Revista Colombiana de Estadística

Volumen 24 (2001) N°2, páginas 73 *a* 89

ESTIMACIÓN DE DATOS FALTANTES EN BLOQUES INCOMPLETOS BALANCEADOS CONECTADOS*

HENRY MENDOZA R.**
LUIS ALBERTO LÓPEZ P. ***

Resumen

La presencia de datos faltantes en un experimento hace que se pierda la estructura original y en consecuencia la ortogonalidad del diseño original planteado. En este trabajo, se estiman los datos faltantes en un diseño de bloques incompletos balanceados tipo I conectados (BIBC), aplicando el método de covarianza de *Bartlett*; se hace la corrección del análisis de varianza y finalmente se aplican los resultados obtenidos para los datos de un experimento realizado bajo un diseño de bloques incompletos en repeticiones. Los resultados obtenidos, muestran que el análisis de varianza corregido depende de los datos observados y de la estimación de los datos faltantes.

***Palabras Claves**: Análisis de Covarianza, Bloques Incompletos, Datos Faltantes, Análisis Intrabloque, Diseños conectados.*

*Esta publicación hace parte de la producción academica del grupo "Estadística Aplicada a la Investigación Experimental, Industrial y Biotecnología "

**Profesor Asistente, Departamento de Estadística, Universidad Nacional de Colombia; email: hmendoza@matematicas.unal.edu.co

***Profesor Asociado, Departamento de Estadística, Universidad Nacional de Colombia; email: alopez@matematicas.unal.edu.co

Abstract

Missing data in experiments cause loss in the original structure and therefore in the orthogonality of the experimental design . This paper estimates missing data in an incomplete connected balanced block design type *I (ICBB)*, applying the Bartlett's covariance method; a correction of the analysis of variance is done, and finally the results are applied in experimental data realized in incomplete blocks design with repetitions. The obtained results show that the corrected analysis of variance depends on the observed data and estimation of the missing data.
***Key Words**: Covariance Analysis, Incomplete Blocks, Missing Data, Within block Analysis, Connected Designs.*

1. Introducción

En la investigación experimental, algunas veces no se pueden obtener todos los datos planeados a observar debido a motivos inherentes al diseño o a que en la construcción del mismo hay pérdida de unidades experimentales, lo que conlleva a tener un diseño no ortogonal. Este hecho no tiene consecuencias cuando el diseño experimental es un completamente al azar, pero en el caso de los diseños de bloques y específicamente en el diseño de bloques incompletos al azar no se cumplen algunas propiedades importantes como el balanceamiento y esto conlleva a que cambia la estructura de bloques incompletos.

2. Bloques incompletos balanceados

Según Hinkelman y Kempthorne (1994), los diseños experimentales se encuentran clasificados en un orden jerárquico de acuerdo al número de factores de control local o bloqueo. Cuando los diseños tienen un factor "extraño", se dice que son *Diseños de Bloques al Azar*, los cuales de acuerdo a algunas características se clasifican en: *Completos al Azar (BCA), al Azar Generalizado (BAG), Incompletos al azar (BI)*. Estos últimos se caracterizan porque no todos los tratamientos pueden ser aplicados en cada bloque. Dentro del tipo de *BI*, cabe resaltar los diseños de *Bloques Incompletos Balanceados*, introducidos por Yates en 1936, los cuales define Raghavarao (1971) de la siguiente manera:

Definición 1. Un diseño de *bloques incompletos balanceados* (BIB), es un arreglo de t símbolos en b conjuntos, cada uno con k $(k < t)$ símbolos que satisfacen las siguientes condiciones:

1. *Todo símbolo aparece a lo más una vez en cada conjunto.*

2. *Todo símbolo aparece en exactamente r conjuntos.*

3. *Cualquier par de símbolos aparecen exactamente λ conjuntos*, donde λ es un entero positivo.

De la definición se sigue que un BIB tiene como parámetros (t, b, k, r, λ), los cuales no son independientes y además cumplen con las siguientes relaciones:

1. $tr = bk$

2. $\lambda\,(t-1) = r\,(k-1)$

3. $b \geq t$ conocida como *Desigualdad de Fisher*.

Como λ debe ser entero, es claro que un *BIB*, no existe para todos los valores de t, k y r; aunque para valores de t, k y r se produzca un λ entero, puede no existir el diseño. En efecto, existe un número limitado de *BIB*. Raghavarao (1971), presenta una lista completa de parámetros de diseños existentes y sus métodos de construcción.

Según Cochran y Cox (1980), existen varios tipos de *BIB*: Tipo I, Tipo II, Tipo III, y Tipo IV. En este trabajo se utilizó el tipo I, el cual se tiene cuando los bloques pueden ser agrupados en repeticiones en repeticiones como se muestra en el siguiente arreglo:

Repeticiones				
1	2	3	4	5
$(1,2)$	$(1,3)$	$(1,4)$	$(1,5)$	$(1,6)$
$(3,4)$	$(2,5)$	$(2,6)$	$(2,4)$	$(2,3)$
$(5,6)$	$(4,6)$	$(3,5)$	$(3,6)$	$(4,5)$

En Raghavarao (1971), John (1980), Dud y Giri (1986) y Dodge (1985) entre otros, se presentan los métodos de construcción de los diseños de bloques incompletos. Clásicamente tales diseños son construidos usando la teoría de cuadrados latinos, geometría finita y la teoría de campos de Galois. Entre los métodos se tienen: Series ortogonales 1 y 2 de Yates (SO1 y SO2), diseños irreducibles, conjuntos diferencias para hallar diseños de bloques incompletos simétricos, diseño complemento $\overline{D}$, diseño residual, y el diseño derivado. El método más sencillo de construir es el *diseño irreducible*, el cuál consiste en formar todos los conjuntos de $k-$uplas de los t tratamientos.

