Revista Colombiana de Estadística
Volumen 29 N° 2. pp. 221 a 238. Diciembre 2006

Imputación de datos en diseños *switchback* usando un modelo mixto con errores correlacionados

Data Imputation in Switchback Designs Using a Mixed Model with Correlated Errors

Luis Fernando Grajales[a], Luis Alberto López[b]

Universidad Nacional de Colombia, Departamento de Estadística, Bogotá

Resumen

Se trata el problema de imputar mediciones individuales en datos provenientes de diseños *switchback* con errores correlacionados, teniendo en cuenta la propuesta de Barroso et al. (1998), donde se considera el BLUP (*Best Linear Unbiased Predictor*) para la imputación de datos. Se hizo uso de los valores propios de las matrices de cuadrados medios de los errores de las predicciones para comparar las estructuras de covarianza $\sigma^2 I$, $AR(1)$ y CS asociadas a los errores. Los resultados sugieren que las dos primeras estructuras son más adecuadas que la tercera.

Palabras clave: datos faltantes, mínimos cuadrados generalizados, BLUP, estructura de covarianza.

Abstract

The problem of predicting individual measurements in switchback designs with correlated errors is considered. The predictions and imputations are done using the BLUP (Best Linear Unbiased Predictions), which have been suggested by Barroso et al. (1998). Three covariance structures were compared by the eigenvalues of the matrices of mean square errors. The results suggest that structures $\sigma^2 I$ and $AR(1)$ are better than CS.

Key words: Missing data, Generalized least squares, Best linear unbiased prediction, Covariance structure.

[a]Profesor Asistente. E-mail: lfgrajalesh@unal.edu.co
[b]Profesor Asociado. E-mail: lalopezp@unal.edu.co

1. Introducción

En muchas situaciones experimentales, la misma unidad experimental recibe dos o más tratamientos en periodos diferentes; cuando esto sucede, el diseño experimental se conoce como *crossover*, el cual puede estudiarse en Jones & Kenward (2003).

Un caso especial, el cual constituye la base de este trabajo, es el diseño *switchback*: en este caso se aplican dos tratamientos a un mismo individuo durante tres periodos; en el primero y el tercero se aplica el mismo tratamiento mientras que en el segundo se aplica otro tratamiento; es decir, se presentan las secuencias ABA o BAB, siendo A y B los tratamientos. Este diseño es uniforme en periodos, pero no en sujetos.

El uso más frecuente de estos diseños se encuentra en los ensayos clínicos (Ebbutt 1984), ensayos de bioequivalencia (Jones & Kenward 2003) y experimentos con ganado lechero (Oman & Seiden 1988, Li 1995, Macchiavelli 1997, Tempelman 2004); la mayoría de los análisis se hacen sobre datos completos.

Matthews (1988) señala que se ha dedicado poca atención en la literatura a los datos faltantes en los diseños *crossover* y aún menos en diseños *switchback*; esto sigue ocurriendo en la literatura posterior consultada. El objetivo del presente trabajo es imputar datos individuales en ensayos provenientes de diseños *switchback*; se proponen dos estrategias de imputación, las cuales son adaptación y modificación de dos propuestas de la literatura: la primera propuesta por Yates (1933), usando regresión lineal simple y efectos fijos; y la segunda por Barroso et al. (1998), usando modelos mixtos y errores correlacionados para tres diferentes estructuras de covarianza.

2. Datos completos en diseño *switchback*

El análisis estadístico de datos completos lo inició Lucas (1956). Inicialmente hizo un ensayo donde comparaba tres dietas A, B, y C, en 12 vacas, agrupadas en tres bloques, uno de tamaño 6 y dos de tamaño 3. La variable respuesta fue la producción de leche (en promedio), corregida por grasa, durante tres periodos de cinco semanas cada uno. Entre cada periodo se dejó un espacio de una semana como tiempo *wash-out*, con lo cual, para efectos del modelo, no necesita tenerse en cuenta el efecto *carry-over* o de rezago. En el momento de analizar los datos, el ensayo no había terminado, por tanto, Lucas simuló los datos para los regímenes de tratamiento ABA, BAB, ACA, CAC, BCB y CBC. Los 36 valores simulados se presentan en la tabla 2.

El análisis de datos de Lucas se basó en el método de *diferencias entre los periodos* para cada vaca: $D = Y_1 - 2Y_2 + Y_3$, siendo Y_i la respuesta (producción) en el i-ésimo periodo, $i = 1, 2, 3$. Con base en los valores de D en los tres bloques obtuvo la tabla ANOVA, considerando un modelo de efectos fijos y errores independientes. Oman & Seiden (1988) presentan la construcción de diseños *switchback* con base en cuadros latinos cíclicos y teoría de grafos; adicionalmente comparan

algunos diseños y sugieren el *switchback* para el caso en que las variaciones de las curvas de respuesta se consideren importantes; reanalizaron los datos de Lucas con un modelo que incluye periodo, tratamiento y añade los efectos producción promedio y curva de lactancia de cada vaca, todos considerados efectos fijos; sin embargo, en las expresiones de sumas de cuadrados dejan abierta la opción de considerar como aleatorio el efecto de curva de lactancia. Los errores se consideraron independientes.

Li (1995) reanalizó los datos de Lucas (1956) aportando la presentación de las expresiones matriciales del ANOVA por medio de IML (*Interactive Matrix Language*) de SAS y la propuesta de considerar matrices de covarianza para los errores.

Macchiavelli (1997) adiciona, al trabajo de Li (1995), la consideración de un modelo mixto con errores correlacionados, teniendo como efectos aleatorios bloque, vaca y pendiente de la curva de lactancia; en la estimación y pruebas de hipótesis considera las estructuras de covarianza $\sigma^2 I$ (*independencia*), $AR(1)$ (*autorregresiva de orden 1*) y CS (*Compound Symmetry*). Utilizó el *Procedure Mixed* de SAS para las estimaciones y los contrastes de hipótesis.

3. Datos faltantes en diseño *switchback*

De acuerdo con la naturaleza del diseño *switchback*, los principales problemas que origina la información faltante son los siguientes: pérdida de ortogonalidad, estimación sesgada y un cálculo inapropiado de las sumas de cuadrados.

