariables de orden $n_0 \times 1$ y e el vector de desviaciones de orden $nt \times 1$. Sin perder generalidad, se puede ordenar el vector de observaciones de forma tal que las primeras componentes correspondan a los tiempos en los cuales se perdió algún dato. Si en total se tienen m_0 datos faltantes en n_0 individuos, entonces el resto de componentes $(n_0 + j)$ con $j = 1, \ldots, n$, corresponden a los individuos con al menos una observación en el tiempo, como se muestra en (2):

$$y = \left(\begin{array}{cc} \left.\begin{array}{c} \tilde{y}^1 \\ \vdots \\ \tilde{y}^{n_0} \end{array}\right\} & \begin{array}{c} \text{Valores iniciales en los} \\ \text{tiempos donde no se obtuvo} \\ \text{información.} \end{array} \\ \left.\begin{array}{c} y_{n_0+1} \\ \vdots \\ y_{n_0+n} \end{array}\right\} & \begin{array}{c} \text{indiviuos con la informa-} \\ \text{ción observada} \end{array} \end{array} \right), \tag{2}$$

En forma equivalente a como se arregla el vector de respuestas, se ordena la matriz diseño y los parámetros del modelo como:

$$X = \begin{pmatrix} X^1 \\ \vdots \\ X^{n_0} \\ X_{n_0+1} \\ \vdots \\ X_{n_0+n} \end{pmatrix}; \gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_{m_0} \end{pmatrix}; \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},$$

con p número de parámetros poblacionales desconocidos, X^l de orden $t_l \times p$, $l = 1, 2, ..., n_0$, matriz diseño asociada a la información faltante en los individuos donde se perdió alguna información, X_i , $i = n_0 + j$; $j = 1, ..., n$ matriz diseño de orden $t_i \times p$ asociada con la información observada. En el modelo en estudio, y de orden $nt \times 1$ es el vector de observaciones, X de orden $nt \times p$ es una matriz de valores conocidos, β de orden $p \times 1$ es el vector de parámetros, γ de orden $m_0 \times 1$ es el vector de coeficientes para las covariables de los valores faltantes, e de orden $nt \times 1$ es la matriz de desviaciones $e = y - E(y)$ no observable, y Z de orden $nt \times m_0$ es la matriz de constantes conocidas de la forma:

$$(3) \qquad Z = \begin{pmatrix} Z^1 \\ \vdots \\ Z^{n_0} \\ Z_{n_0+1} \\ \vdots \\ Z_{n_0+n} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \cdots & 1 \\ & & 0_{(t_{n_0+1} \times m_0)} & & \\ & & \vdots & & \\ & & 0_{(t_{n_0+n} \times m_0)} & & \end{pmatrix} = \begin{pmatrix} I_{(m_0)} \\ 0_{(k \times m_0)} \end{pmatrix},$$

con $k = \sum_{i=1}^{n} t_i$. La notación propuesta se ilustra con la información del ejemplo 2.1.1.

Ejemplo 2.1.1: La siguiente información tomada de Crowder y Hand [5] presenta el efecto de una dieta suplementaria de vitamina E en el crecimiento de cerdos raza guinea. El peso corporal de cada animal fue registrado al final de las semanas 1, 3, 4, 5, 6 y 7. A cada uno de estos animales se les dio una sustancia inhibidora durante la semana uno, la terapia de la vitamina E se comenzo en la semana cinco. Tres grupos de animales, cinco en cada grupo, recibieron dosis de vitamina E: cero, baja y alta, respectivamente. Para la comprensión de este modelo solo se registra en la Tabla 1, el peso corporal (en gramos) de las semanas uno, tres y cuatro, con cinco animales del grupo uno y cuatro animales del grupo dos, eliminando en forma aleatoria cuatro datos del conjunto de información.

Tabla 1: Efecto de dietas suplementarias sobre las tazas de crecimiento en cerdos guinea con pérdida aleatoria de datos.

		Semanas		
Grupo	Animal	1	3	4
1	1	455	•	510
	2	467	565	610
	3	445	•	580
	4	485	542	594
	5	480	500	550
2	6	514	•	•
	7	440	480	536
	8	495	570	569
	9	520	590	610

Fuente: Datos Adaptados de Crowder y Hand [5]. Ejemplo 3.1 pág. 27

• : Datos que fueron eliminados.

De la tabla se tiene que: $n = 9, t = 3, m_0 = 4, n_0 = 3, p = 4$; $\beta_1, \beta_2, \beta_3$ parámetros asociados con el efecto de semana 1, 3 y 4, respectivamente, y β_4 parámetro asociado con el efecto del grupo.