3. Análisis intrabloque

En este trabajo se aplicó el *análisis intrabloque*, en el cual se hacen comparaciones entre las unidades experimentales del mismo bloque, estos son usados para comparar los efectos de tratamientos y se recomienda en cualquier experimento con arreglo de BI, ya que los bloques sean fijos o aleatorios; pero, según Kirk (1968) es más eficiente cuando se utiliza con bloques de efectos fijos, ya que los estimados de los efectos de bloques son óptimos. Este análisis permite además eliminar las diferencias entre los bloques. El análisis intrabloque es aplicado bajo el modelo

$$y_{ij} = n_{ij}(\mu + \tau_i + \beta_j + \varepsilon_{ij}) \qquad i = 1, 2, \ldots, t \quad j = 1, 2, \ldots k \tag{1}$$

donde, n_{ij} hace referencia al número de veces que el i-ésimo tratamiento aparece en el j-ésimo bloque, solo toma los valores 1 o 0, cuando $n_{ij} = 0$ significa que no hay dato registrado en el experimento; μ es la media global de los tratamientos; τ_i es el efecto del i-ésimo tratamiento y además $\sum_i \tau_i = 0$; β_j es el efecto del j-ésimo bloque, con la restricción $\sum_j \beta_j = 0$; ε_{ij} son las variables aleatorias para el error, igualmente distribuidas, normales, independientes con media 0 y varianza σ_ε^2.

Para la estimación de parámetros se aplicó el método de mínimos cuadrados, a partir del cuál se obtuvieron las siguientes *ecuaciones normales:*

$$y_{..} = n\widehat{\mu} + r\sum_{i=1}^{t}\widehat{\tau}_i + k\sum_{j=1}^{b}\widehat{\beta}_j \tag{2}$$

$$y_{i.} = r\widehat{\mu} + r\widehat{\tau}_i + \sum_{j=1}^{b} n_{ij}\widehat{\beta}_j, \qquad i = 1, 2, ..., t \tag{3}$$

$$y_{.j} = k\widehat{\mu} + \sum_{i=1}^{t} n_{ij}\widehat{\tau}_i + k\widehat{\beta}_j, \qquad j = 1, 2, ..., b \tag{4}$$

Despejando $\widehat{\beta}_j$ de (4), reemplazando en (3) y haciendo algunas simplificaciones se tiene

$$\left(y_{i.} - \sum_{j=1}^{b} n_{ij}\overline{y}_{.j}\right) = \widehat{\tau}_i\left(r - \frac{r}{k}\right) - \frac{\lambda}{k}\sum_{m\neq i}^{t}\widehat{\tau}_m \tag{5}$$

haciendo

$$Q_i = y_{i.} - \sum_{j=1}^{b} n_{ij}\overline{y}_{.j}$$

y se reemplaza en (5) se tiene

$$Q_i = \widehat{\tau}_i\left(r - \frac{r}{k}\right) - \sum_{m\neq i}^{t}\widehat{\tau}_m \qquad (i = 1, 2, ..., t) \tag{6}$$

Las ecuaciones anteriores se llaman *ecuaciones normales reducidas*. Q_i es llamado el total ajustado del i-ésimo tratamiento. Las t ecuaciones en (6) no son independientes, porque cuando se suman miembro a miembro a ambos lados se anulan; es decir,

$$\sum_{i=1}^{t} Q_i = \sum_{i=1}^{t} \left[\widehat{\tau}_i \left(r - \frac{r}{k} \right) - \frac{\lambda}{k} \sum_{m \neq i}^{t} \widehat{\tau}_m \right] = 0$$

Para que la solución sea única se impone la restricción $\sum_{i=1}^{t} \widehat{\tau}_i = 0$; esto implica que los $\widehat{\tau}_i$ son estimados como desviaciones de sus medias; si además se suma y resta $\frac{\lambda}{k}\widehat{\tau}_i$ en el lado derecho de las ecuaciones normales reducidas en (6) y se aplica la restricción anterior:

$$\begin{aligned} Q_i &= \widehat{\tau}_i \left(r - \frac{r}{k} \right) + \frac{\lambda}{k}\widehat{\tau}_i - \frac{\lambda}{k}\widehat{\tau}_i - \sum_{m \neq i}^{t} \widehat{\tau}_m \left(\frac{\lambda}{k} \right) \\ &= \widehat{\tau}_i \left(r - \frac{r}{k} + \frac{\lambda}{k} \right) - \frac{\lambda}{k} \sum_{m=i}^{t} \widehat{\tau}_m \end{aligned}$$

y así el estimador del i-ésimo tratamiento $\widehat{\tau}_i$ es dado por:

$$\widehat{\tau}_i = \frac{Q_i}{\left(r - \frac{r}{k} + \frac{\lambda}{k} \right)} \qquad (i = 1, 2, ..., t) \tag{7}$$