El primer problema se debe a que un sujeto con el régimen de tratamiento ABA y un dato faltante ya no podría compararse con otro sujeto con el tratamiento BAB, pues los valores de la respuesta del primero serían de una de las formas $(Y_1, Y_2, \circ)$, $(Y_1, \circ, Y_3)$, $(\circ, Y_2, Y_3)$, donde $\circ$ indica dato faltante.

De otro lado, el sesgo de los análisis de datos incompletos ya ha sido discutido por varios autores; se destaca principalmente a Little & Rubin (2002) y Richardson & Flack (1996); estos últimos, después de analizar solo los casos completos de un diseño *switchback*, encontraron los sesgos de dicho análisis y en la discusión afirman que debería evitarse el análisis de los casos completos en presencia de datos faltantes en los estudios.

Finalmente, cuando faltan datos, las sumas de cuadrados y los estadísticos de prueba tradicionales ya no son adecuados (Little & Rubin 2002).

La primera consideración de datos faltantes en el diseño *switchback* se hace en Lucas (1956); allí propone una fórmula para imputar un dato, pero el proceso de imputación presenta dos situaciones inadecuadas: i) La fórmula presentada no imputa el dato faltante en sí, lo que hace es completar, para el sujeto del dato faltante, el valor de una transformación, y ii) la propuesta se hace únicamente para modelo de efectos fijos y errores independientes. Li (1964) imputó datos de diseños *switchback* usando métodos de cuadros latinos y bloques aleatorizados, pero no consideró efectos aleatorios.

A nivel general, en otros diseños, la imputación de datos usando regresión y ANOVA, la literatura es bastante amplia. Se resaltan los trabajos de Yates (1933), el cual se adapta en este estudio; Carriere (1994) analiza medidas repetidas incompletas, considerando el diseño *crossover* con las secuencias ABB y BAA y pérdidas *dropout* (caso donde la información falta en el último periodo); toma en cuenta únicamente el modelo de efectos fijos. En Barroso et al. (1998) se propone una metodología para la imputación con modelo mixto y errores correlacionados, la cual también se adaptó en el presente trabajo.

Con base en las dos secciones anteriores se tomaron los elementos para proponer la imputación de datos en el diseño *switchback*. Esto se detallará en la sección 5. En la siguiente sección se presentan la notación y el modelo que se utilizarán.

4. Modelo y notación

4.1. El modelo mixto

El modelo propuesto para este trabajo es el siguiente:

$$y_{ijk} = \mu + \pi_i + \tau_j + c_k + (i-2)s_k + \varepsilon_{ijk} \tag{1}$$

donde:

- y_{ijk} hace referencia a la producción de leche de la k-ésima vaca, después de recibir la j-ésima dieta, en el i-ésimo periodo.
- $i, j = 1, 2, 3$ y $k = 1, \ldots, 12$.
- μ es la media global.
- π_i es el efecto del i-ésimo periodo, $i = 1, 2, 3$.
- τ_j es el efecto del j-ésimo tratamiento, $j = 1, 2, 3$.
- c_k y s_k son, respectivamente, el intercepto y la pendiente de la curva de lactancia de la k-ésima vaca, y
- ε_{ijk} es un término de error aleatorio para la k-ésima vaca, después de recibir la j-ésima dieta, en el i-ésimo periodo.

En el modelo (1) los efectos μ, π_i y τ_j se consideran fijos y los efectos c_k y s_k se consideran aleatorios. La escritura $i-2$ garantiza la ortogonalidad entre las filas de la matriz diseño que se asociará más adelante a los efectos aleatorios; ver detalles en Macchiavelli (1997).

Matricialmente, el modelo (1) se escribe como:

$$Y = X_0\beta_0 + Z\gamma + \xi \tag{2}$$

donde:

- Y es el vector de respuesta.
- X_0 es la matriz diseño asociada con los efectos fijos.
- $\beta_0 = [\mu \quad \pi_1 \quad \pi_2 \quad \pi_3 \quad \tau_1 \quad \tau_2 \quad \tau_3]^t$ es un vector desconocido de efectos fijos.
- Z es la matriz diseño asociada a los efectos aleatorios.
- γ un vector, desconocido, de efectos aleatorios, y
- ξ es el vector de errores aleatorios no observados.

Por facilidad en el análisis, el modelo (2) se reparametrizó en el modelo

$$Y = X\beta + Z\gamma + \xi \tag{3}$$

La matriz X va a ser de tamaño 36×5, con

$$\beta^t = \begin{bmatrix} \mu & \phi_1 & \phi_2 & \theta_1 & \theta_2 \end{bmatrix}$$

donde los nuevos parámetros son: $\phi_1 = \pi_1 - \pi_2$ $\phi_2 = \pi_2 - \pi_3$ $\theta_1 = \tau_1 - \tau_2$ $\theta_2 = \tau_2 - \tau_3$.

En el modelo (3) se supone que:

a) Los vectores γ y ξ siguen una distribución normal.

b) El vector de respuestas Y tiene distribución normal multivariada con vector de medias y matriz de covarianzas dados por las expresiones:

$$E[Y] = X\beta \qquad \text{y} \qquad Var[Y] = V = ZGZ^t + R \tag{4}$$

donde G y R son matrices no singulares, simétricas y desconocidas.

El tercer supuesto es que, conjuntamente, γ y ξ tienen distribución normal, con vector de medias y matriz de covarianzas dados por:

$$E\begin{bmatrix}\gamma \\ \varepsilon\end{bmatrix} = \begin{bmatrix}0 \\ 0\end{bmatrix} \quad \text{y} \quad Var\begin{bmatrix}\gamma \\ \varepsilon\end{bmatrix} = \begin{bmatrix}G & 0 \\ 0 & R\end{bmatrix} \tag{5}$$

Como se observa en el modelo (4), la matriz V se puede modelar tomando la matriz diseño Z de efectos aleatorios y especificando las estructuras de covarianza para las matrices G y R. Teniendo en cuenta (4), la matriz V, para cada vaca, cumple:

$$Var\begin{bmatrix} Y_{1jk} & Y_{2jk} & Y_{3jk} \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} G_k \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \end{bmatrix} + R_k$$

donde:

$$Z_k^t = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \end{bmatrix} \quad \text{y} \quad G_k = \begin{bmatrix} \sigma_c^2 & \sigma_{cs} \\ \sigma_{cs} & \sigma_s^2 \end{bmatrix}$$

Los componentes de la matriz no-estructurada G_k son: σ_c^2, varianza del intercepto; σ_s^2, varianza de la pendiente de la curva de lactancia, y σ_{cs}, covarianza entre el intercepto y la pendiente.

