SIMULACIÓN PARA EVALUAR EL TRATAMIENTO DE DATOS FALTANTES DE ESTRUCTURA LONGITUDINAL EN EL CONTEXTO DE ENSAYOS CLÍNICOS

Rolando Uranga*[1], Sira Allende[2]**, Geert Molenberghs[3]***
*Centro Nacional Coordinador de Ensayos Clínicos.**
Facultad de Matemática y Computación
3. Instituto Internacional de Bioestadística y Bioinformática Estadística. Universidades de Hasselt y de Lovania

ABSTRACT
Longitudinal studies often suffer from missing data. In this work different methods are compared, by means of a simulation study, to address the estimation task under missing data in longitudinal settings. Completely random, random and non-random missingness mechanisms are studied. Methods used include complete cases, zero imputation, imputation by the mean and last observation carried forward. These methods are refined by the conditional mean imputation, regression and stochastic regression. Such methods are contrasted with maximum likelihood. The goodness of each one in terms of the proportion of missing values is assessed. A qualitative approach that allows a quick and clear assessment of each method is introduced, by means of their classification as Excellent, Good, Regular, or Bad. In this way a simulator written in the IML (Interactive Matrix Language) language of SAS (Statistical Analysis System) is obtained, which allows the evaluation. We conclude that naive imputation methods produce biased estimators. Simple refined methods such as stochastic regression imputation show an acceptable behavior for low proportions of missing values but underestimate variability parameters. The complete case method shows good behavior. Maximum likelihood behaves the best.

KEYWORDS: longitudinal models, maximum likelihood, missingness mechanism

MSC: 62-02; 46N30; 62-07; 00A72; 65C20; 68U20; 62J99

RESUMEN
En estudios longitudinales no se suele contar con la completitud de los datos. En este trabajo se comparan, mediante un estudio de simulación, distintos métodos para abordar la estimación bajo datos faltantes. Se consideran mecanismos de pérdidas completamente aleatorios, aleatorios y no aleatorios. Los métodos empleados incluyen casos completos, imputación por cero, imputación por la media y última observación llevada adelante. Se refinan los métodos mediante imputación por la media condicional, regresión y regresión estocástica. Tales métodos se contrastan con el de máxima verosimilitud. Se valora la bondad de cada uno en términos del porciento de pérdidas. Se introduce un criterio cualitativo que permite realizar una valoración rápida y clara de cada método, mediante su clasificación en Excelente, Bien, Regular o Mal. Se obtiene de este modo un simulador, escrito en el lenguaje IML (Interactive Matrix Language) de SAS (Statistical Analysis System), que permite realizar la evaluación. Se concluye que los métodos de imputación ingenua producen estimadores sesgados. Métodos de imputación simple refinados como regresión estocástica, muestran comportamiento aceptable para bajos porcientos de pérdidas aunque subestiman los parámetros de variabilidad. El método de casos completos muestra buen comportamiento. Máxima verosimilitud es el método de mejor comportamiento.

1. INTRODUCCIÓN

La presencia de datos faltantes es un problema común en ensayos clínicos. En una revisión de lo publicado sobre el tema entre Julio y Diciembre de 2001 (Wood, White et al. 2004) en revistas médicas de alto impacto, se detectan 63 ensayos con presencia de datos faltantes de un total de 71, para un 89%. En una revisión del

[1] rolando.uranga@cencec.sld.cu
[2] sira@matcom.uh.cu
[3] . geert.molenberghs@uhasselt.be

2014 de ensayos aleatorizados por grupos (Díaz-Ordaz, Kenward et al. 2014), se detectan 95 de un total de 132, para un 72%; sin embargo aquí se destacan 41 ensayos (31%) en los cuales "no pudo verificarse la presencia de datos faltantes a partir del reporte publicado". Buuren (2012) critica la pobreza de los reportes de datos faltantes en las publicaciones. Ejemplifica, mediante una colección de pequeñas bases de datos (Hand, Daly et al. 1994), la baja calidad de estos reportes: de las 510 bases de la colección, sólo 13 contenían un código para los datos faltantes. La Tabla 1 muestra la incidencia del problema de datos faltantes en cuatro estudios realizados en Cuba.

Tabla 1
Incidencia de datos faltantes en ensayos clínicos cubanos

Estudio	Variable de respuesta	Proporción de pérdidas
Sertralina[1]	Puntuación de Hamilton	15%
Vacuna contra Melanoma cutáneo[2]	Hemoglobina	69%
Migraprecol[3]	Calidad de vida	30%
Surfacén[4]	Índice de Oxigenación	34%

Fuentes: 1: CENCEC (2014), 2: Uranga (2011), 3: Marrero (2012), 4: Uranga and Molenberghs (2014)

Un incorrecto manejo de los datos faltantes puede tener consecuencias importantes para la validez de las inferencias de algunos estudios clínicos. Métodos que resultan en inferencias válidas bajo ciertos supuestos pueden conducir a falsas conclusiones si el mecanismo de pérdidas no responde a tales supuestos.
Resulta por tanto de interés, realizar un estudio experimental donde se evidencien, de manera precisa, las consecuencias del uso de uno u otro método. En este trabajo proponemos un acercamiento a tal objetivo mediante un estudio de simulaciones de datos longitudinales. El objetivo es estudiar el comportamiento de los métodos de imputación en relación con el mecanismo de pérdidas de la información y el porciento de datos faltantes.
Se valora el desempeño de diferentes métodos para la obtención de buenas estimaciones de los parámetros de dos modelos estadísticos bajo datos faltantes. Se consideran patrones de pérdidas de tipos monótono (a partir del primer valor no observado tampoco se observan los valores sucesivos) e intermitente (la secuencia de valores faltantes es arbitraria).
El trabajo se ha organizado en cinco secciones. En la segunda sección se presentan los modelos longitudinales y los mecanismos de pérdidas considerados. En la sección 3 se presentan los métodos empleados para el tratamiento de datos faltantes. La simulación del modelo longitudinal bajo diferentes mecanismos de pérdidas es descrita en la sección 4; en la misma sección, se introducen diferentes criterios para evaluar el comportamiento de los distintos métodos de manejo de datos faltantes. En la sección 5 se muestran los resultados experimentales de la simulación y se presenta un caso de estudio.