Ahora, simplificando el denominador de (7) y aplicando algunas de las relaciones de los *BIB* se tiene que el estimador del efecto del i-ésimo tratamiento τ_i es dado por

$$\widehat{\tau}_i = \frac{k}{\lambda t} Q_i \qquad (i = 1, 2, ..., t)\,. \tag{8}$$

En Garza (1988), se describe la tabla de análisis de varianza para este diseño:

Tabla 1. Análisis de varianza intrabloque para un diseño de bloques incompletos

Causa de variación	Grados de libertad	Sumas de cuadrados	Cuadrados medios	F calculada
Tratamientos (ajustados)	$t-1$	$\frac{k}{\lambda t}\sum_{i=1}^{t} Q_i^2$	$\mathrm{CM_{ttos}}$	$\frac{\mathrm{CM_{tto}}}{\mathrm{CM_{error}}}$
Bloques (no ajustados)	$b-1$	$\sum_{j=1}^{b} \frac{B_j}{k} - \frac{y_{...}^2}{rt}$		
Error Intrabloque	$rt-t-b+1$	Por diferencia	$\mathrm{CM_{error}}$	
Total	$rt-1$	$\sum y_{ij}^2 - \frac{y_{...}^2}{rt}$		

Para la estimación de la información faltante se requiere que el diseño sea *conectado*. En Raghavarao (1971) se cita el trabajo de Chakrabarti quien en 1963

definió un *diseño conectado* como aquel en el cual la matriz C tiene rango $t-1$. El estudio de la *conectes* también puede hacerse en Searle (1987), Murray-Smith (1985) y Hocking (1996).

4. Propuesta metodológica para imputar datos en BIB

Según Searle (1987), los datos en un diseño de *BIB* son por su estructura de construcción datos desbalanceados-planeados, debido a que no poseen el mismo número de observaciones por celda, sin embargo si por circunstancias ajenas al experimento hay pérdida de observaciones, los datos se clasifican como *desbalanceados faltantes*.

En este trabajo se proporciona una metodología basada en el método de *Bartlett* para imputar datos faltantes en bloques incompletos balanceados, considerándolos cuando se presentan en la variable respuesta.

La pérdida de datos en un *BIB*, tiene las siguientes consecuencias:

1. Se pierde la estructura del diseño, en el sentido que el número de veces que aparece cada par de tratamientos en el diseño (λ) no es constante.
2. El diseño se vuelve no balanceado; es decir, no todos los contrastes elementales $\tau_i - \tau_{i'}$ $(i \neq i')$ se pueden estimar con la misma precisión, es decir,
$$var\,(\tau_i - \tau_{i'}) \neq var\,(\tau_k - \tau_{k'}) \quad (i \neq i'\,;\, k \neq k')$$
3. Puede perderse la propiedad de conectes; es decir, no todos los contrastes elementales, $\tau_i - \tau_{i'}$ $(i \neq i')$ son estimables.

Para el desarrollo de la propuesta se consideró el modelo para un diseño de bloques incompletos balanceados con t tratamientos, arreglados en b bloques que contienen k unidades experimentales y r réplicas por tratamiento. El modelo sin considerar los datos faltantes es

$$\mathbf{y} = \mathbf{X}\theta + \varepsilon \tag{9}$$

donde

$$\mathbf{X} = [\mathbf{1}_{n\times 1} \vdots \mathbf{X}_{\tau(t-1)} \vdots \mathbf{X}_{\beta(b-1)}]$$
$$\theta' = \left[\mu, \tau_1, \ldots, \tau_{(t-1)}, \beta_1, \ldots, \beta_{(b-1)}\right]$$

$\mathbf{X}$ es la matriz diseño reparametrizada de orden $n \times (t+b-1)$ sujeto a las restricciones

$$\sum_{i=1}^{t}\tau_i = 0 \;\text{ y }\; \sum_{j=1}^{b}\beta_j = 0$$

$\mathbf{y}_{n\times 1}$ vector de observaciones; $\mathbf{1}$ es un vector columna de unos de tamaño n; $\mathbf{X}_\tau$ y $\mathbf{X}_\beta$ representan respectivamente las matrices diseño para los tratamientos y los bloques; μ es la media global; τ es el vector de efectos de tratamientos y β es el vector de efectos de bloques y $\varepsilon_{n\times 1}$ es el vector de errores aleatorios ε_{ij}, $i = 1, 2, \ldots, t$; $j = 1, 2, \ldots, b$, se asume que las variables aleatorias ε_{ij} son igualmente distribuidas normales, independientes, con media 0 y varianza constante σ_ε^2.

Cuando en (9) se asume m datos faltantes, el modelo estadístico propuesto por Bartlett, el cual puede estudiarse con más detalle en Little y Rubin (1987), es

$$\mathbf{y} = \mathbf{X}\theta + \mathbf{Z}\gamma + \varepsilon \tag{10}$$

con la matriz $\mathbf{Z} = (z_1, \ldots, z_n)$ de orden $d_z = n \times m$, la l-ésima fila z_l está conformada por ceros excepto en la posición del i-ésimo dato faltante donde toma el valor de -1. La matriz $\mathbf{Z}$ se puede expresar como

$$\mathbf{Z} = \begin{pmatrix} \oslash \\ -\mathbf{I}_m \end{pmatrix}$$

Además el método de *Bartlett* en los *BIB* con información faltante considera el modelo particionado

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \theta + \mathbf{Z}\gamma + \varepsilon \tag{11}$$

donde $\mathbf{y}_1$ representa el vector de datos observados y $\mathbf{y}_2$ representa el vector de los m datos faltantes. La metodología de Bartlett inicia con la imputación de un vector cualquiera $\widetilde{\mathbf{y}}_2$, como estimación inicial de los datos faltantes.