Finalmente, en (4), para la matriz de errores, R_k, se consideraron en este trabajo las siguientes estructuras, para cada vaca:

$$\sigma_\varepsilon^2 I = \begin{bmatrix} \sigma_\varepsilon^2 & 0 & 0 \\ 0 & \sigma_\varepsilon^2 & 0 \\ 0 & 0 & \sigma_\varepsilon^2 \end{bmatrix} \qquad AR(1) = \sigma_\varepsilon^2 \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix} \qquad CS = \sigma_\varepsilon^2 \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

4.2. Ecuaciones de estimación

Dado que una forma de imputación usada en este trabajo se hace a través del BLUP, es necesario caracterizar las ecuaciones de estimación tanto de la parte fija como de la aleatoria. Si se tiene en cuenta el modelo (3), bajo el supuesto de normalidad, la función de densidad conjunta $f(y, \beta, \gamma)$ es:

$$(2\pi)^{-\frac{N+q}{2}}(|R||G|)^{-\frac{1}{2}} \times \exp\Big[-\frac{1}{2}(y - X\beta - Z\gamma)^t R^{-1}(y - X\beta - Z\gamma) - \frac{1}{2}\gamma^t G^{-1}\gamma \Big]$$

Al derivar parcialmente el logaritmo de la función de verosimilitud con respecto a β y γ e igualar a cero se obtienen las dos soluciones:

$$\widehat{\beta} = \left(X^t V^{-1} X\right)^{-1} X^t V^{-1}$$
$$\widehat{\gamma} = \left(Z^t R^{-1} Z + G^{-1}\right)^{-1} \left[Z^t R^{-1} - Z^t R^{-1} X (X^t V^{-1} X)^{-1} X^t V^{-1}\right]$$

Generalmente no se conocen G ni R y deben estimarse a partir de los datos. En este trabajo se escogen matrices de acuerdo con la literatura para poder comparar los resultados: para G, una matriz no-estructurada y, para la matriz R, estructuras $\sigma^2 I$, $AR(1)$ y CS. Sin embargo, existen muchos textos y métodos en la literatura para encontrar estas estimaciones; se destacan principalmente los métodos MIVQUE0, ML y REML (*Restricted Estimation Maximum Likelihood*), los cuales pueden estudiarse en Searle et al. (1992).

En el procedimiento *Mixed* de SAS (2005) se usan los valores iniciales de MIVQUE0 para conseguir las estimaciones del REML. El proceso de estimación depende del conocimiento que se tenga de la estructura de la matriz de covarianza; en Wolfinger (1993) se presentan diferentes estrategias para tomar una decisión sobre la estructura más adecuada; en la comparación y elección de estas estructuras se puede hacer uso del criterio de información de Akaike (AIC), el criterio bayesiano de Schwarz (BIC) y el criterio basado en el error cuadrado medio (ECM) de los BLUP. En el presente trabajo se utilizó este último.

En la sección siguiente se presenta la metodología utilizada en el trabajo que consiste en simulación de datos, posteriormente la eliminación, para finalizar con la imputación de datos.

5. Metodología

Muchos detalles de la metodología completa del trabajo no pueden darse acá. El trabajo de simulación, eliminación de datos e imputación se realizó con el paquete estadístico SAS; los detalles de programación en IML y *Mixed* pueden hallarse en Grajales (2006); el proceso se presenta a continuación.

5.1. Proceso de simulación

En el proceso de simulación se utilizaron algunos valores de medias y varianzas hallados en la literatura de referencia, especialmente los trabajos de Oman & Seiden (1988) y Macchiavelli (1997). En la simulación se llevaron a cabo los siguientes pasos:

1. Se generó un vector Y, de tamaño 36×1, el cual se supone que se distribuye normal con vector de medias $[\mu_1 \quad \mu_2 \quad \mu_3] = [40 \quad 35 \quad 30]^t$ y matriz de covarianzas

$$Var(Y) = V_{36\times 36} = R_{36\times 36} + Z_{36\times 72}\, G_{72\times 72}\, Z^t_{72\times 36}$$

donde: $Z = [Z_1 \; Z_2]$ con:

$$Z_1 = I_{12} \bigotimes \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad \text{y} \quad Z_2 = I_{12} \bigotimes \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}$$

$$G = \begin{bmatrix} G_C & G_{CS} \\ G_{CS} & G_S \end{bmatrix} \quad \text{con} \quad \begin{cases} G_C = 58\, I_{12} \bigotimes I_3 \\ G_{CS} = (-0.6)\, I_{12} \bigotimes I_3 \\ G_S = (0.9)\, I_{12} \bigotimes I_3 \end{cases}$$

$\bigotimes$ indica el producto Kronecker.

Esto se realizó para las tres matrices de correlación de errores: $R = \sigma^2\, I_{36}$, $AR(1)$ y CS.

2. Se simularon los 108 datos (36 valores para cada tipo de error), los cuales se muestran en la tabla 3.

Los datos de Lucas (1956) se usan como ejemplo para la aplicación.

5.2. Eliminación *MCAR* de datos

Para cada estructura de covarianza de los errores, con los 36 datos simulados, se eliminó un dato a la vez, de la variable respuesta. De esta forma se obtuvieron 36 bases de datos, de 35 individuos cada una. Acá se considera el mecanismo MCAR (*Missing completely at random*) para eliminación de datos. La idea básica del mecanismo es que el dato faltante no depende de los valores presentes de la variable respuesta; véase Little & Rubin (2002).