Inicialmente, el vector respuesta esta dado por:

$$y = \left(\underbrace{455, ., 510}_{Ind,1}; \underbrace{467, 565, 610}_{Ind,2}; \underbrace{445, ., 580}_{Ind,3}; \cdots ; \underbrace{520, 590, 610}_{Ind,9} \right)^t ,$$

Al ordenarlo y reemplazar los datos faltantes por valores iniciales cero, se tiene:

$$y^t = \left(\underbrace{0}_{\tilde{y}^1}; \underbrace{0}_{\tilde{y}^2}; \underbrace{0, 0}_{\tilde{y}^3}; \underbrace{455, 510}_{y_{3+1}}; \underbrace{467, 565, 610}_{y_{3+2}}; \underbrace{445, 580}_{y_{3+3}}; \cdots ; \underbrace{520, 590, 610}_{y_{3+9}} \right) ,$$

$$X = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ & I_3|J_3 & & \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ & I_3|J_3 & & \\ & I_3|J_3 & & \\ 1 & 0 & 0 & -1 \\ & I_3|(-1)J_3 & & \\ & I_3|(-1)J_3 & & \\ & I_3|(-1)J_3 & & \end{pmatrix} \begin{matrix} \} \rightarrow X^1 \\ \} \rightarrow X^2 \\ \} \rightarrow X^3 \\ \} \rightarrow X_{3+1} \\ \} \rightarrow X_{3+2} \\ \} \rightarrow X_{3+3} \\ \} \rightarrow X_{3+4} \\ \} \rightarrow X_{3+5} \\ \} \rightarrow X_{3+6} \\ \} \rightarrow X_{3+7} \\ \} \rightarrow X_{3+8} \\ \} \rightarrow X_{3+9} \end{matrix}, e = \begin{pmatrix} e^{12} \\ e^{32} \\ e^{62} \\ e^{63} \\ e_{11} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \\ \vdots \\ e_{61} \\ e_{71} \\ \vdots \\ e_{93} \end{pmatrix},$$

$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$; $\gamma = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{pmatrix}$; $Z_{(27\times 4)} = \begin{pmatrix} I_{(4)} \\ 0_{(23\times 4)} \end{pmatrix}$ con I_s matriz identidad de orden s y J_r vector de unos de tamaño $r \times 1$ y donde $I_s|J_r$ es una matriz aumentada.

El estimador mínimos cuadrados generalizados de β se obtiene minimizando la forma cuadrática $\sum_{i=1}^{n_0+n} Q_i(\beta, \Sigma_i)$, donde Σ_i de orden $t_i \times t_i$ es una submatriz de Σ de componentes de varianzas asociadas a los tiempos donde hay información para y_i.

Si Σ es conocida, entonces β tiene como estimador a:

$$\hat{\beta} = \left(\sum_{i=1}^{n_0+n} X_i^t \Sigma_i^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{n_0+n} X_i^t \Sigma_i^{-1} y_i \right). \tag{4}$$

Si Σ es desconocida, la estimación de β se obtiene a partir de la expresión:

$$\hat{\beta} = \left(\sum_{i=1}^{n_0+n} X_i^t \hat{\Sigma}_i^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{n_0+n} X_i^t \hat{\Sigma}_i^{-1} y_i \right). \tag{5}$$

Crowder y Hand [5] muestran que si hay datos faltantes entonces no hay soluciones explícitas para $\hat{\beta}$ y $\hat{\Sigma}$ en forma separada, y así, la solución para las ecuaciones debe hacerse en forma iterativa. Para efectos de este trabajo, se tomó como estimación de la matriz de covarianza (Σ) las estimaciones de las

componentes dadas en el PROC MIXED de SAS[1] y la matriz de covarianza combinada.

Al considerar el modelo 1, con las características descritas y tener en cuenta que existen datos faltantes se tiene como función a minimizar:

$$\sum_{i=1}^{n_0+n} Q_i\left(\beta,\gamma,\Sigma_i\right) = \sum_{i=1}^{n_0+n} \left(y_i - X_i\beta - Z_i\gamma\right)^t \Sigma_i^{-1} \left(y_i - X_i\beta - Z_i\gamma\right). \tag{6}$$

Nuevamente, teniendo en cuenta los resultados de Bartlett, se separa la información en dos partes, una con los tiempos en los cuales se presentan datos faltantes y la otra con los individuos y tiempos con datos observados, es decir:

$$\begin{aligned}\sum_{i=1}^{n_0+n} Q_i\left(\hat\beta,\hat\gamma,\Sigma_i\right) &= \sum_{l=1}^{n_0} Q_l\left(\hat\beta,\hat\gamma,\Sigma^l\right) + \sum_{i=n_0+1}^{n} Q_i\left(\hat\beta,\hat\gamma,\Sigma_i\right) \\ &= \sum_{l=1}^{n_0} \left(y^l - X^l\hat\beta - Z^l\hat\gamma\right)^t \left(\Sigma^l\right)^{-1} \left(y^l - X^l\hat\beta - Z^l\hat\gamma\right) \\ &\quad + \sum_{i=n_0+1}^{n_0+n} \left(y_i - X_i\hat\beta - Z_i\hat\gamma\right)^t \Sigma_i^{-1} \left(y_i - X_i\hat\beta - Z_i\hat\gamma\right).\end{aligned} \tag{7}$$