Del modelo (10), por mínimos cuadrados se obtienen las ecuaciones normales

$$\mathbf{X}'\mathbf{X}\widehat{\theta} + \mathbf{X}'\mathbf{Z}\widehat{\gamma} = \mathbf{X}'\mathbf{y} \tag{12}$$

$$\mathbf{Z}'\mathbf{X}\widehat{\theta} + \mathbf{Z}'\mathbf{Z}\widehat{\gamma} = \mathbf{Z}'\mathbf{y} \tag{13}$$

Estas fueron simplificadas utilizando el vector de respuesta $\mathbf{y}$ bajo el modelo particionado (11) con el vector de supuestos iniciales $\widetilde{\mathbf{y}}_2 = \mathbf{0}$ con los cuales se obtienen las siguientes equivalencias:

$$\begin{array}{llcll}
\text{(i)} & \mathbf{Z}'\mathbf{y} & = & \begin{pmatrix} \oslash & -I_m \end{pmatrix} \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{0} \end{pmatrix} & = \quad \mathbf{0} \\
\text{(ii)} & \mathbf{Z}'\mathbf{X} & = & \begin{pmatrix} \oslash & -I_m \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} & = \quad -\mathbf{I}_m\mathbf{X}_2 \quad = -\mathbf{X}_2 \\
\text{(iii)} & \mathbf{Z}'\mathbf{Z} & = & \begin{pmatrix} \oslash & -I_m \end{pmatrix} \begin{pmatrix} \oslash \\ -\mathbf{I}_m \end{pmatrix} & = \quad \mathbf{I}_m
\end{array}$$

Reemplazando (i), (ii) y (iii) en (12) y (13) se obtienen las ecuaciones normales

simplificadas

$$(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)\,\widehat{\theta} - \mathbf{X}_2'\widehat{\gamma} = \mathbf{X}_1'\mathbf{y}_1 \tag{14}$$

$$-\mathbf{X}_2\widehat{\theta} + \widehat{\gamma} = \mathbf{0} \tag{15}$$

Despejando $\widehat{\theta}$ de (14) y reemplazando en (15) se tiene:

$$\widehat{\theta} = (\mathbf{X}'\mathbf{X})^{-1}\,(\mathbf{X}_1'\mathbf{y}_1 + \mathbf{X}_2'\widehat{\gamma}) \tag{16}$$

y

$$\left[-\mathbf{X}_2\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}_2' + \mathbf{I}_m\right]\widehat{\gamma} = \mathbf{X}_2\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}_1'\mathbf{y}_1$$

haciendo

$$\mathbf{W} = \mathbf{I}_m - \mathbf{X}_2\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}_2' \tag{17}$$

$$\mathbf{V}' = \mathbf{X}_2\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}_1' \tag{18}$$

así se obtiene como estimador de γ:

$$\widehat{\gamma} = -\mathbf{W}^{-1}\mathbf{V}'\mathbf{y}_1 \tag{19}$$

Este resultado permite imputar la información, con lo cual, finalmente el vector estimado para los datos faltantes es dado por

$$\widehat{\mathbf{y}}_0 = -\mathbf{W}^{-1}\mathbf{V}'\mathbf{y}_1$$

de este resultado se observa que *la estimación de los datos faltantes sólo depende del vector de valores observados* $\mathbf{y}_1$ y la matriz diseño $\mathbf{X}$.

Lema 4.1 $\mathbf{W}$ y $\mathbf{V}'$ definidas en (17) y (18) respectivamente satisfacen

$$\mathbf{V}'\mathbf{V} + \mathbf{W}^2 = \mathbf{W} \tag{20}$$

La anterior estimación de $\widehat{\mathbf{y}}$ es importante de manera analítica, pero para comodidad computacional se determinó la estimación de este vector utilizando el modelo general (10), para este modelo las ecuaciones normales son dadas en (12) y (13) donde al despejar $\widehat{\theta}$ se obtiene

$$\widehat{\theta} = (\mathbf{X}'\mathbf{X})^{-1}\left(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{Z}\widehat{\gamma}\right) \tag{21}$$

Premultiplicando la ecuación anterior por $\mathbf{X}$ y reemplazando en (13) se tiene:

$$\mathbf{Z}'\mathbf{R}_\mathbf{X}\mathbf{Z}\widehat{\gamma} = \mathbf{Z}'\mathbf{R}_\mathbf{X}\mathbf{y}$$

donde $\mathbf{R_X} = (\mathbf{I} - \mathbf{P_X})$, teniendo así la solución para $\widehat{\gamma}$

$$\widehat{\gamma} = (\mathbf{Z}'\mathbf{R_X}\mathbf{Z})^{-1}\,\mathbf{Z}'\mathbf{R_X}\mathbf{y} \tag{22}$$

y en consecuencia la estimación del vector de datos faltantes en términos del modelo general (11) es

$$\widehat{\mathbf{y}}_2 = -\,(\mathbf{Z}'\mathbf{R_X}\mathbf{Z})^{-1}\,\mathbf{Z}'\mathbf{R_X}\mathbf{y}$$