5.3. Método de imputación de Yates

El método considera que si falta una observación en la variable respuesta, dígase y_k, entonces se debe: a) estimar $\widehat{\beta}_*$ en la información completa, y b) imputar el dato faltante mediante la expresión $\widehat{y}_k = X_k\widehat{\beta}_*$.

La justificación del procedimiento se basa en dos ventajas: i) produce estimaciones *correctas* de β, vía mínimos cuadrados, y ii) se logra una estimación *correcta* de la suma de cuadrados de los errores.

Además, aunque el método de Yates produce estimaciones sesgadas de la matriz de covarianza de $\widehat{\beta}$, los sesgos son relativamente menores cuando se tiene una fracción pequeña de datos faltantes; véase Little & Rubin (2002). Como en este trabajo se asume que hay alguna estructura de correlación de los errores, debe modificarse el método propuesto por Yates, teniendo en cuenta las matrices $AR(1)$ y CS. La imputación con este método se realizó asumiendo que la curva de lactancia de cada vaca se ajusta mediante una línea recta; esto se justifica porque las vacas ingresan al estudio después del pico máximo de la curva, como se observa en la figura 1. El modelo ajustado fue entonces $\widehat{y}_i = b_0 + b_1 x_i$, siendo x_i el tiempo y y_i la producción de leche.

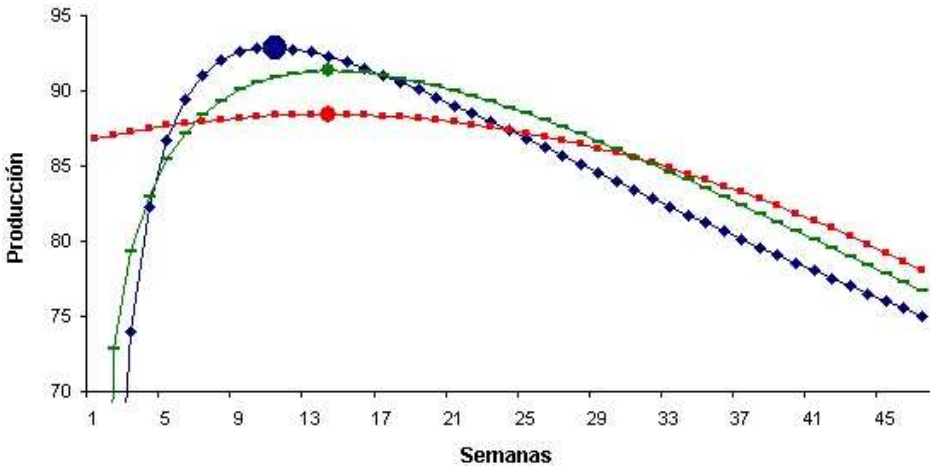

FIGURA 1: Ejemplos de curvas de lactancia con pico máximo.

Como ilustración del procedimiento, si suponemos que faltó el dato de la semana 5, entonces:

$$\text{i) } \widehat{\beta}_5 = \begin{bmatrix} \widehat{\beta}_{0,5} \\ \widehat{\beta}_{1,5} \end{bmatrix} = (X_5^t V_5^{-1} X_5)^{-1} X_5^t V_5^{-1} \begin{bmatrix} Y_{10} \\ Y_{15} \end{bmatrix},$$

$$\text{con } X_5 = \begin{bmatrix} 1 & 10 \\ 1 & 15 \end{bmatrix} \text{ y } V_5 = 1.2 * \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}$$

$$\text{ii) } Y_{5,pred} = \begin{bmatrix} 1 & 5 \end{bmatrix} \begin{bmatrix} \widehat{\beta}_{0,5} \\ \widehat{\beta}_{1,5} \end{bmatrix}$$

Con errores CS, la cual coincide con la matriz V_5 para los errores $AR(1)$. Se usaron $\sigma_\varepsilon^2 = 1.2$ y $\rho = 0.2$; estos valores se toman con base en el trabajo de Macchiavelli (1997). Las imputaciones con el método de Yates se presentan en la tabla 1; estas imputaciones coinciden, para los tres tipos de errores, porque las tres matrices de covarianzas coinciden debido a que las matrices X y V son 2×2 e invertibles; por tanto, se cumple la siguiente relación matricial, cuyo resultado final no depende de V:

$$\left(X^t V^{-1} X\right)^{-1} X^t V^{-1} y = X^{-1} V \left(X^t\right)^{-1} X^t V^{-1} y = X^{-1} I y = X^{-1} y$$

Finalmente, puede verse que el método de Yates solo tiene en cuenta la pendiente y el intercepto de la curva de lactancia, ambos como efectos fijos.

5.4. Método de imputación de Barroso et al. (1998)

Este método considera la imputación de datos en modelos mixtos, cuando se presenta algún tipo de correlación de los errores.

En el proceso de imputación, cuando se tienen m datos faltantes, el método es el siguiente:

1. Se parte de un modelo como el (3), con los siguientes órdenes de vectores y matrices:
   - y un vector de orden $d \times 1$,
   - β un vector de tamaño $p \times 1$, asociado a los efectos fijos,
   - γ es un vector de tamaño $q \times 1$, asociado a los efectos aleatorios,
   - X y Z matrices conocidas, y
   - ε es el vector de errores, de orden $d \times 1$.
2. Las matrices G y R son definidas positivas, de rango completo, conocidas, con dimensiones $q \times q$ y $d \times d$, respectivamente.

En el arreglo del ensayo utilizado en este trabajo, se tiene que: $d = 36$; $p = 5$; $q = 72$ y $m = 1$.