Por la construcción de Z, la expresión (7) es equivalente a:

$$\begin{aligned}&\sum_{l=1}^{n_0} \left(y^l - X^l\hat\beta - Z^l\hat\gamma\right)^t \left(\Sigma^l\right)^{-1} \left(y^l - X^l\hat\beta - Z^l\hat\gamma\right) \\ &\qquad + \sum_{i=n_0+1}^{n_0+n} \left(y_i - X_i\hat\beta\right)^t \Sigma_i^{-1} \left(y_i - X_i\hat\beta\right).\end{aligned} \tag{8}$$

Al minimizar la segunda parte de la expresión (8) y tener Σ desconocida, la estimación de β se obtiene a partir de la expresión:

$$\hat\beta = \left(\sum_{i=n_0+1}^{n_0+n} X_i^t \hat\Sigma_i^{-1} X_i\right)^{-1} \left(\sum_{i=n_0+1}^{n_0+n} X_i^t \hat\Sigma_i^{-1} y_i\right). \tag{9}$$

Para los datos de la Tabla 1, se muestra la estimación obtenida. Como se desconoce la matriz de covarianzas (Σ), se estima usando la matriz de covarianza combinada, obtenida a partir de un procedimiento iterativo implementado en SAS-IML (véase González L. M. [7]), el resultado de esta estimación es:

[1]Vease la guia del usuario de SAS [12].

$$\hat{\Sigma} = \begin{pmatrix} 728{,}56429 & 917{,}95238 & 721{,}2 \\ 917{,}95238 & 2092{,}1905 & 1434{,}619 \\ 721{,}2 & 1434{,}619 & 1484{,}5429 \end{pmatrix}.$$

en forma iterativa, la estimación de (9) dió los siguientes resultados:

$$\hat{\beta}^t = \begin{pmatrix} 479{,}13503 & 540{,}26207 & 572{,}93697 & -11{,}21525 \end{pmatrix}.$$

Estas estimaciones fueron usadas para la imputación de la información faltante. Con este estimador, y despejando de las ecuaciones normales asociadas al modelo (1), se tiene que $Z\hat{\gamma} = y - X\hat{\beta}$ y para las primeras m_0 componentes se satisface:

$$Z^l\hat{\gamma} = \tilde{y}^l - X^l\hat{\beta}, \tag{10}$$

con $l = 1, \ldots, n_0$, para las demás componentes $Z_i\hat{\gamma} = 0$, con $i = n_0+1, \ldots, n_0+n$ al tenerse en cuenta que $Z_i = 0$ para todo $i > n_0$, reemplazando esta estimación en (8) se obtiene:

$$\begin{aligned} \sum_{l=1}^{n_0} \left(\tilde{y}^l - X^l\hat{\beta} - \left(\tilde{y}^l - X^l\hat{\beta}\right)\right)^t \left(\hat{\Sigma}^l\right)^{-1} \left(\tilde{y}^l - X^l\hat{\beta} - \left(\tilde{y}^l - X^l\hat{\beta}\right)\right) \\ + \sum_{i=n_0+1}^{n_0+n} \left(y_i - X_i\hat{\beta}\right)^t \hat{\Sigma}_i^{-1} \left(y_i - X_i\hat{\beta}\right) \\ = \sum_{i=n_0+1}^{n_0+n} \left(y_i - X_i\hat{\beta}\right)^t \hat{\Sigma}_i^{-1} \left(y_i - X_i\hat{\beta}\right), \end{aligned} \tag{11}$$

al minimizar (11) respecto a β, se llega a la solución encontranda en (9), con esta solución y despejando γ de (10) se halla que:

$$\begin{pmatrix} \underbrace{0}_{1} & \cdots & \underbrace{0}_{k-1} & \underbrace{1}_{k} & \underbrace{0}_{k+1} & \cdots & \underbrace{0}_{m_0} \end{pmatrix} \begin{pmatrix} \hat{\gamma}_1, & \cdots, & \hat{\gamma}_{k-1}, & \hat{\gamma}_k, & \hat{\gamma}_{k+1}, & \cdots, & \hat{\gamma}_{m_0} \end{pmatrix}^t =$$

$$\begin{aligned} \tilde{y}^k - X_k \begin{pmatrix} \hat{\beta}_1 & \cdots & \hat{\beta}_p \end{pmatrix}^t; \\ \hat{\gamma}_k = \tilde{y}^{[k]} - X_{[k]}\hat{\beta}. \end{aligned} \tag{12}$$

Nótese que $k = 1, ..., m_0$, donde $\tilde{y}^{[k]}$ es el valor inicial "conjeturado" para el k-ésimo valor faltante, $X_{[k]}$ es la fila de la matriz diseño asociada al k-ésimo

valor faltante y $\hat{\gamma}_k$ es el coeficiente estimado de la covariable para el k-ésimo valor faltante.