4.1. Estimación corregida de parámetros

El estimador del vector de parámetros θ cuando se tiene como supuestos iniciales para los datos faltantes el vector nulo ($\widetilde{\mathbf{y}}_2 = \mathbf{0}$), se obtuvo reemplazando (19) en (16) y así

$$\begin{aligned}\widehat{\theta} &= (\mathbf{X}'\mathbf{X})^{-1}\left[\mathbf{X}_1'\mathbf{y}_1 + \mathbf{X}_2'\widehat{\gamma}\right] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\left[\mathbf{X}_1'\mathbf{y}_1 - \mathbf{X}_2'\mathbf{W}^{-1}\mathbf{V}'\mathbf{y}_1\right]\end{aligned} \tag{23}$$

En (23), se observa que el vector de parámetros estimado $\widehat{\theta}$ depende únicamente del vector de valores observados $\mathbf{y}_1$.

Para efectos computacionales, es más conveniente expresarla en términos de la estimación de $\widehat{\gamma}$ obtenida en (22), es decir,

$$\begin{aligned}\widehat{\theta} &= (\mathbf{X}'\mathbf{X})^{-1}\left[\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{Z}\widehat{\gamma}\right] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\left[\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{Z}\,(\mathbf{Z}'\mathbf{R_X}\mathbf{Z})^{-1}\,\mathbf{Z}'\mathbf{R_X}\mathbf{y}\right] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\left[\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{Z}\,(\mathbf{Z}'\mathbf{R_X}\mathbf{Z})^{-1}\,\mathbf{Z}'\mathbf{R_X}\mathbf{y}\right]\end{aligned} \tag{24}$$

Una vez obtenido el vector estimado para la información faltante, se procede a imputar esta información y posteriormente realizar el análisis de los datos en un BIB ajustando el respectivo análisis de varianza como se muestra en la siguiente sección.

4.2. Corrección al Análisis de Varianza

Obtenida la expresión para la estimación de los datos faltantes se procedió a llevar a cabo la corrección de la suma de cuadrados del error con la cual se realiza en forma adecuada la hipótesis sobre efectos de tratamientos. Este resultado se obtiene reemplazando (22) y (24) en la suma de cuadrados del error (SC_{error}), del modelo (10), con lo cual se obtiene:

$$\begin{aligned}SC_{\text{error modelo completo}} &= \mathbf{y}'\mathbf{y} - (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y} - \widehat{\gamma}'\mathbf{Z}'\mathbf{R_X}\mathbf{y} \\ &= SC_{\text{error modelo 1}} - \widehat{\gamma}'\mathbf{Z}'\mathbf{R_X}\mathbf{y}\end{aligned} \tag{25}$$

donde $SC_{\text{error modelo 1}}$ es la suma de cuadrados del error para el modelo sin datos faltantes. La expresión (25) muestra que la suma de cuadrados del error para el modelo (10) cuando se corrigen los datos imputados es más pequeña que la suma de cuadrados del error del modelo (9) esta reducción es dada por $\widehat{\gamma}'\mathbf{Z}'\mathbf{R}_{\mathbf{X}}\mathbf{y}$; y el estimador de la varianza del error experimental es dado por

$$\widehat{\sigma}_{\varepsilon}^{2} = \frac{SC_{error}}{(bt - d_x - d_z)}$$

donde $d_z = ran(\mathbf{Z})$ (número de datos faltantes) y $d_x = ran(\mathbf{X})$.

Para obtener la estadística de prueba para la hipótesis de no efectos de tratamientos $H_0 : \tau = \mathbf{0}$, se ajustó el modelo de covarianza general bajo H_0 y se obtuvo la suma de cuadrados del modelo reducido, $SC_{\text{modelo reducido}}$ de la misma manera que la obtenida para el modelo (10). Si la matriz asociada a la hipótesis nula es de rango $d = t - 1$, entonces la estadística de prueba para la hipótesis es:

$$F_o = \frac{\left(SC_{\text{error modelo reducido}} - SC_{\text{error modelo completo}}\right)/d}{\widehat{\sigma}_{\varepsilon}^{2}} \sim F(d; bt - d_x - d_z)$$

4.3. Varianza de contrastes lineales de tratamientos

Debido a que el interés del experimento se centra en la comparación de contrastes de efectos de tratamientos se desarrolla a continuación el procedimiento para la estimación de la varianza de contrastes de tratamientos, para ello se debe tener en cuenta las siguientes equivalencias:

$$\mathbf{X}'\mathbf{X} = \mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2 = \mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_2'\mathbf{W}^{-1}\mathbf{V}'\mathbf{X}_1$$

$$\mathbf{X}_2'\mathbf{W}^{-1}\mathbf{V}'\mathbf{X}_1 = -\left[\mathbf{X}'\mathbf{X} - \mathbf{X}_1'\mathbf{X}_1\right] = \mathbf{X}_2'\mathbf{X}_2 \tag{26}$$

y teniendo en cuenta (23) se obtiene que:

$$var\left(\mathbf{X}'\mathbf{X}\widehat{\theta}\right) = \left[\mathbf{X}'\mathbf{X} + \mathbf{X}_2'\mathbf{W}^{-1}\mathbf{X}_2\right]\sigma_{\varepsilon}^{2}$$

De este resultado se concluye que si un contraste no involucra un tratamiento con información faltante, entonces la varianza de combinaciones lineales será la misma que se hubiera obtenido del diseño completo.