2. MODELOS LONGITUDINALES Y MECANISMOS DE PÉRDIDAS EN ENSAYOS CLÍNICOS

2.1. Modelos longitudinales

Se asume la ejecución de un ensayo clínico para el que han sido seleccionados $N = N_0 + N_1$ individuos divididos en dos grupos de estudio, un grupo de tratamiento estándar (N_0 sujetos) y otro de tratamiento combinado con un nuevo producto (N_1 sujetos); o bien definidos a partir de los valores de una variable basal. La variable de respuesta es cuantitativa; digamos una medición de laboratorio clínico como presión arterial sistólica o hemoglobina; o bien, en el caso de una afección respiratoria, un medidor cuantitativo de la función pulmonar como el índice de oxigenación. Por cada paciente se planifica colectar mediciones de la variable de respuesta en T momentos predefinidos de tiempo. El estudio se dice entonces longitudinal. Se define el perfil de un individuo hipotético i como un vector de T componentes $Y_i = (Y_{i1}, \ldots, Y_{iT})$ donde cada componente Y_{ij} representa la medición de la variable de respuesta en la ocasión j.
Interesa evaluar el efecto del grupo en la variable de respuesta y un efecto temporal, es decir, evolución de la respuesta en el tiempo. Por ello se asume que la respuesta depende linealmente de una covariable indicadora de grupo G_i, un efecto temporal t_j, e interacción grupo-tiempo. La correlación entre mediciones de un mismo individuo en el tiempo puede capturarse de dos modos:

1. Asumiendo que la fuente de la dependencia radica en un proceso externo al sujeto.

2. Asumiendo que la fuente de la dependencia radica en un proceso estocástico operante dentro del sujeto.

El primer caso se formula mediante un modelo de interceptos aleatorios y el segundo, mediante un modelo de transición de primer orden. Ambos modelos se describen analíticamente como sigue.

Modelo de interceptos aleatorios

$$Y_{ij} = \beta_1 + \beta_2 G_i + \beta_3 t_j + \beta_4 G_i t_j + b_i + \varepsilon_{ij}$$

$b_i \sim N(0, d^2), \varepsilon_{ij} \sim N(0, \sigma^2), b_i, \varepsilon_{ij}$ independientes (1)

En estas notaciones G_i, variable indicadora de grupo, toma los valores 0 y 1; la variable temporal t_j toma el valor j/T, $1 \leq j \leq T$, j entera; el vector de parámetros es $\theta = (\beta_1, \beta_2, \beta_3, \beta_4, d^2, \sigma^2)$; la función de verosimilitud responde a una distribución normal multivariada proveniente de contribuciones individuales independientes (Verbeke and Molenberghs 2000).

Modelo de transición de primer orden

$$Y_{ij} = \beta_1 + \beta_2 G_i + \beta_3 t_j + \beta_4 G_i t_j + \delta_{ij}$$

$\delta_{ij} = \alpha \delta_{i,j-1} + z_{ij}, \delta_{ij} \sim N(0, \lambda^2), z_{ij}$ independiente de $\delta_{i1}, \ldots, \delta_{i,j-1}$ (1)

El vector de parámetros es $\theta = (\beta_1, \beta_2, \beta_3, \beta_4, \alpha, \lambda^2)$. El modelo suele denominarse *AR(1)*, abreviatura de "*Auto-regresivo de primer orden*" y la función de verosimilitud responde a una distribución normal multivariada (Verbeke 2005).

En los ensayos clínicos son de especial interés ambos modelos. Al modelo a simular, (1) ó (1), se le denominará en lo que sigue *modelo de ajuste*.

2.2. Mecanismos de pérdidas

Rubin (1976) introduce el concepto de mecanismo de pérdidas, abriendo así el camino hacia un estudio formal del problema de datos faltantes. La taxonomía de Rubin consta de tres categorías: pérdidas completamente aleatorias (MCAR), pérdidas aleatorias (MAR) y pérdidas no aleatorias (MNAR). Para describirlas se asume que, con las notaciones introducidas en 2.1, $Y_i = (Y_{i1}, \ldots, Y_{iT})$ denota el perfil de observaciones temporales del individuo i. En el vector indicador de pérdidas del individuo i, $R_i = (R_{i1}, \ldots, R_{iT})$, la componente $R_{ij} = 1$ si Y_{ij} se observa, $R_{ij} = 0$ si no. Se define entonces:

Tipo de mecanismo de pérdidas	**Se define cuando**
MCAR	$P(R_i \mid Y_i) = P(R_i)$
MAR	$P(R_i \mid Y_i) = P(R_i \mid Y_{iO})$
MNAR	$P(R_i \mid Y_i) \neq P(R_i \mid Y_{iO})$

Y_{iO} = Subvector de componentes observadas

Verbalmente, el mecanismo es completamente aleatorio si la probabilidad de pérdidas no depende de las mediciones; es aleatorio si la probabilidad depende de lo observado; y es no aleatorio si la probabilidad depende de lo no observado. Por ejemplo, supongamos que en un ensayo clínico se mide presión arterial sistólica (PAS) en enero y febrero. Se tiene:

- MCAR: Hay datos faltantes aleatorios en el mes de febrero, que no guardan relación con el nivel de PAS en enero o febrero u otra variable del estudio. Los datos faltantes constituyen una muestra aleatoria de los datos originales.
- MAR: Hay datos faltantes en el mes de febrero porque la medición de enero no excedía 140. Los datos faltantes son aleatorios dentro del grupo "PAS en enero≤140".