Como:

$$\hat{y}^{[k]} = X_{[k]}\hat{\beta}, \tag{13}$$

al reemplazar en (12) se tiene que el predictor $\hat{y}^{[k]}$ para el k-ésimo dato faltante es igual al valor conjeturado para el k-ésimo dato faltante menos el coeficiente de la covariable para el k-ésimo valor faltante, es decir $\hat{y}^{[k]} = \tilde{y}^{[k]} - \hat{\gamma}_k$. Utilizando la estimación de β se encuentran los valores estimados para la información faltante. Esta predicción es presentada en la tabla (2).

Tabla 2: Resultados de la predición de la información faltante usando análisis de covarianza.

k	Grupo	Animal	Tiempo	$\hat{y}^{[k]}$
1	1	1	Sem. 3	527.15865
2	1	3	Sem. 3	527.15865
3	2	6	Sem. 3	549.82606
4	2	6	Sem. 4	586.21752

2.2. Enfoque multivariado para la imputación de información

Una alternativa para el problema propuesto consiste en imputar la información haciendo uso de un enfoque multivariado. Para ello se utilizan los desarrollos encontrados en Timm y Mieczkowski [14] quienes muestran inicialmente un modelo lineal multivariado para analizar medidas repetidas cuando no se ha perdido información. Así, en la subsección 2.2.1 se sigue la metodología presentada por ellos usando solo la información de aquellas unidades que se observaron en su totalidad, y posteriormente, en la subsección 2.2.2, se utiliza toda la información disponible a la vez que se encuentra una relación entre ellas. Debido a que el segundo método es iterativo, la relación se busca a nivel de la primera iteración antes de imputar los datos. Estos resultados se ilustran con los datos de la tabla 1.

2.2.1. Enfoque multivariado - Casos completos

Timm y Mieczkowski [14] muestran que un diseño en medidas repetidas univariado con información completa puede ser presentado como un modelo

lineal multivariado. Partiendo de este resultado, se ajusta un modelo donde solamente se tienen en cuenta los individuos que tienen información completa $(n - n_0)$ y reordenando las observaciones se llega a un modelo univariado para medidas repetidas:

$$y_C = X_C\beta^{(1)} + e_C, \tag{14}$$

con $E(y_C) = X_C\beta^{(1)}$ y $Cov(y_C) = I_{n-n_0} \otimes \Sigma = \Omega_C$ donde y_C es el vector de respuestas de orden $(t(n - n_0)) \times 1$, X_C es la matriz diseño de orden $(t(n - n_0)) \times p$, con $p = p^*t$, $\beta^{(1)}$ es el vector de parámetros desconocidos de orden $p \times 1$, e_C vector de errores de orden $(t(n - n_0)) \times 1$ y Σ es la matriz de covarianzas. Ahora, si se tiene en cuenta que el haber observado la información completa significa que todos los individuos fueron observados en todas las ocasiones de evaluación (t-tiempos), entonces el vector de respuestas y_C, se puede escribir como una matriz Y de orden $n - n_0$ filas por t columnas, $X_C\beta^{(1)}$ como el producto de tres matrices: X_W de orden $t \times t$ que corresponde a la matriz diseño de los tiempos en un modelo reparametrizado, B de orden $p^* \times t$ matriz de parámetros desconocidos, X_B de orden $(n - n_0) \times p^*$ matriz diseño correspondiente a los factores en un modelo reparametrizado y e_C como $U_{((n-n_0)\times t)}$ matriz de errores. Con lo anterior (14) se reescribe como:

$$Y^t = X_W B^t X_B^t + U^t. \tag{15}$$

El hecho de utilizar sólo los casos completos permite que la matriz asociada a estos se pueda escribir como $X_C = X_B \otimes X_W$, es decir, X_C es separable [2], entonces el mejor estimador lineal insesgado (MELI) de B es:

$$\begin{aligned} \hat{B} &= \left(X_B^t X_B\right)^{-1} X_B^t Z \\ &= \left(X_B^t X_B\right)^{-1} X_B^t Y \left(X_W^t\right)^{-1}, \end{aligned} \tag{16}$$

que es el estimador multivariado.