4.4. Propiedades del Vector $\widehat{\gamma}$

El vector estimado de coeficientes de regresión es un estimador insesgado.

Entretanto de (19) se sigue que

$$\begin{aligned} var\left(\widehat{\gamma}\right) &= var\left(-\mathbf{W}^{-1}\mathbf{V}'\mathbf{y}_1\right) \\ &= \mathbf{W}^{-1}\mathbf{V}'\mathbf{V}\mathbf{W}^{-1}\sigma_{\varepsilon}^2 \end{aligned}$$

y teniendo en cuenta el *lema* se tiene que

$$\begin{aligned} var\left(\widehat{\gamma}\right) &= \mathbf{W}^{-1}\left(\mathbf{W}-\mathbf{W}^2\right)\mathbf{W}^{-1}\sigma_{\varepsilon}^2 \\ &= \left(\mathbf{W}^{-1}-\mathbf{I}\right)\sigma_{\varepsilon}^2 \end{aligned}$$

y de (22) se tiene finalmente la expresión general de la varianza, es decir,

$$\begin{aligned} Var\left(\widehat{\gamma}\right) &= Var\left[\left(\mathbf{Z}'\mathbf{R}_{\mathbf{X}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{R}_{\mathbf{X}}\mathbf{y}\right] \\ &= \left(\mathbf{Z}'\mathbf{R}_{\mathbf{X}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{R}_{\mathbf{X}}Var\left(\mathbf{y}\right)\left[\left(\mathbf{Z}'\mathbf{R}_{\mathbf{X}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{R}_{\mathbf{X}}\right]' \\ &= \left(\mathbf{Z}'\mathbf{R}_{\mathbf{X}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{R}_{\mathbf{X}}\mathbf{Z}\left[\mathbf{Z}'\mathbf{R}_{\mathbf{X}}'\mathbf{Z}\right]^{-1}Var\left(\mathbf{y}\right) \\ &= \left(\mathbf{Z}'\mathbf{R}_{\mathbf{X}}\mathbf{Z}\right)^{-1}\sigma_{\varepsilon}^2 \end{aligned}$$

por ser $\mathbf{R}_{\mathbf{X}}$ una matriz simétrica e idempotente.

5. Ejemplo

En esta sección se ilustran los resultados obtenidos en la *sección 5*, los datos se reportan en Cochran y Cox (1980) en donde se enuncia que:

El experimento lo llevó a cabo el doctor Pauline Paul en el colegio del estado de Iowa. Su propósito fue comparar los efectos de la longitud del período de almacenamiento a baja temperatura en la suavidad y sabor de carne de res. Se probaron 6 períodos de almacenamiento (0, 1, 2, 4, 9 y 18 días). Estos están identificados por los símbolos de tratamiento 1, 2, 3, 4, 5 y 6, respectivamente. Se utilizaron treinta piezas de carne; 6 de ellas se obtuvieron de cada uno de 4 músculos, mientras que 3 de éstos dieron dos piezas cada uno. Las piezas de cualquier músculo se agrupan naturalmente en pares, ya que a cada pieza del lado izquierdo del animal le corresponde otra pieza en el lado derecho. Por la experiencia se creía que las piezas de cualquier par darían resultados muy similares. Se esperaba que la variación entre diferentes pares del mismo músculo fuera algo mayor, y la variación entre músculos, aún más grande.

Lo anterior sugirió un diseño en bloques incompletos balanceado tipo I, con parámetros $t = 6, k = 2, r = 5, b = 15, \lambda = 1$ y $E = 0{,}6$ según el siguiente arreglo:

Bloque	Rep. I		Bloque	Rep. II		Bloque	Rep. III	
(1)	1	2	(4)	1	3	(7)	1	4
(2)	3	4	(5)	2	5	(8)	2	6
(3)	5	6	(6)	4	6	(9)	3	5

Bloque	Rep. IV		Bloque	Rep.V	
(10)	1	5	(13)	1	6
(11)	2	4	(14)	2	3
(12)	3	6	(15)	4	5

Los dos tratamientos en cada bloque se aplicaron sobre las piezas del lado izquierdo y derecho como un par para cada músculo (repetición). Al agrupar los bloques en repeticiones, lo más natural fue poner piezas del mismo músculo en la misma repetición. La evaluación fue hecha por cuatro jueces y cada uno de ellos marcaba sobre una escala de 0 a 10. Las evaluaciones mostradas son sus totales (de un máximo de 40), donde una alta puntuación indica una pieza de carne muy suave.

El arreglo de los datos fue:

Bloque	Rep. I (músculo 1)			Rep. II (músculo 2)			Rep. III (músculo 3)	
(1)*	7[1]**	17[2]	(4)	17[1]	27[3]	(7)	10[1]	25[4]
(2)	26[3]	25[4]	(5)	23[2]	27[5]	(8)	26[2]	37[6]
(3)	33[5]	29[6]	(6)	29[4]	30[6]	(9)	24[3]	26[5]

Bloque	Rep. IV (músculo 4)			Rep. V (músculo 5)	
(10)	25[1]	40[5]	(13)	11[1]	27[6]
(11)	25[2]	34[4]	(14)	24[2]	21[3]
(12)	34[3]	32[6]	(15)	26[4]	32[5]

$*$ El Bloque (i) $(i = 1, 2, \ldots, 15)$, corresponde a un par de piezas de carne de un mismo músculo

$**$ el número entre paréntesis rectangulares representa el tratamiento aplicado.