- MNAR: Hay datos faltantes en el mes de febrero porque los casos que presentaron la medición de la PAS de febrero menor o igual que 140 no fueron registrados. (PAS realizada, pero no recogida si era $\leq$140.)

Una clasificación alternativa distingue entre patrón monótono e intermitente. Cuando los valores de una respuesta repetida en el tiempo se observan en su totalidad, o bien se observan parcialmente pero de manera tal que a partir del primer valor no observado tampoco se observan los valores sucesivos, se habla de un patrón de pérdidas de tipo monótono. Si por el contrario, la secuencia de valores faltantes es arbitraria, entonces el patrón de pérdidas se dice de tipo intermitente.

3. TRATAMIENTO DE DATOS FALTANTES Y AJUSTE DEL MODELO

Para la estimación bajo información incompleta pueden considerarse tres estrategias:

1. Eliminación
2. Imputación
3. Optimización

Entre las estrategias de eliminación se incluye el *método de casos completos*. En un escenario longitudinal (ver sección 2.1), el análisis se limita a los individuos con todo el perfil observado. Es conocido que, si el mecanismo de pérdidas es MCAR, con el empleo de este método se obtienen estimadores insesgados para estructuras de medias (Verbeke 2005, Buuren 2012).

La estrategia de imputación conlleva un cálculo previo a la de estimación de parámetros. Métodos elementales de imputación son el de sustituir los datos faltantes por *cero*, la sustitución de cada dato faltante por que le precede (*método de la última observación llevada adelante*) y el *método de sustitución por la media*, en que cada observación faltante se sustituye por la media de ciertos valores observados de la variable en cuestión. En este trabajo se distingue el caso en que se promedian individuos de un mismo grupo, *sustitución por medias agrupadas*.

Métodos más refinados de imputación son el de *sustitución por media condicional*, *regresión* y *regresión estocástica* (Buck 1960, Buuren 2012). En la imputación por regresión la observación actual se expresa como combinación lineal de covariables u observaciones previas, usando la estructura del modelo generador de los datos. La regresión estocástica agrega una componente aleatoria a esta dependencia. Para una descripción completa de estos métodos, véase Buck (1960), Rubin (1976), Verbeke and Molenberghs (2000), Buuren (2012), Carpenter and Kenward (2013).

Se experimenta también la inclusión de una componente aleatoria al método de media condicional. En este caso, en lugar de sustituir los datos faltantes por la media condicional de la distribución normal multivariada de las componentes faltantes dadas las observadas (Verbeke and Molenberghs 2000), se genera una instancia aleatoria de esa distribución. De manera que se aprovecha no sólo la información contenida en la media condicional, sino también en la matriz condicional de varianzas y covarianzas.

Las estrategias 1 y 2 contienen una segunda etapa posterior al tratamiento de las pérdidas: la estimación de parámetros. La estrategia 3 consta de esta única etapa y se diferencia de la 1 y 2 en que no altera previamente la información disponible. La estimación se obtiene en todos los casos mediante el ajuste del modelo seleccionado por el método de máxima verosimilitud (Verbeke and Molenberghs 2000, Verbeke 2005, Uranga and Molenberghs 2014). Se espera que la estrategia 3 se comporte de manera óptima bajo MAR (Verbeke and Molenberghs 2000).

3.1. Simulación de pérdidas en datos longitudinales

Dada una población y los valores de sus mediciones siguiendo el modelo de ajuste, se aplican distintos patrones de pérdida de información. Se aplican los métodos de eliminación, imputación y optimización antes descritos para lidiar con los datos faltantes. Con la información que resulta se obtiene el estimador $\hat{\theta}$ del vector θ de los parámetros del modelo. El desempeño de los métodos es evaluado a partir de la calidad de la estimación.

La simulación consta por tanto de cinco procesos principales:

1. Selección de un tamaño de muestra.
2. Generación de muestras aleatorias a partir del modelo de ajuste.
3. Selección del mecanismo de pérdidas y generación de estas, para lo que se requiere el control de la proporción de datos faltantes.

4. Aplicación de la estrategia de tratamiento de los datos faltantes. Estimación del vector θ.
5. Evaluación de la estimación mediante distintos criterios.

Prefijados un tamaño de muestra y un modelo de ajuste, llamaremos *escenario* a la especificación de un mecanismo de pérdidas, una proporción de datos faltantes y una estrategia de tratamiento de pérdidas. Dado un escenario, los procesos 2, 3 y 4 constituyen una iteración que se repite I veces. El proceso 5 se realiza sólo una vez, al finalizar las I iteraciones. A continuación se discuten los 5 pasos de manera detallada.

3.2. Selección de un tamaño de muestra

Este paso es elemental y se refiere a fijar los valores de N, N_0 y N_1 (ver sección 2.1). Note que es muy diferente lo afirmado aquí al clásico problema de justificación de un tamaño de muestra en un ensayo clínico real. La diferencia radica en que en el presente trabajo el ensayo clínico es hipotético, *simulado*. En lo que concierne al estudio de simulación, la selección del tamaño de la muestra es en principio *arbitraria* y se trata precisamente, de evaluar las propiedades de tal elección.

3.3. Generación de muestras aleatorias a partir del modelo de ajuste

Se simula la pérdida de información en un ensayo clínico sobre una muestra aleatoria integrada por $N = N_0 + N_1$ individuos. La muestra es construida mediante un generador de números pseudo-aleatorios a partir del modelo asumido. Se modela la situación en que N_0 individuos siguen el tratamiento propuesto y por tanto se les asigna el valor $G_i = 0$ y a los otros N_1 el valor $G_i = 1$; alternativamente, la agrupación puede definirse a partir de los valores de una variable basal. A continuación se generan pérdidas sobre la muestra.