Ahora, al aplicar el operador $Vec(.)$ a la traspuesta de la ecuación (16), se tiene:

$$\hat{\beta}_M^{(1)} = \left[\left(X_B^t X_B\right)^{-1} X_B^t \otimes X_W^{-1}\right] Vec\left(Y^t\right). \tag{17}$$

Por otro lado, teniendo en cuenta que $X_C = X_B \otimes X_W$, el estimador univariado

[2] la condición que señala que la matriz diseño univariada X puede ser representada como el producto kronecker $X = X_B \otimes X_W$ es llamada condición de separabilidad

de mínimos cuadrados generalizados de (14) es:

$$\hat{\beta}^{(1)} = \left(X_C^t \hat{\Omega}_C^{-1} X_C\right)^{-1} \left(X_C^t \hat{\Omega}_C^{-1} y_C\right); \text{con } \hat{\Omega} = I_C \otimes \hat{\Sigma} = I_{n-n_0} \otimes \hat{\Sigma}$$

$$\hat{\beta}^{(1)} = \left\{ \left[\left(X_B^t X_B\right)^{-1} X_B^t \right] \otimes \left[\left(X_W^t \hat{\Sigma}^{-1} X_W\right)^{-1} X_W^t \hat{\Sigma}^{-1} \right] \right\} y_C. \tag{18}$$

Para efectos de estimación de $\hat{\beta}^{(1)}$, la matriz de covarianza Σ se puede estimar usando la información completa (casos completos) o toda la información disponible.

Si la matriz diseño X_W es de rango completo y X_C es separable, entonces se satisface que $\left(X_W^t \Sigma^{-1} X_W\right)^{-1} X_W^t \Sigma^{-1} = X_W^{-1} \Sigma \left(X_W^t\right)^{-1} X_W^t \Sigma^{-1} = X_W^{-1} \Sigma \Sigma^{-1} = X_W^{-1}$. Este resultado muestra la equivalencia entre la estimación multivariada y univariada, es decir, (17) y (18) producen resultados idénticos, así,

$$\hat{\beta}^{(1)} = \left\{ \left(X_B^t X_B\right)^{-1} X_B^t \otimes X_W^{-1} \right\} y_C = \hat{\beta}_M^{(1)}. \tag{19}$$

Obtenida la estimación de $\beta^{(1)}$ a partir de (19), se procedio a encontrar la estimación del vector de predicción, a partir de la siguiente expresión:

$$\begin{aligned} \hat{y} &= \left(X_{B:C} \otimes X_W\right) \left\{ \left(X_B^t X_B\right)^{-1} X_B^t \otimes X_W^{-1} \right\} y_C \\ &= \left(X_{B:C} \left(X_B^t X_B\right)^{-1} X_B^t \otimes I_t \right) y_C, \end{aligned} \tag{20}$$

siendo $X_{B:C}$ la matriz diseño con toda la información.

La matriz de covarianza estimada cuando se tiene la información completa es obtenida a partir de la expresión:

$$\begin{aligned} \hat{Cov}\left(\hat{\beta}^{(1)}\right) &= \left(X_C^t \hat{\Omega}_C^{-1} X_C\right)^{-1} \\ \hat{Cov}\left(\hat{\beta}^{(1)}\right) &= \left(X_B^t X_B\right)^{-1} \otimes \left(X_W^t \hat{\Sigma}^{-1} X_W\right)^{-1}. \end{aligned} \tag{21}$$

siguiendo con los datos propuestos para ilustrar este trabajo, se sigue que la estimación de la matriz de covarianzas con el conjunto completo de datos es:

$$\hat{\Sigma} = \begin{pmatrix} 880{,}66667 & 1100{,}0833 & 659{,}33333 \\ 1100{,}0833 & 2259{,}8333 & 1503{,}8333 \\ 659{,}33333 & 1503{,}8333 & 1169{,}83333 \end{pmatrix}. \tag{$*$}$$

y la estimación de $\beta^{(1)}$, con la ecuación (19), arrojo los siguientes resultados:

$$\hat{\beta}^{(1)} = \begin{pmatrix} 530{,}24167 & -50{,}91667 & 10{,}9250 & -6{,}619444 & -6{,}305556 & 1{,}119444 \end{pmatrix}^t.$$

Finalmente, obtenida la estimación de $\hat{\beta}^{(1)}$, se encontraron los valores de predicción a partir de la ecuación (20); estos resultado se ilustran en la tabla 3.

Tabla 3: Resultados de la predición de la información faltante usando casos completos.

Grupo	Animal	Tiempo	$\hat{y}_{ij}^{(1)}$
1	1	Sem. 3	535.6667
1	3	Sem. 3	535.6667
2	6	Sem. 3	546.6667
2	6	Sem. 4	571.6667

2.2.2. Método alternativo de estimación

En esta sección se propone una variante al método de estimación de información faltante presentado en la sección 2.2.1, el método tiene en cuenta toda la información disponible. En este proceso de estimación, se complementa el modelo (14) incluyendo los individuos que tenían alguna información, esto llevó a plantear el modelo:

$$y = X\beta^{(2)} + e, \tag{22}$$

con $E(y) = X\beta^{(2)}$ y $E\left(ee^t\right) = \Omega$.

En (22) se satisface que $\begin{pmatrix} y_F \\ y_C \end{pmatrix}, X = \begin{pmatrix} X_F \\ X_C \end{pmatrix}$ con y_F vector respuesta asociado con los individuos observados parcialmente, X_F matriz diseño de los mismos individuos observados parcialmente, y_C y X_C como se definieron en la sección 2.2.1, esto es, $X_C = X_B \otimes X_W$.