Al aplicar el *algoritmo R*, programado por Melo y Lozano (1998), y con la metodología de Chakrabarti estudiadas en Mendoza (2002), se encontró que esa estructura de *BIB* era conectada, con lo cual se garantiza que la metodología propuesta se pudo aplicar a ese conjunto de datos experimentales. Los resultados del análisis de varianza se obtuvieron con los programas desarrollados por Mendoza (2002).

Tabla 2. Resumen del Análisis de Varianza

Causa de variación	G.L.	Sumas de cuadrados	Cuadrados medios	F_0	Valor p
Tratamientos (ajustados)	5	520,17	104,033	13,45	0,0004
Repeticiones	4	298,5	74,6	9,65	0,0018
Bloques	10	753,0			
Error	10	77,3			
Total	29	1649,0			

La *suma de cuadrados ajustada para los tratamientos* es obtenida de la suma de cuadrados tipo III de la salida de *SAS* (1990) y la *suma de cuadrados no ajustada para los tratamientos* es obtenida de la suma de cuadrados tipo I de la salida de *SAS* y así la *suma de cuadrados no ajustada para los bloques* es obtenida por diferencia, de lo cual resulta que el valor de *suma de cuadrados ajustada para bloques* es 753,0.

Para el análisis de los datos faltantes, inicialmente, se presenta el método de *Cornish* para el caso de un dato faltante y posteriormente se presenta el método propuesto. La aplicación del método de *Cornish* se llevó a cabo suponiendo que el dato del tratamiento 5 en la repetición IV está ausente, en Cochran y Cox (1980), se presenta la fórmula de *Cornish* para la estimación del dato faltante:

$$x = \frac{tr\,(k-1)\,y_{.j^*} + k\,(t-1)\,Q - (t-1)\,Q'}{(k-1)\,[tr\,(k-1) - k\,(t-1)]}$$

donde $y_{.j^*}$ es el total del bloque que contiene el dato faltante, Q es el valor del total del tratamiento ajustado donde está el dato faltante $\left(Q_i = ky_{i.} - \frac{1}{k}\sum_{j=1}^{b} n_{ij}y_{.j}\right)$, y Q' es la suma de los valores Q para todos los tratamientos que están en el bloque donde se encuentra el dato faltante.

El valor estimado para el dato faltante fue 41.75 obtenido con el programa desarrollado por González (1997), este resultado coincide con el obtenido por el método de *Bartlett*. Se observa que este valor es muy cercano al valor real 40.

Luego de estimar el dato faltante se procedió a insertarlo y realizar el análisis de varianza. Análogamente se estimaron los datos faltantes para los casos de dos, tres, cinco y diez datos faltantes y se hizo una comparación entre los ANOVAS respectivos insertando el dato faltante y corrigiendo el ANOVA, insertando el dato faltante y no corrigiendo, y cuando los datos estaban completos. Los resultados comparativos se presentan en las tablas 3 a 6. El primero, segundo y tercer valores de cada renglón de las tablas de ANOVA corresponden respectivamente a los datos completos, insertando el dato faltante y no corrigiendo el ANOVA e insertando el dato faltante y corrigiendo el ANOVA.

Tabla 3. ANOVA Comparativo cuando hay un dato faltante

Causa de variación	G.L.	Sumas de cuadrados	Cuadrados medios	F_0	Valor p
Tratamientos (ajustados)	5	520,11	104,03	13,45	0,00036
		548.97	109.79	14.39	0.00027
		408.69	81.74	9.64	0.0020
Repeticiones	4	298,47	74,62	9.65	0.0018
		319.99	79.99	10.48	0.0013
		253.84	58.96	6.95	0.0077
Bloques	10	753,00	75,3	9,737	0,00064
		756.94	75.69	9.92	0.00059
		1307.46	13.07	1.54	0.2640
Error	10	77,33	7,73		
	10	76.3125	7.63		
	9	76.3125	8.48		
Total	29	1648,97			
	29	1702.21			
	28	2046.3			

Tabla 4. ANOVA comparativo cuando hay tres datos faltantes

Causa de variación	G.L.	Sumas de cuadrados	Cuadrados medios	F_0	Valor p
Tratamientos (ajustados)	5	520,11	104,03	13,45	0,00036
		537.15	107.43	14.12	0.0003
		358.92	71.78	6.60	0.0139
Repeticiones	4	298,47	74,62	9.65	0.0018
		30.1	7.52	0.99	0.4500
		120.63	30.16	2.77	0.1136
Bloques	10	753,00	75,3	9,737	0,00064
		1554.57	155.46	20.43	0.00002
		2337.75	233.78	21.50	0.00025
Error	10	77,33	7,73		
	10	76.08	7.608		
	7	76.07	10.87		
Total	29	1648,97			
	29	2197.89			
	26	2893.37			

Tabla 5. ANOVA Comparativo cuando hay cinco datos faltantes

Causa de variación	G.L.	Sumas de cuadrados	Cuadrados medios	F_0	Valor p
Tratamientos (ajustados)	5	520,11	104,03	13,45	0,00036
		543.40	108.68	15.10	0.0002
		362.50	72.50	5.04	0.0501
Repeticiones	4	298,47	74,62	9.65	0.0018
		119.56	29.89	4.15	0.0309
		98.20	24.55	1.71	0.2833
Bloques	10	753,00	75,3	9,737	0,00064
		1956.13	195.61	27.17	0.0000
		3332.11	333.21	23.16	0.0014
Error	10	77,33	7,73		
	10	71.995	7.20		
	5	71.99	14.39		
Total	29	1648,97			
	29	2691.09			
	20	3864.8			