Adicionalmente se impone la condición que los errores correspondientes a las mediciones observadas y faltantes no tienen correlación; supuesto que se cumple en las condiciones del modelo del presente trabajo. El método de Barroso et al. (1998) supone que si m de las d mediciones son faltantes, sin pérdida de generalidad, el modelo se puede con:

$$y = \begin{bmatrix} y_0 \\ y_m \end{bmatrix} = \begin{bmatrix} X_0 \\ X_m \end{bmatrix} \beta + \begin{bmatrix} Z_0 \\ Z_m \end{bmatrix} \gamma + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_m \end{bmatrix} \tag{6}$$

La expresión (6) es similar a la que se presenta en el método de análisis de covarianza de Bartlett (Little & Rubin 2002); sin embargo, la diferencia conceptual consiste en que la matriz Z particionada en (6) indica la matriz diseño de los efectos aleatorios del modelo mixto, mientras que en el método de Bartlett, la matriz Z representa la matriz diseño de los datos faltantes. Con la notación usada en (6) se encuentra el BLUP:

1. Haciendo una partición de la matriz de covarianzas del vector de efectos aleatorios:
$$Var \begin{bmatrix} \gamma \\ \varepsilon_0 \\ \varepsilon_m \end{bmatrix} = \begin{bmatrix} G & 0 & 0 \\ 0 & R_0 & 0 \\ 0 & 0 & R_m \end{bmatrix} \sigma^2$$
donde los subíndices en ε y R hacen referencia a los datos observados y faltantes, respectivamente.

2. Considerando la relación
$$ECM \begin{bmatrix} \widehat{\gamma} \\ \widehat{\varepsilon} \end{bmatrix} = \begin{bmatrix} \widehat{\gamma} \\ \widehat{\varepsilon} \end{bmatrix}^t \left[Var \begin{bmatrix} \widehat{\gamma} \\ \widehat{\varepsilon} \end{bmatrix} \right]^{-1} \begin{bmatrix} \widehat{\gamma} \\ \widehat{\varepsilon} \end{bmatrix}$$

Con estas dos relaciones, el BLUP se obtiene minimizando la forma cuadrática:

$$H = \begin{bmatrix} \gamma \\ y_0 - X_0\beta - Z_0\gamma \\ y_m - X_m\beta - Z_m\gamma \end{bmatrix}^t \begin{bmatrix} G^{-1} & 0 & 0 \\ 0 & R_0^{-1} & 0 \\ 0 & 0 & R_m^{-1} \end{bmatrix} \begin{bmatrix} \gamma \\ y_0 - X_0\beta - Z_0\gamma \\ y_m - X_m\beta - Z_m\gamma \end{bmatrix} \tag{7}$$

La expresión (7) se simplifica al tomar la igualdad $y_m = X_m\beta - Z_m\gamma$ y, por tanto, la minimización de H sobre β, γ y y_m se logra al tomar los datos observados y ajustarles el modelo para conseguir los BLUP de β y γ.

De esta forma se imputan los valores en y_m reemplazando las cantidades desconocidas en $y_m = X_m\beta - Z_m\gamma + \varepsilon_m$ por el respectivo BLUP.

Las estimaciones, de acuerdo con Barroso et al. (1998), son $\widehat{y}_m$, $\widehat{\beta}$ y $\widehat{\gamma}$, las cuales vienen dadas por:

$$\widehat{y}_m = \Pi \times \Theta$$

donde:

$$\Pi = \left\{I_m - X_m\left(X^tW^{-1}X\right)^{-1}\left[X_m^t - X^tR^{-1}ZAZ^t\right]R_m^{-1} - Z_mA\left[Z_m^t - Z^tR^{-1}X\left(X^tW^{-1}X\right)^{-1}\left(X_m^t - X^tR^{-1}ZAZ_m^t\right)\right]R_m^{-1}\right\}^{-1}$$

$$\Theta = \left\{X_m\left(X^tW^{-1}X\right)^{-1}\left[X_0^t - X^tR^{-1}ZAZ_0^t\right] + Z_mA\left[Z_0^t - Z^tR^{-1}X\left(X^tW^{-1}X\right)^{-1}\left(X_0^t - X^tR^{-1}ZAZ_0^t\right)\right]\right\}R_0^{-1}y_0$$

$$\widehat{\beta} = \left(X^tW^{-1}X\right)^{-1}X^tW^{-1}\widehat{y} \quad \text{y}$$

$$\widehat{\gamma} = \left(Z^tR^{-1}Z + G^{-1}\right)^{-1}\left(Z^tR^{-1} - Z^tR^{-1}X\left(X^tW^{-1}X\right)^{-1}X^tW^{-1}\right)\widehat{y}$$

con:

$$A = (Z^tR^{-1}Z)^{-1} \qquad W = \frac{1}{\sigma^2}V \qquad \text{y} \qquad \widehat{y} = \begin{bmatrix} y_0 \\ \widehat{y}_m \end{bmatrix}$$

La imputación de datos se realiza con la fórmula:

$$\widehat{y}_m = X_m\widehat{\beta} + Z_m\widehat{\gamma} \tag{8}$$

Estos estimadores también pueden hallarse de manera simplificada, usando matrices indicadoras de valores observados y valores faltantes. Dicha simplificación se utilizó en los análisis del presente trabajo. La simplificación parte de la siguiente notación: E es la matriz indicadora de valores faltantes, tiene dimensión $d \times m$, donde cada columna corresponde a uno de los valores faltantes. Cada columna de E tiene $d-1$ ceros y un 1, ubicado en la fila correspondiente al valor faltante.
F es la matriz indicadora de los valores observados, con dimensión $d \times (d-m)$ y es análoga a E pero indicando los valores observados. Con la notación usada para $\widehat{y}_m$, $\widehat{\beta}$ y $\widehat{\gamma}$ se tienen los siguientes resultados:

$$\begin{aligned} y_0 &= F^ty & y_m &= E^ty \\ X_0 &= F^tX & X_m &= E^tX \\ Z_0 &= F^tZ & Z_m &= E^tZ \\ R_0^{-1} &= F^tR^{-1}F & R_m^{-1} &= E^tR^{-1}E \\ E^tE &= I_m & F^tF &= I_{d-m} \\ E^tF &= 0 & EE^t + FF^t &= I_d \\ FF^tR^{-1}FF^t + EE^tR^{-1}EE^t &= R^{-1} \end{aligned}$$