3.4. Selección del mecanismo de pérdidas y generación de estas. Control de la proporción de pérdidas

Se generan datos faltantes siguiendo uno de los siguientes mecanismos de pérdidas: intermitente de tipo completamente aleatorio, monótono de tipo aleatorio, monótono no aleatorio logístico, monótono no aleatorio por punto de corte e intermitente no aleatorio.

En el mecanismo intermitente completamente aleatorio, prefijado un porciento de datos faltantes, se eliminan los datos de la base correspondientes a posiciones generadas a partir de una distribución uniforme.

El mecanismo intermitente no aleatorio se basa en establecer un valor umbral o de corte de la variable de respuesta. Se eliminan las observaciones por debajo del valor umbral. El valor umbral o de corte se fija de modo que se garantice la proporción de pérdidas deseada.

Los mecanismos monótono de tipo aleatorio y monótono no aleatorio logístico se basan en el modelo de regresión logística descrito en Diggle and Kenward (1994). Se asume que la probabilidad de abandono en la ocasión j ($j = 1, \ldots, T$), dado que el sujeto se encontraba aún en el estudio en la ocasión anterior, satisface el modelo de regresión logística

$$logit[P(D_i = j \mid D_i \geq j, y_i)] = \psi_0 + \psi_1 y_{ij} + \psi_2 y_{i,j-1}. \quad (2)$$

Para garantizar pérdidas aleatorias se impone la restricción $\psi_0 = \psi_1 = 0$ y se deja el parámetro ψ_2 libre; para pérdidas no aleatorias se impone $\psi_0 = \psi_2 = 0$ y se deja ψ_1 libre. El mecanismo monótono no aleatorio por punto de corte se obtiene sustituyendo el miembro izquierdo de

(2) por una constante ("punto de corte"), haciendo $\psi_0 = \psi_2 = 0$ y dejando ψ_1 libre. Para garantizar la proporción de pérdidas deseada, se genera una muestra de 10 bases a partir del modelo de ajuste y se mueve el parámetro libre en una u otra dirección hasta alcanzar la proporción.

4. APLICACIÓN DE LA ESTRATEGIA DE TRATAMIENTO DE LOS DATOS FALTANTES. ESTIMACIÓN DEL VECTOR θ.

El simulador posibilita la aplicación de las estrategias de tratamiento de los datos faltantes descritas en la sección 0 sobre una base de datos, cualquiera sea el mecanismo de pérdida de la información que ésta presente tratados en la segunda sección del trabajo. En dependencia de la estrategia utilizada, se obtiene el estimador de los parámetros de interés (θ).

4.1. Criterios para la evaluación de la estimación obtenida

Se asume conocido el vector θ que describe el modelo original. Cada una de las muestras generadas a partir del modelo de ajuste, se mutila siguiendo diferentes mecanismos de pérdidas y se aplica un método de tratamiento a ese mecanismo. Se obtiene entonces una estimación de θ. Se trata ahora de fijar criterios para evaluar la calidad de la estimación obtenida.
Para ello se compara el vector estimado con el original siguiendo dos criterios: sesgo y cobertura. El sesgo se define como la diferencia entre el valor original y el estimado; estas diferencias se promedian a través de las I iteraciones. La cobertura se define como sigue. En cada iteración se genera, junto con la estimación $\hat{\theta}$, un intervalo confidencial del 95% para cada componente de $\hat{\theta}$. La cobertura asociada a cada componente se define como la proporción con que el intervalo cubre al valor original a través de las I iteraciones. Note que la *cobertura* constituye una estimación de la *cobertura real*, definida como el valor límite de la cobertura cuando el número de iteraciones I crece indefinidamente; y a su vez la cobertura real difiere de la *cobertura nominal*, que en el presente caso es 95% y viene dada por el percentil 95 de la distribución normal, empleado en la construcción del intervalo. La cobertura real debe aproximarse a la nominal cuando el tamaño de muestra N crece suficientemente y en ausencia de datos faltantes; o bien en presencia de una proporción acotada de datos faltantes tratados adecuadamente, por ejemplo, mediante el método de máxima verosimilitud y ante un mecanismo MAR.
A continuación se calcula el sesgo máximo en valor absoluto y la cobertura mínima, con respecto a todas las componentes de $\hat{\theta}$ (caso identificado como *estructura general*), o bien con respecto a los coeficientes $\beta_1, \beta_2, \beta_3, \beta_4$ solamente (caso identificado como *estructura de medias*). Si el sesgo máximo no supera 0.05 se asigna clasificación excelente (E) según este criterio; bien (B), si no supera 0.1 y no clasifica como E; regular (R), si no supera 0.2 y no clasifica como E ni B; mal (M) en otro caso. Se genera una clasificación análoga para la cobertura mínima, favoreciendo valores altos, y tomando como puntos de corte 0.9, 0.8 y 0.7. Se arriba a una clasificación global según el criterio (Figura 1):

- ***E***: Sesgo=E y Cobertura=E
- ***B***: Sesgo=E ó B, Cobertura =E ó B, y no clasifica como excelente global
- ***R***: Sesgo=E ó B ó R, Cobertura =E ó B ó R, y no clasifica como excelente ni bien global
- ***M***: Otro caso

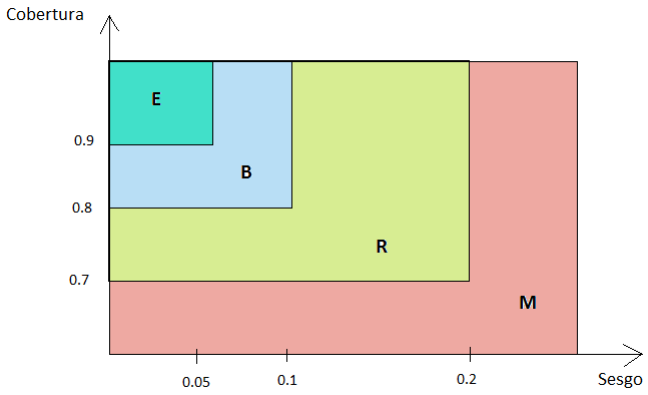

Figura 1. Clasificación cualitativa de la calidad de una estimación

5. Resultados y Discusión

5.1. Simulador

El simulador se implementa en el lenguaje IML (Interactive Matrix Language) de SAS (Statistical Analysis System) versión 9.3. Se programan sobre ese lenguaje los métodos de regresión y regresión estocástica. También los algoritmos para el control de pérdidas bajo los diferentes mecanismos. Se utiliza la instrucción PROC MIXED para el ajuste de los modelos longitudinales.