Tabla 6. ANOVA Comparativo cuando hay diez datos faltantes

Causa de variación	G.L.	Sumas de cuadrados	Cuadrados medios	F_0	Valor p
Tratamientos (ajustados)	5	520,11	104,03	13,45	0,00036
		590.54	118.11	18.10	0.0001
		311.26	62.25	ND	ND
Repeticiones	4	298,47	74,62	9.65	0.0018
		2195.04	548.76	84.17	0.0000
		83.36	20.84	ND	ND
Bloques	10	753,00	75,3	9,737	0,00064
		1907.16	190.71	29.23	0.0000
		82.69	8.27	ND	ND
Error	10	77,33	7,73		
	10	65.24	6.52		
	0	65.24	ND		
Total	29	1648,97			
	29	4757.98			
	19				

ND = No se puede determinar

De las comparaciones realizadas en las tablas se puede observar que :

1. Cuando se corrige mediante el método de Bartlett

(a) El cuadrado medio del error tiende a incrementar a medida que el número de datos faltantes aumenta

(b) El cuadrado medio de los tratamientos tiende a disminuir a medida que el número de datos faltantes aumenta

(c) La estadística de prueba F para los tratamientos ajustados disminuye

a medida que el número de datos faltantes aumenta

2. Cuando no se corrige el ANOVA

(a) El cuadrado medio del error tiende a disminuir a medida que el número de datos faltantes aumenta.

(b) El cuadrado medio de los tratamientos se mantiene estable a medida que el número de datos faltantes aumenta.

(c) La estadística de prueba calculada para los tratamientos es mayor que cuando se corrige el ANOVA y tiende a incrementar a medida que el número de datos faltantes aumenta.

3. La decisión sobre la prueba de hipótesis de tratamientos tiende en todos los casos a rechazarse al nivel de 5 %.

6. Conclusiones

Los resultados de este trabajo muestran que el método de análisis de covarianza de Bartlett es una metodología adecuada para la estimación de datos faltantes en Bloques Incompletos Balanceados; la estimación de estos depende de los datos observados. Estas estimaciones son equivalentes a las obtenidas por el método de Cornish.

El método de Bartlett tiene la ventaja sobre el de Cornish de permitir obtener la correción del análisis de varianza.

Una desventaja de la metodología de Bartlett consiste en que a medida que se aumenta el número de datos faltantes, no es posible calcular el cuadrado medio del error.

Se recomienda realizar un análisis comparativo del método de Bartlett con otros métodos utilizados en la estimación de datos incompletos como por ejemplo el propuesto por Onukogu, o con métodos basados en los estimadores EM.

Referencias

[1] COCHRAN, W. G. and COX, G. M. (1980). *Experimental Designs.* John Wiley & Sons.

[2] DAS, M. N., and GIRI N. C. (1986). *Design and Analysis of Experiments.* John whiley & Sons.

[3] DODGE, Y. (1985). *Analysis of Experiments with Missing Data.* John Wiley & Sons.

[4] GARZA, A. M. (1988). *Diseños Experimentales. Métodos y Elementos de Teoría*. Trillas, LIMUSA.

[5] GONZÁLEZ, L.M. (1997). *Ajuste de Bloques Incompletos mediante el Procedimiento IML de SAS*. Tesis de Estadístico. Universidad Nacional de Colombia.

[6] HINKELMAN. K. and KEMPTHORNE. O. (1994). *Design and Analysis of Experiments* volumen I. John Wiley & Sons.

[7] JOHN, P. W. M. (1980). *Incomplete Blocks Designs*. Marcel Dekker.

[8] KIRK, R. E. (1968). *Experimental Design: Procedures for the Behavioral Sciences*. Belmon, Calif.: Brooks/ Cole.

[9] LITTLE, J. A., and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons.

[10] MELO O., y LOZANO R. (1998). *Funciones Estimables en Modelos de clasificación con Datos Desbalanceados a través del Algoritmo de Cholesky*. Tesis de Estadístico. Universidad Nacional de Colombia.

[11] MENDOZA R. H. (2002). *Estimación de Datos Faltantes en Bloques Incompletos Balanceados*. Tesis de Maestría en Ciencias. Departamento de Estadística. Universidad Nacional de Colombia.

[12] MURRAY, L. W. and SMITH, D.W. (1985) *Estimability, Testability and Connectedeness in the cells means model*. Communications in StatisticsPart A Theory and Methods. Vol. 14 pag. 1889-1915.

[13] ONUKOGU, I. B. (1976). *An Alternative to Least Squares for Estimation the Missing Values in a BIB*. Metron 34. 181-186.

[14] RAGHAVARAO, D. (1971). *Constructions and Combinatorial Problems in Design of Experiments*. John Wiley international.

[15] SEARLE,S. R. (1987). *Linear Model for Unbalanced Data*. John Wiley & Sons.

[16] SAS Institute Inc., SAS/STAT User´s Guide (1990). Version 6, Fourth Edition, Vol. 1, Cary, NC: SAS Institute Inc.