Con la notación simplificada de las matrices E y F, los errores cuadrados medios de $\widehat{y}_m$, $\widehat{\beta}$ y $\widehat{\gamma}$ son:

$$ECM(\widehat{y_m}) = \Delta \times \Upsilon$$

donde:

$$\begin{aligned}
\Delta &= \left[J^{-1}E^tQFF^tR^{-1}FF^t - E^t\right]V \\
\Upsilon &= \left[FF^tR^{-1}FF^tQE\left(I_m - E^tR^{-1}EE^tQE\right)^{-1} - E\right]\sigma^2 \\
ECM(\widehat{\beta}) &= BVB^t\sigma^2 \\
B &= \left(X^tV^{-1}X\right)^{-1}X^tV^{-1}R\left[I_d + EE^tR^{-1}EJ^{-1}E^tQ\right]FF^tR^{-1}FF^t \\
ECM(\widehat{\gamma}) &= \left(CVC^t - CZG - GZ^tC^t + G\right)\sigma^2
\end{aligned}$$

con

$$\begin{aligned}
J &= I_m - E^tQEE^tR^{-1}E \\
C &= AZ^t\left[I_d - R^{-1}X\left(X^tV^{-1}X\right)^{-1}X^tV^{-1}R\right]\left[I_d + EE^tR^{-1}EJ^{-1}E^tQ\right]FF^tR^{-1}FF^t
\end{aligned}$$

Para la imputación las matrices E y F utilizadas para este trabajo fueron las siguientes:

$$E = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}_{36\times 1} = \overrightarrow{e}_{36} \qquad F = \begin{bmatrix} I_{35} \\ \cdots \\ 0_{1\times 35} \end{bmatrix}$$

Es decir, de acuerdo con la notación del trabajo de Barroso et al. (1998), $m = 1$, la matriz E tiene tamaño 36×1 y la matriz F tiene tamaño 36×35. Se utilizaron estas dos matrices ubicando el valor faltante en la última fila. Con estas matrices pueden verificarse las propiedades de simplificación mencionadas en dicho trabajo.

5.5. Método de comparación de BLUP

Con el fin de comparar los BLUP, para cada estructura de covarianzas, se procedió de la siguiente manera, con cada dato faltante:

1. Se calcularon los BLUP de efectos fijos.

2. Se calcularon las matrices 5×5 del ECM.

3. Para cada estructura de covarianza, se calcularon las tres posibles diferencias de matrices de ECM.

4. Se hallaron los valores propios de las diferencias de matrices calculadas en 3. y se determinó, con el mínimo y máximo valor propio de las matrices, si ellas eran definidas no-negativas.

Estos pasos también se realizaron para los BLUP de efectos aleatorios (matrices 72×72).

6. Resultados y discusión

Esta sección presenta los resultados básicos con su respectiva interpretación y discusión. En la parte final se presentan consideraciones finales, a modo de conclusiones y temas a seguir.

1. **Imputaciones de los datos de Lucas (1956)**

   La última columna de la tabla 1 presenta los resultados de los 36 valores imputados con el método de Yates para los datos simulados por Lucas. En la sección 5 se mencionó la razón por la cual coinciden las imputaciones para las tres estructuras de covarianza de los errores.

   De otro lado, la tabla 2 presenta los valores imputados con el método de Barroso et al. (1998) para los mismos datos de Lucas (1956).

   Llama la atención el hecho de que las imputaciones de Yates hayan sido más *cercanas* que las de Barroso et al. (1998). Una razón para esto es que en el método de Yates se están tomando como fijos los efectos intercepto y pendiente de la curva de lactancia y la estimación mínimo cuadrática es *insesgada* $E(\widehat{\beta}) = \beta$ y es el único parámetro a estimar; en cambio, en Barroso et al. (1998) estos efectos son aleatorios y aunque las predicciones son *insesgadas*: $E(\widehat{\beta}) = E(\beta)$, es necesario estimar otros efectos. Sin embargo, aunque sale favorecido en estos datos, el método de Yates tiene la limitante de admitir solo efectos fijos. En este punto queda abierta la discusión de la comparación de estos dos métodos en datos con otras magnitudes o con otras estructuras y estimaciones de los efectos aleatorios.

2. **ECM en los datos de Lucas**

   El método de Barroso-Bussab-Knott permite el cálculo del ECM del valor imputado. La parte derecha de la tabla 2 presenta los ECM del valor imputado para los datos de Lucas. Allí puede verse que, en general, la imputación con errores independientes produce un ECM menor que la imputación con las estructuras $AR(1)$ y CS. Este resultado no sorprende teniendo en cuenta que la idea original de los datos de Lucas consideraba errores independientes. La ventaja de la estructura $\sigma^2 I$ sobre $AR(1)$ también fue hallada por Macchiavelli (1997) para datos completos y con prueba de razón de verosimilitud. La ventaja de la estructura $\sigma^2 I$ sobre CS es opuesta a la hallada por Li (1995), aunque allí lo hace con datos completos y usando un modelo que incluye bloque y la interacción *periodo* $\times$ *tratamiento*. La ventaja de $AR(1)$ sobre CS, aunque no se tienen resultados en la literatura para comparar, podría explicarse en la diferencia entre $\rho^2 = 0.02$ para $AR(1)$ y $\rho = 0.4$ para CS; estos son los valores usados para la correlación de errores de los tiempos 1 y 3. Como puede verse, la correlación más grande afecta negativamente la imputación con los errores CS. En la tabla 2 se tiene una razón más para pensar que la estructura CS tiene menor ECM que la estructura $AR(1)$ cuando se imputa el dato del periodo 2; sin embargo, esto queda para explorar.

Tabla 1: Datos originales de Lucas (1956) e imputación con el método de Yates.