5.2. Poblaciones hipotéticas

Se consideran dos problemas.
Problema 1. Asume que:
1. El efecto de grupo supera al efecto temporal.
2. La correlación entre mediciones sucesivas se explica por un proceso externo al sujeto.
Para garantizar la condición 1 se elige $\boldsymbol{\beta} = (10,9,4,4)$; para la 2 se elige el modelo de interceptos aleatorios (1) con parámetros $d^2 = 1, \sigma^2 = 2$. La correlación entre mediciones sucesivas es 0.33.
Problema 2. Asume que:
1. El efecto temporal supera al efecto de grupo.
2. La correlación entre mediciones sucesivas se explica por un proceso operante dentro del sujeto.
Para garantizar la condición 1 se elige $\boldsymbol{\beta} = (9,4,8,3)$; para la 2 se elige el modelo de transición (1) con parámetros $\alpha = 0.6, \lambda^2 = 2$. La correlación entre mediciones sucesivas es 0.6.

5.2. Escenarios simulados

La Tabla 2 describe los 20 escenarios simulados en este trabajo. Como se explica en la Sección 0, cada escenario se determina por el mecanismo generador de pérdidas a simular, la proporción de pérdidas seleccionada y el método de tratamiento de datos faltantes a evaluar (columnas Método, Mecanismo y Proporción de la Tabla 2). Por ejemplo, el escenario # 18 concierne al método de casos completos bajo un 35% de pérdidas aleatorias. El tamaño de muestra seleccionado para la implementación del paso 1 del proceso de simulación de cada escenario (ver Sección 0) es $N_0 = N_1 = 25$ individuos por grupo (Sección 0), para una muestra total de $N = 50$ individuos hipotéticos. La longitud del perfil es $T = 3$ (Sección 2.1). El número de iteraciones de los pasos 2, 3 y 4 (Sección 0) se fija en $I = 500$.
Cada escenario se evalúa en el contexto de los problemas 1 y 2. La figura 2 contiene los resultados generados por los 20 escenarios al ser evaluados en el contexto del problema 1, y a ella hace referencia la sección 0. La figura 3 contiene los resultados generados por los 20 escenarios al ser evaluados en el contexto del problema 2, y a ella hace referencia la sección 0. Las tablas A1 y A2 del anexo contienen las salidas numéricas correspondientes.

Tabla 2
Escenarios simulados en los Problemas 1 y 2

Escenario	Método	Mecanismo	Proporción	Escenario	Método	Mecanismo	Proporción
1	cero	mcar	20	11	cc	mcar	20
2	locf	mcar	20	12	cc	mar	20
3	mean	mcar	20	13	ml	mar	20
4	gmean	mcar	20	14	ml	mnar	20
5	gmean	mar	20	15	ml	cutmnari	20
6	cmean	mar	20	16	ml	cutmnar	20
7	cmeane	mar	20	17	cmeane	mar	35
8	reg	mcar	20	18	cc	mar	35
9	regx	mar	20	19	cc	mar	50
10	rege	mar	20	20	ml	mar	50

cero = imputación por cero; locf = última observación llevada adelante; mean = imputación por la media; gmean = imputación por la media distinguiendo covariable grupo; cmean = imputación por la media condicional; cmeane = imputación por la media condicional con componente aleatoria; reg = regresión sin covariables (incluye sólo variable dependiente previa); regx = regresión con covariables; rege = regresión estocástica; cc = casos completos; ml = máxima verosimilitud; mcar = pérdidas completamente aleatorias (missing at random); mar = pérdidas aleatorias; mnar = pérdidas no aleatorias; cutmnari = pérdidas no aleatorias intermitentes por punto de corte; cutmnar = pérdidas no aleatorias monótonas por punto de corte; método = método de tratamiento de datos faltantes; mecanismo = mecanismo generador de las pérdidas; proporción = proporción de pérdidas

5.3. Resultados: Problema 1

Esta sección hace referencia a la figura 2.

Estructura de medias

Los escenarios 1, 2, 3, 8, 16, 19 reciben calificación de Mal. Ellos son: imputación por cero, última observación llevada adelante, media no agrupada, regresión sin covariables, máxima verosimilitud bajo MNAR monótono por punto de corte, y casos completos bajo MAR con 50% de pérdidas. Los cuatro primeros escenarios se generan bajo MCAR con 20% de pérdidas. Esto ilustra la debilidad de los cuatro primeros métodos, que podemos llamar "imputaciones ingenuas"; el peligro de acudir a máxima verosimilitud cuando se viola fuertemente el supuesto MAR; y la falta de validez del método de casos completos cuando los porcientos de pérdida son altos.
Los escenarios 13, 15 y 20 reciben calificación de Excelente. Ellos son máxima verosimilitud bajo MAR con 20% de pérdidas, bajo MNAR intermitente por punto de corte con 20% y bajo MAR con 50% de pérdidas, respectivamente. Se confirma la fortaleza de ML bajo MAR, aún cuando los porcientos de pérdidas son elevados, y su robustez ante ciertas desviaciones hacia MNAR. El escenario 4, medias agrupadas, tiene excelente comportamiento en términos de sesgo y muy altas coberturas; por tanto es un método totalmente admisible bajo MCAR y bajas proporciones de pérdidas; bajo MAR (escenario 5) los sesgos abandonan la zona de excelencia y las coberturas disminuyen; este comportamiento se corrige por el método de medias condicionales (escenarios 6 y 7), donde los sesgos recuperan la zona de excelencia. La anomalía del escenario 8 se corrige al incluir covariables en el método de regresión (escenarios 9 y 10). Los escenarios 11 y 12 muestran que CC cede ante MAR (sesgos más altos que en MCAR) aunque se mantiene robusto (sesgos en la zona de Bien). Un comportamiento similar lo muestra ML bajo MNAR monótono logístico, escenario 14, confirmando la pertinencia del supuesto MAR. Media condicional con moderadas proporciones de pérdidas (escenario 17) cede en las coberturas pero se mantiene robusto en los sesgos. Casos completos con moderadas proporciones de pérdidas (escenario 18) cede en los sesgos; su aparente robustez en cuanto a coberturas se alcanza probablemente a costa de intervalos confidenciales amplios.