La matriz Ω se particiona como,

$$\Omega = \begin{pmatrix} \Omega_F & \Omega_{FC} \\ \Omega_{CF} & \Omega_C \end{pmatrix},$$

donde $\Omega_F = Cov\left(y_F\right)$ y Ω_C como se definió en la sección 2.2.1. Para efectos de este trabajo se asume independencia entre y_F y y_C; por tanto se tiene que $\Omega_{FC} = \Omega_{CF} = 0$. El estimador de mínimos cuadrados generalizados para $\beta^{(2)}$ en el modelo (22) es:

$$\hat{\beta}^{(2)} = \left(X^t\hat{\Omega}^{-1}X\right)\left(X^t\hat{\Omega}^{-1}y\right);$$

$$\hat{\beta}^{(2)} = \left(X_F^t \hat{\Omega}_F^{-1} X_F + \left(X_B^t X_B \otimes X_W^t \hat{\Sigma}^{-1} X_W\right)\right)^{-1} \left(X_F^t \hat{\Omega}_F^{-1} y_F + \left(X_B^t \otimes X_W^t \hat{\Sigma}^{-1}\right) y_C\right). \tag{23}$$

Los resultados (24), (25) y (26) son de Henderson y Searle [8]:

$$(A + UBV)^{-1} = A^{-1} - A^{-1}\left(I + UBVA^{-1}\right)^{-1} UBVA^{-1}, \tag{24}$$

para A matriz no singular, U, B y V matrices rectangulares o cuadradas;

$$(I + P)^{-1} = I - P\,(I + P)^{-1} = I - (I + P)^{-1} P, \tag{25}$$

con $I + P$ no singular e I matriz idéntica;

$$(I + PQ)^{-1} P = P\,(I + QP)^{-1}, \tag{26}$$

con $I + PQ$ y $I + QP$ no singulares.

Se puede reescribir (24) como: $U = X_F^t \hat{\Omega}_F^{-1} X_F$; $B = V = I$ y $A = X_C^t \hat{\Omega}^{-1} X_C = \left(X_B^t X_B \otimes X_W^t \hat{\Sigma}^{-1} X_W\right)$ y por (25) y (26) entonces (23) es estimado como:

$$\begin{aligned}
\hat{\beta}^{(2)} = &\left\{\left(X_B^t X_B \otimes X_W^t \hat{\Sigma}^{-1} X_W\right)^{-1} - \left(X_B^t X_B \otimes X_W^t \hat{\Sigma}^{-1} X_W\right)^{-1}\right. \\
&\left(I + \left(X_F^t \hat{\Omega}_F^{-1} X_F\right)\left(X_B^t X_B \otimes X_W^t \hat{\Sigma}^{-1} X_W\right)^{-1}\right)^{-1} \left(X_F^t \hat{\Omega}_F^{-1} X_F\right) \\
&\left.\left(X_B^t X_B \otimes X_W^t \hat{\Sigma}^{-1} X_W\right)^{-1}\right\} \left(X_F^t \hat{\Omega}_F^{-1} y_F + \left(X_B^t \otimes X_W^t \hat{\Sigma}^{-1}\right) y_C\right) \\
= &\left\{\left(\left(X_B^t X_B\right)^{-1} \otimes X_W^{-1} \hat{\Sigma} \left(X_W^{-1}\right)^t\right) - \left(\left(X_B^t X_B\right)^{-1} \otimes X_W^{-1} \hat{\Sigma} \left(X_W^{-1}\right)^t\right)\right. \\
&\left(I + \left(X_F^t \hat{\Omega}_F^{-1} X_F\right)\left(\left(X_B^t X_B\right)^{-1} \otimes X_W^{-1} \hat{\Sigma} \left(X_W^{-1}\right)^t\right)\right)^{-1} \left(X_F^t \hat{\Omega}_F^{-1} X_F\right) \\
&\left.\left(\left(X_B^t X_B\right)^{-1} \otimes X_W^{-1} \hat{\Sigma} \left(X_W^{-1}\right)^t\right)\right\} \left(X_F^t \hat{\Omega}_F^{-1} y_F + \left(X_B^t \otimes X_W^t \hat{\Sigma}^{-1}\right) y_C\right)
\end{aligned} \tag{27}$$

Por facilidad, en (27) se usa (25) y (26), con $Q = M = X_F^t \hat{\Omega}_F^{-1} X_F$ y $P = N = \hat{Cov}\left(\hat{\beta}^{(1)}\right) = \left(X_C^t \hat{\Omega}_C^{-1} X_C\right)^{-1} = \left(X_B^t X_B\right)^{-1} \otimes X_W^{-1} \hat{\Sigma}_C \left(X_W^{-1}\right)^t$

obteniendo entonces:

$$\hat{\beta}^{(2)} = (I + NM)^{-1} N \left(X_F^t \hat{\Omega}_F^{-1} y_F + \left(X_B^t \otimes X_W^t \hat{\Sigma}^{-1}\right) y_C\right). \tag{28}$$