Vaca	Periodo	Dieta	Leche	Imputación
1	1	1	34.6	36.10
1	2	2	32.3	31.55
1	3	1	28.5	30.00
10	1	1	38.7	40.40
10	2	2	37.4	36.55
10	3	1	34.4	36.10
2	1	2	22.8	23.40
2	2	3	21.0	20.70
2	3	2	18.6	19.20
11	1	2	25.7	28.80
11	2	3	26.1	24.55
11	3	2	23.4	26.50
3	1	3	32.9	38.70
3	2	1	33.1	30.20
3	3	3	27.5	33.30
12	1	3	21.4	24.60
12	2	1	22.0	20.40
12	3	3	19.4	22.60
4	1	1	48.9	51.80
4	2	3	46.9	45.45
4	3	1	42.0	44.90
7	1	1	30.4	32.30
7	2	3	29.5	28.55
7	3	1	26.7	28.60
5	1	2	21.8	26.10
5	2	1	23.9	21.75
5	3	2	21.7	26.00
8	1	2	35.2	32.20
8	2	1	33.5	31.80
8	3	2	28.4	31.80
6	1	3	25.4	28.10
6	2	2	26.0	24.65
6	3	3	23.9	26.60
9	1	3	30.8	32.20
9	2	2	29.3	28.60
9	3	3	26.4	27.80

3. Imputaciones de los datos simulados

La tabla 3 presenta los datos simulados en el presente trabajo, con su respectiva imputación (Barroso et al. 1998) y el ECM del dato imputado. Los resultados de ECM son similares a los hallados en los datos de Lucas y tienen la misma interpretación y discusión anterior. Debe hacerse una discusión adicional, teniendo el hecho de que los ECM sean tan *grandes*, en especial para la estructura CS; como estos son ECM de datos imputados, teóricamente no pueden compararse con el ECM (único) de los datos completos. Una forma de explorar esto sería por medio de las covarianzas estimadas de la matriz G_k con cada dato faltante. Como se tendrían 36 valores para cada parámetro,

sería necesario buscar una medida resumen para que haya un único valor de comparación entre las estructuras de covarianzas.

4. **Comparaciones de los BLUP**

   Al hallar los valores propios de las matrices correspondientes a las predicciones:

   Para efectos aleatorios. Al hallar los valores propios, no se puede decir que el BLUP de una estructura de covarianza de los errores sea mejor que el BLUP de otra estructura.

Tabla 2: Datos de Lucas e imputación con el método de Barroso et al.

Y-Lucas	$Yimp$ $\sigma^2 I$	$Yimp$ $AR(1)$	$Yimp$ CS	ECM $\sigma^2 I$	ECM $AR(1)$	ECM CS
34.6	44.53	50.80	58.01	0.5	15.6	73.1
32.3	45.29	69.50	64.20	0.6	119.5	112.3
28.5	41.18	47.71	56.02	0.5	15.6	73.1
38.7	52.27	59.58	68.19	0.5	15.6	73.1
37.4	52.22	80.26	74.10	0.6	119.5	112.3
34.4	47.20	54.66	64.05	0.5	15.6	73.1
22.8	29.96	34.14	38.75	0.5	15.6	73.1
21.0	31.31	47.32	43.71	0.6	119.5	112.3
18.6	24.80	29.13	34.50	0.5	15.6	73.1
25.7	37.08	42.22	48.07	0.5	15.6	73.1
26.1	36.11	55.02	50.81	0.6	119.5	112.3
23.4	30.05	35.28	41.54	0.5	15.6	73.1
32.9	43.85	50.27	57.29	0.5	15.6	73.1
33.1	44.45	67.71	62.53	0.6	119.5	112.3
27.5	39.82	46.47	54.47	0.5	15.6	73.1
21.4	31.18	35.58	40.37	0.5	15.6	73.1
22.0	31.37	47.18	43.59	0.6	119.5	112.3
19.4	23.86	28.33	33.49	0.5	15.6	73.1
48.9	63.58	72.65	83.29	0.5	15.6	73.1
46.9	64.33	99.15	91.53	0.6	119.5	112.3
42.0	61.00	70.37	82.25	0.5	15.6	73.1
30.4	41.79	47.55	54.26	0.5	15.6	73.1
29.5	40.97	62.88	58.06	0.6	119.5	112.3
26.7	36.12	42.02	49.40	0.5	15.6	73.1
21.8	34.38	39.15	44.50	0.5	15.6	73.1
23.9	32.96	49.78	45.99	0.6	119.5	112.3
21.7	25.21	29.98	35.35	0.5	15.6	73.1
35.2	44.42	50.96	58.16	0.5	15.6	73.1
33.5	47.21	71.72	66.24	0.6	119.5	112.3
28.4	41.76	48.59	57.00	0.5	15.6	73.1
25.4	37.42	42.57	48.49	0.5	15.6	73.1
26.0	36.42	55.41	51.18	0.6	119.5	112.3
23.9	29.61	34.82	41.01	0.5	15.6	73.1
30.8	40.87	46.66	53.23	0.5	15.6	73.1
29.3	42.23	64.27	59.36	0.6	119.5	112.3
26.4	35.88	41.84	49.21	0.5	15.6	73.1

$Yimp$: valor imputado.
ECM: Error cuadrado medio de imputación.

Para efectos fijos. En algunos datos faltantes, el BLUP ($AR(1)$) es mejor que el BLUP (CS) (ver tabla 3).
En términos de ECM, no hay forma de comparar este resultado con la literatura consultada. En efectos fijos, no se puede concluir sobre las otras dos estructuras de correlación de errores.

Tabla 3: Datos simulados e imputados con tres estructuras de covarianza de errores.