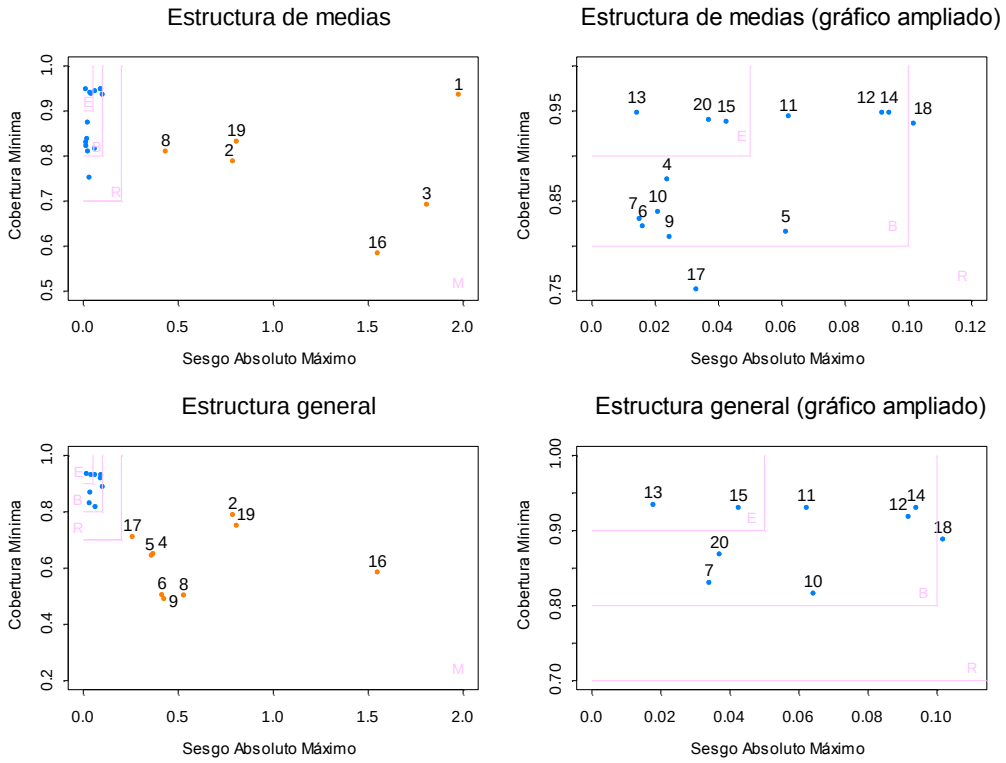

Figura 2. Clasificación cualitativa de los 20 escenarios. Problema 1.

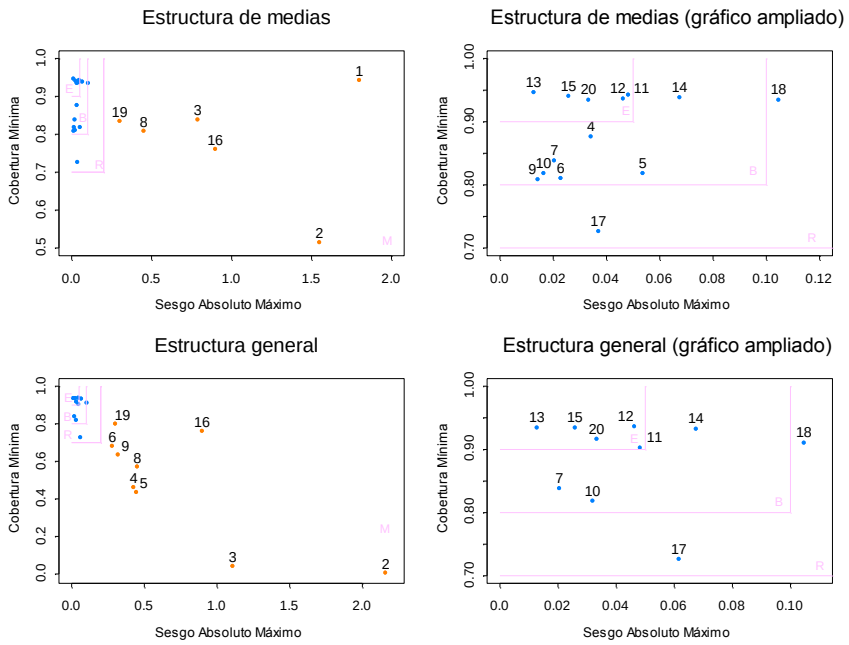

Figura 3. Clasificación cualitativa de los 20 escenarios. Problema 2.