Reemplazando M y N, se tiene finalmente:

$$\hat{\beta}^{(2)} = \left(I + \widehat{Cov}\left(\hat{\beta}^{(1)}\right)\left(X_F^t \hat{\Omega}_F^{-1} X_F\right)\right)^{-1} \left(\hat{\beta}^{(1)} + \widehat{Cov}\left(\hat{\beta}^{(1)}\right)\left(X_F^t \hat{\Omega}_F^{-1} y_F\right)\right). \tag{29}$$

Así, $\hat{\beta}^{(2)}$ se puede expresar usando $\hat{\beta}^{(1)}$ y la varianza de $\hat{\beta}^{(1)}$. Se observa en la expresión anterior que si no hay información faltante, $\hat{\beta}^{(2)}$ es igual a $\hat{\beta}^{(1)}$.

Por otro lado, la varianza de $\hat{\beta}^{(2)}$ es:

$$\widehat{Cov}\left(\hat{\beta}^{(2)}\right) = \left(X^t \hat{\Omega}^{-1} X\right)^{-1} = \left(X_F^t \hat{\Omega}_F^{-1} X_F + X_B^t X_B \otimes X_W^t \hat{\Sigma}^{-1} X_W\right)^{-1},$$

usando (27) y reemplazando a M y N se tiene finalmente que:

$$\widehat{Cov}\left(\hat{\beta}^{(2)}\right) = \left(I + \widehat{Cov}\left(\hat{\beta}^{(1)}\right)\left(X_F^t \hat{\Omega}_F^{-1} X_F\right)\right)^{-1} \widehat{Cov}\left(\hat{\beta}^{(1)}\right) \tag{30}$$

Del resultado anterior, se concluye que la covarianza de $\hat{\beta}^{(2)}$ puede expresarse en términos de la covarianza de $\hat{\beta}^{(1)}$, y si la información esta completa, éstas coinciden.

Con los datos del ejemplo y $\hat{\Sigma}$ obtenida en $(*)$ se encontraron los siguientes valores de estimación para $\beta^{(2)}$:

$$\left(\hat{\beta}^{(2)}\right)^t = \begin{pmatrix} 529{,}7779 & -50{,}4529 & 9{,}3395 & -8{,}1188 & -4{,}8062 & -1{,}2213 \end{pmatrix},$$

En la tabla 4 se muestran los valores inputados por este método.

Tabla 4: Resultados de la predicción de la información faltante usando el método alternativo.

Grupo	Animal	Tiempo	$\hat{y}_{ij}^{(2)}$
1	1	Sem. 3	520.9435
1	3	Sem. 3	520.9435
2	6	Sem. 3	553.8301
2	6	Sem. 4	576.9306

2.2.3. Relación entre sumas de cuadrados del enfoque multivariado-casos completos y el método alternativo.

Finalmente se presenta en esta sección una relación entre las dos propuestas del enfoque multivariado, basada en la comparación de las sumas de cuadrados

del modelo y del error, considerando únicamente la primera iteración, es decir, sin tener en cuenta los datos inputados.

Se inicia con la suma de cuadrados del modelo y a partir de desarrollos algebraicos (véase González L. M. [7]), se encuentra que:

$$\begin{aligned} SCM^{(2)} &= SCM^{(1)} + y_F^t \hat{\Omega}_F^{-1} X_F \hat{Cov}\left(\hat{\beta}^{(2)}\right) X_F^t \hat{\Omega}_F^{-1} y_F \\ &+ \left(\hat{\beta}^{(1)}\right)^t (I + MN)^{-1} \left(2X_F^t \hat{\Omega}_F^{-1} y_F - M\hat{\beta}^{(1)}\right) \end{aligned} \tag{31}$$

Y la suma de cuadrados del error, cuando se usan todos los datos es:

$$\begin{aligned} SCE^{(2)} &= SCE^{(1)} + y_F^t \hat{\Omega}_F^{-1} y_F - 2\left(\hat{\beta}^{(1)}\right)^t X_F^t \hat{\Omega}_F^{-1} y_F \\ &- y_F^t \hat{\Omega}_F^{-1} X_F N X_F^t \hat{\Omega}_F^{-1} y_F + \left(y_F^t \hat{\Omega}_F^{-1} X_F + y_C^t \hat{\Omega}_C^{-1} X_C\right) \\ &NM\hat{Cov}\left(\hat{\beta}^{(2)}\right)\left(X_F^t \hat{\Omega}_F^{-1} y_F + X_C^t \hat{\Omega}_C^{-1} y_C\right) \end{aligned} \tag{32}$$

Las ecuanciones (31) y (32) permiten encontrar una relación entre las sumas de cuadrados de los dos enfoques multivariados, esto es, se expresan las sumas de cuadrados (del modelo y del error) del enfoque multivariado-método alternativo en términos de las sumas de cuadrados del enfoque multivariado-casos completos. De estos resultados se observa que la $SCE^{(2)}$ es igual a $SCE^{(1)}$, siempre que no haya pérdida de información.