Original $\sigma^2 I$	Imputado	ECM	Original $AR(1)$	Imputado	ECM	Original CS	Imputado	ECM
39.64	39.80	2.3	39.64	53.20	84.9	39.64	68.30	369.2
34.54	33.73	2.7	34.55	69.04	649.2	34.55	67.95	567.4*
29.32	29.88	2.3	29.33	41.77	84.9	29.34	57.06	369.2
38.12	39.28	2.3	38.12	52.43	84.9	38.12	67.21	369.2
33.42	33.01	2.7	33.40	67.48	649.2	33.40	66.37	567.4*
29.20	28.39	2.3	29.17	39.77	84.9	29.15	54.44	369.2
27.12	30.87	2.3	27.12	40.90	84.9	27.12	51.90	369.2
24.55	23.98	2.7	24.42	48.81	649.2	24.42	47.92	567.4*
21.11	17.89	2.3	21.13	25.61	84.9	21.01	35.70	369.2
41.49	39.87	2.3	41.49	53.51	84.9	41.49	68.81	369.2
34.43	36.58	2.7	34.61	73.60	649.2	34.60	72.63	567.4*
30.60	30.38	2.3	30.45	42.71	84.9	30.59	58.34	369.2
34.54	37.30	2.3	34.54	49.97	84.9	34.54	63.71	369.2
32.37	31.13	2.7	32.16	63.07	649.2	32.16	61.91	567.4*
27.06	25.38	2.3	27.20	35.89	84.9	27.02	49.22	369.2
48.16	47.35	2.3	48.16	63.89	84.9	48.16	82.20	369.2
43.08	43.50	2.7	43.03	87.70	649.2	43.03	86.31	567.4*
37.37	37.80	2.3	37.40	52.85	84.9	37.35	71.65	369.2
47.22	47.33	2.3	47.22	63.53	84.9	47.22	81.76	369.2
42.40	40.94	2.7	42.33	83.89	649.2	42.33	82.46	567.4*
36.56	37.68	2.3	36.61	52.35	84.9	36.55	71.01	369.2
45.27	43.83	2.3	45.27	58.78	84.9	45.27	75.72	369.2
38.34	39.28	2.7	38.48	79.84	649.2	38.48	78.69	567.4*
34.00	34.61	2.3	33.87	48.30	84.9	33.98	65.71	369.2
52.47	51.36	2.3	52.47	69.51	84.9	52.47	89.45	369.2
48.56	46.22	2.7	48.37	94.01	649.2	48.37	92.41	567.4*
39.96	43.03	2.3	40.21	59.80	84.9	40.07	80.77	369.2
26.82	26.42	2.3	26.82	26.82	84.9	26.82	35.31	369.2
21.43	22.88	2.7	21.56	45.55	649.2	21.56	45.06	567.4*
16.98	16.18	2.3	16.94	23.48	84.9	17.08	32.89	369.2
43.10	41.60	2.3	43.10	55.88	84.9	43.10	71.89	369.2
36.30	38.19	2.7	36.44	76.91	649.2	36.44	75.86	567.4*
32.20	32.10	2.3	32.09	45.05	84.9	32.20	61.42	369.2
39.71	38.67	2.3	39.71	51.96	84.9	39.71	66.61	369.2
34.14	34.11	2.7	34.19	69.17	649.2	34.19	68.17	567.4*
28.41	29.47	2.3	28.43	41.37	84.9	28.49	56.51	369.2

*Indica BLUP ($AR(1)$) mejor que BLUP(CS) en efectos fijos, cuando falta el dato.

7. Consideraciones finales

- De acuerdo con lo analizado en el presente trabajo, el método de Barroso et al. (1998) se considera una herramienta valiosa para la imputación de datos cuando se tiene en cuenta un modelo de efectos mixtos.
- En el trabajo se consideró el caso de un dato faltante, pero el método puede generalizarse a más de uno; para ello puede seguirse a Barroso et al. (1998).
- Los datos provenientes de diseños *crossover* en algunos casos consideran modelos que incluyen el efecto residual *(carryover)* del tratamiento; esto puede verse en Oman & Seiden (1988). El presente trabajo deja planteada la posibilidad de imputar datos en dichos casos, modificando la matriz diseño de los efectos aleatorios.
- Uno de los temas para continuar es el estudio de las distribuciones de los estadísticos de prueba, después de imputación.

Agradecimientos

Agradecemos a los evaluadores por sus valiosos aportes que permitieron mejorar el artículo. Este trabajo hace parte de la producción del Grupo de Investigación en Estadística Aplicada del Departamento de Estadística.

Recibido: septiembre de 2006
Aceptado: octubre de 2006

Referencias

Barroso, L., Bussab, W. & Knott, M. (1998), 'Best Linear Unbiased Predictor in the Mixed Model with Incomplete Data', *Commun. Statis. - Theory Meth.* **27**(1), 121–129.

Carriere, K. (1994), 'Incomplete Repeated Measures Data Analysis in the Presence of Treatments Effects', *J. Am. Stat. Assoc.* **89**, 680–686.

Ebbutt, A. (1984), 'Three-Period Crossover Design for Two Treatments', *Biometrics* **40**, 219–224.

Grajales, L. (2006), Imputación de datos en modelos mixtos con errores correlacionados: caso de diseños *switchback*, *Tesis de maestría, Ciencias-Estadística*, Universidad Nacional de Colombia, sede Bogotá.

Jones, B. & Kenward, M. (2003), *Design and Analysis of Cross-Over Trials*, 2 edn, Chapman and Hall.

Li, C. (1964), *Introduction to Experimental Statistics*, McGraw-Hill.

Li, J. (1995), Analysis of Switchback Designs, Technical report, Louisiana State University.

Little, R. & Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2 edn, Jonh Wiley & Sons.

Lucas, H. L. (1956), 'Switchback Trials for more than two Treatments', *Journal of Dairy Science* **39**, 146–154.

Macchiavelli, R. (1997), Analysis of switch-back designs in dairy experiments, *in* '5^{th} Meeting of the International Biometric Society Network for Central America, the Caribbean, Mexico, Colombia and Venezuela'.

Matthews, J. (1988), 'Recent Developments in Crossover Designs', *International Statistical Review* **56**(2), 117–127.

Oman, S. & Seiden, E. (1988), 'Switch-back Designs', *Biometrika* **75**, 81–89.

Richardson, B. & Flack, V. (1996), 'The Analysis of Incomplete Data in Three-period Two-treatment Cross-over Design for Clinical Trials', *Statistics in Medicine* **15**, 127–143.

SAS (2005), *SAS/STAT software: changes and enhancements*.

Searle, S., Casella, G. & McCulloch, C. (1992), *Variance Components*, Jonh Wiley & Sons.

Tempelman, R. (2004), 'Experimental Design and Statistical Methods for Classical and Bioequivalence Hypothesis Testing with an Application to Dairy Nutrition Studies', *J. Anim. Sci.* **82 (E. Suppl.)**, E162–E172.

Wolfinger, R. D. (1993), 'Covariance Structure Selection in General Mixed Models', *Communication in Statistics, Simulation and Computation* **22**, 1079–1106.

Yates, F. (1933), 'The Analysis of Replicated Experiments when the Field Results are Incomplete', *Annals of Eugenics* pp. 27–33.