3. Conclusiones

En este artículo se llevó a cabo la implementación del método basado en el análisis de covarianza para la estimación de parámetros en medidas repetidas cuando se pierden datos en forma aleatoria encontrando que la estimación del vector de parámetros β no depende de los valores iniciales çonjeturados"para los datos perdidos.

Tanto para el enfoque multivariado conocido en el texto como casos completos, como para el método alternativo, se muestran las expresiones algebraicas que permiten encontrar las predicciones para el vector respuesta, las covarianzas de $\hat{\beta}^{(1)}$ y $\hat{\beta}^{(2)}$, y las expresiones algebraicas para las sumas de cuadrados del modelo y del error, respectivamente.

Finalmente, en la Tabla 5 se comparan los resultados de las predicciones frente a los datos originales, observando que la predicción del método alternativo es la que más se acerca a los datos originales.[3].

[3]Los métodos de imputación se programaron en SAS/IML (véase González, L. M. [7])

Tabla 5: Resultados de la predición de la información faltante usando el método alternativo.

Grupo-Animal-Tiempo	Valores originales	Método de covarianza	Casos completos	Método alternativo
1-1-Sem. 3	460	527.15865	535.6667	520.9435
1-3-Sem. 3	530	527.15865	535.6667	520.9435
2-6-Sem. 3	560	549.82606	546.6667	553.8301
2-6-Sem. 4	565	586.21752	571.6667	576.9306

Bibliografía

[1] Afifi, A. and Elashoff, R. Missing Observations in Multivariate Statistics I: Review of the Literature, *Journal of the American Statistical Association*, 61, 595-604 (1966).

[2] Andrade, D. y Singer, J. Análise de Dados Longitudinais, *VII Simpósio Nacional de Probabilidade e Estatística*, Universidade de Sao Paulo, Brasil (1986).

[3] Andreoni, S. Modelos de Efeitos Aleatórios para Análise de Datos Longitudunais Não Balanceados em Relacão ao Tempo, *Dissertacão Apresentada ao Instituto de Matemática e Estatística da Universidade de São Paulo para Obtencão do Grau de Mestre em Estatística*,São Paulo, Brasil (1989).

[4] Bartlett, M. Some Examples of Statitical Methods of Research in Agriculture, *Journal of the Royal Statistical Society Supplement*, 4, 137-183 (1937). Citado por Affifi y Elashoff (1966.

[5] Crowder, M. y Hand, D. *Analysis of Repeated Measures*, Chapman and Hall (1990).

[6] Dear, R. E. A Principal-Component Missing-Data Method for Multiple Regression Models . SP-86. *System Developed Corporation*, Santa Monica, California (1959). Citado por: Affifi y Elashoff (1966).

[7] González, L. M. Medidas Repetidas con Datos Faltantes: Estimación de Parámetros Vía Análisis de Covarianza, *Tesis de Maestría en Estadística*. Departamento de Estadística. Facultad de Ciencias. Universidad Nacional de Colombia (2002).

[8] Henderson, H. and Searle, S. *On Deriving the Inverse of a Sum of Matrices*. SIAM Review. Society for Industrial and Applied Mathematics. Vol 23 No. 1. 53-60 (1981).

[9] Laird, N., Lange, N. and Stram D. Maximum Likelihood Computations With Repeated Measures: Application of the EM Algorithm. *Journal of the American Statistical Association*, Vo. 82, No. 397 (1987).

[10] Laird, N. and Ware, J. Random-Effects Models for Longitudinal Data. *Biometrics* 38, 963-974 (1982).

[11] Liang, K. and Zeger, S. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73, 1, 13-22 (1986).

[12] *SAS Institute Inc. SAS/STAT User's Guide*, Release 6.03 Edition. Cary, NC: SAS Institute Inc. 1028 pp. (1988).

[13] Searle, S. *Linear Models*, John Wiley and Sons. (1971).

[14] Timm, N. and Mieczkowski, T. *General Linear Models*, SAS. (1997).

[15] Tocher, K. The Design and Analysis of Block Experiments. *Journal of the Royal Statistical Society.* Series B. 14, 45-100 (1952). Citado por Affifi y Elashoff (1966).

[16] Ware, J. Linear Models for the Analysis of Longitudinal Studies". *The American Statistician,* Vol. 39 No. 2 (1985).

[17] Wilkinson, G. Estimation of the Missing Value for the Analysis of Incomplete Data. *Biometrics*, 14, 257-86 (1958). Citado por Affifi y Elashoff (1966).

[18] Yates, F. The Analysis of Replicated Experimental when the Field Results are Incomplete. *The Empire Journal of Experimental Agriculture*, 1, 129-142 (1933). Citado por Affifi y Elashoff (1966).