Estructura general

A los escenarios que calificaron mal en la estimación de los coeficientes de la estructura de medias se suman ahora 4, 5, 6, 9 y 17. Medias agrupadas bajo MCAR (escenario 4), que clasificaba de manera casi excelente, cede ahora totalmente, mostrando que este método no procede si el objetivo del estudio incluye la estimación de los coeficientes
de la estructura de covarianzas, o sea d^2, σ^2 para el modelo
(1) y α, λ^2 para el modelo (1), porque produce sesgos muy altos y coberturas muy bajas. Medias agrupadas bajo MAR (escenario 5) cede también. Media condicional y regresión (escenarios 6 y 9) tampoco proceden si se pretenden estimar parámetros de covarianzas. Media condicional con componente aleatoria incluida (escenario 17) no procede ante proporciones de pérdidas moderadas (35%).
Sólo los escenarios 13 y 15, ML bajo MAR 20% y bajo MNAR intermitente por punto de corte 20%, obtienen calificación de excelente; ML bajo MAR 50% (escenario 20) roza la zona de excelencia. Media condicional con componente aleatoria incluida (escenario 7) roza también la zona de excelencia, de manera que se logra corregir el defecto del escenario 6. Regresión estocástica (escenario 10) corrige el defecto de regresión no estocástica (escenario 9) que había clasificado como Mal, al insertarse en la zona de Bien. Los escenarios 11, 12, 14 y 18 mantienen un comportamiento idéntico al ya descrito en la subsección *Estructura de medias*.

5.4. Resultados: Problema 2

Esta sección hace referencia a la figura 3.

Estructura de medias

El comportamiento de los escenarios es cualitativamente idéntico al problema 1, con la única excepción de los escenarios 11 y 12 que reciben ahora calificación de Excelente. Ellos son casos completos bajo MCAR y MAR con 20% de pérdidas, evidenciando la robustez de CC ante MAR para bajos porcientos de pérdidas. La interpretación de los escenarios restantes coincide con la expuesta en el problema 1.

Estructura general

El comportamiento es idéntico al descrito en el problema 1, con las siguientes excepciones:
- El escenario 17, media condicional con componente aleatoria incluida y 35% de pérdidas, que calificaba como Mal en el problema 1, ahora califica como Regular.
- Los escenarios 11, 12 y 20, que calificaban como Bien, ahora califican como Excelente. Ellos son casos completos bajo MCAR y MAR con 20% de pérdidas, y ML bajo MAR con 50% de pérdidas. Recordemos que los escenarios 13 y 15, que mantienen comportamiento excelente, son ML bajo MAR con 20% de pérdidas y ML bajo MNAR intermitente por punto de corte con 20% de pérdidas.
- El escenario 14, ML bajo MNAR monótono con 20% de pérdidas, califica nuevamente como Bien, pero en el problema 1 se acercaba a la zona Regular con sesgos moderados, y ahora se acerca a la zona Excelente con sesgos menores.

5.5. Caso de estudio

Ilustremos nuestros hallazgos mediante un caso de estudio. En 2014 se presentan los resultados finales de un ensayo clínico sobre un medicamento cubano destinado a tratar la anemia asociada a enfermedad renal crónica en prediálisis. Una de las variables de interés es nivel de hemoglobina, con valores altos indicando buen estado del paciente (ausencia de anemia) y valores bajos, mal estado. Se realizan 13 mediciones por individuo en el tiempo, una basal y 12 mensuales, pero nos limitamos a modelar las mediciones de los meses 2, 9 y 12. Se dispone de una muestra de tamaño 184, que se divide en dos grupos según nivel de hemoglobina inicial alto (Hb ≥ 10g/dL; 82 pacientes, grupo A) o bajo (Hb < 10g/dL; 102 pacientes, grupo B). El interés se centra en la detección de un posible efecto de grupo en la variable de interés, que favorezca a los pacientes con nivel inicial de hemoglobina bajo, y en la obtención de conclusiones sobre la evolución de esta variable en el tiempo.
Las proporciones de pérdidas son 38% a nivel individual (70 pacientes no reportan al menos una medición) y 22% a nivel de medición (de las $184 \cdot 3 = 552$ mediciones planificadas, no se reportan 121). La figura 4 muestra la evolución promedio en el tiempo del nivel de hemoglobina por grupo de estudio, aplicando, con carácter exploratorio, el método de casos disponibles por su cómoda implementación. Visualmente, el perfil del grupo B es ascendente, mientras que el del grupo A primero asciende y luego desciende levemente.
La Tabla 3 muestra los resultados de aplicar varios métodos de tratamiento de datos faltantes. Se parte del modelo (1) como inicial, tomando como Y la variable nivel de hemoglobina. La codificación de la variable grupo es 1 para el A y 0 para el B. La longitud del perfil es $T = 3$. Se elige este modelo, en consonancia con el Problema 2, porque la variable de interés se genera por un proceso operante dentro del individuo.
Nuestro objetivo es comparar el comportamiento de los métodos, pero no estamos ya en una simulación, sino en un caso de estudio. La diferencia radica en que en una simulación se conoce el valor original de los parámetros, mientras que en una situación clínico-práctica se desconocen. Puede resolverse el inconveniente tomando como patrón de referencia el método de máxima verosimilitud, ya que en el estudio de simulación es el método favorecido, con estimaciones reportadas muy cercanas a los valores originales de los parámetros.
Para ello se ha agregado una fila inferior a la Tabla 3, donde se tabula la estimación $\hat{\delta}$ de cierto indicador δ de la distancia de cada método M al método de máxima verosimilitud ML. Formalmente se tiene:
1. Vector de parámetros originales: $\theta = (\beta_1, \beta_2, \beta_3, \beta_4, \alpha, \lambda^2)$.
2. Vector de parámetros asociados al método ML: $\theta_{ML} = \theta$ porque se asume que ML reporta estimadores consistentes.
3. Vector de parámetros asociados al método M: $\theta_M = (\beta_1^M, \beta_1^M, \beta_1^M, \beta_1^M, \alpha_M, \lambda_M^2)$. Se tiene $\theta_M \neq \theta$ porque el método M no reporta en general estimadores consistentes.
4. Estimador reportado por el método ML: $\hat{\theta}_{ML} = \left(\hat{\beta}_1^{ML}, \hat{\beta}_1^{ML}, \hat{\beta}_1^{ML}, \hat{\beta}_1^{ML}, \hat{\alpha}_{ML}, \hat{\lambda}_{ML}^2\right)$.
5. Estimador reportado por el método M: $\hat{\theta}_M = \left(\hat{\beta}_1^M, \hat{\beta}_1^M, \hat{\beta}_1^M, \hat{\beta}_1^M, \hat{\alpha}_M, \hat{\lambda}_M^2\right)$.
6. Definición del indicador δ:

$$\delta = d(\theta, \theta_M) = \sqrt{(\beta_1 - \beta_1^M)^2 + (\beta_2 - \beta_2^M)^2 + (\beta_3 - \beta_3^M)^2 + (\beta_4 - \beta_4^M)^2 + (\alpha - \alpha_M)^2 + (\lambda^2 - \lambda_M^2)^2}$$

7. Definición del estimador $\hat{\delta}$ que se tabula:

$$\hat{\delta} = d\left(\hat{\theta}_{ML}, \hat{\theta}_M\right)$$

$$\hat{\delta} = \sqrt{\left(\hat{\beta}_1^{ML} - \hat{\beta}_1^M\right)^2 + \left(\hat{\beta}_2^{ML} - \hat{\beta}_2^M\right)^2 + \left(\hat{\beta}_3^{ML} - \hat{\beta}_3^M\right)^2 + \left(\hat{\beta}_4^{ML} - \hat{\beta}_4^M\right)^2 + (\hat{\alpha}_{ML} - \hat{\alpha}_M)^2 + \left(\hat{\lambda}_{ML}^2 - \hat{\lambda}_M^2\right)^2}$$

A grandes rasgos, dado un método M, se define el vector de parámetros asociado θ_M como el límite del estimador $\hat{\theta}_M$ cuando el tamaño de la muestra crece indefinidamente. Si $\theta_M = \theta$, el estimador $\hat{\theta}_M$ se dice *consistente*; si $\theta_M \neq \theta$, el estimador $\hat{\theta}_M$ se dice in*consistente*. El indicador δ constituye una guía acerca de la bondad del método M, asumiendo que el método ML es correcto, o sea, asumiendo que el mecanismo generador de pérdidas es MAR y por tanto, ML reporta estimadores consistentes. Note que en la Tabla 3 M es uno de los métodos siguientes: ml, mean, gmean, cmean, cmeane, reg, regx, rege, cc, locf.
Pasando comparar los métodos según el indicador estimado $\hat{\delta}$, se tiene en primer lugar $\hat{\delta} = 0$ para ML, como es obvio. El método más aberrado es locf, con un indicador que supera la unidad. Imputación por la media (mean) tiene el indicador más alto entre los métodos restantes; medias agrupadas (gmean) supera a mean, con un indicador algo más bajo; media condicional (cmean) produce una disminución mayor; y media condicional con componente aleatoria (cmeane) casi divide a la mitad el indicador de mean.

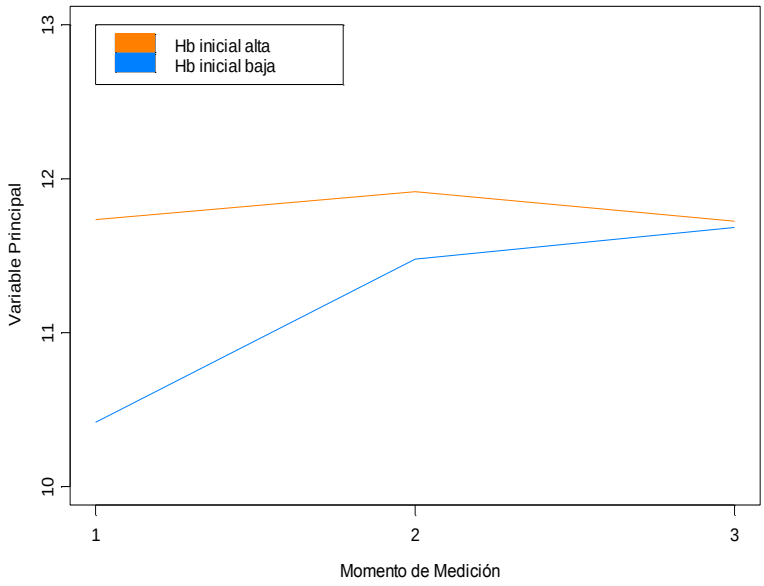

Figura 4. Perfiles promedio de la variable hemoglobina según valor inicial.

Regresión sin covariables (reg) produce un indicador alto que lo coloca en tercer lugar en insuficiencia, detrás de locf y mean; regresión con covariables (regx) produce una leve mejora al reducir el indicador $\hat{\delta}$; regresión estocástica (rege) produce una mejora sustancial, al extremo de ser el método de indicador más bajo (después de ml); finalmente, casos completos ocupa el lugar 4 entre los mejores métodos, después de ml, rege y cmeane.
La concordancia con el estudio de simulación es total

Tabla 3
Ajuste de un modelo de transición a los datos del estudio de hemoglobina

	Método estimación (error estándar)				
Parámetro	ml	mean	gmean	cmean	cmeane
β_1	9,91 (0,21)	9,94 (0,18)	9,93 (0,18)	9,92 (0,18)	9,88 (0,2)
β_2	1,75 (0,31)	1,75 (0,27)	1,75 (0,27)	1,74 (0,27)	1,76 (0,29)
β_3	1,75 (0,29)	1,87 (0,24)	1,85 (0,24)	1,82 (0,23)	1,96 (0,25)
β_4	-1,62 (0,44)	-1,79 (0,36)	-1,73 (0,36)	-1,67 (0,34)	-1,73 (0,38)

α	0,48 (0,05)	0,34 (0,05)	0,34 (0,05)	0,48 (0,04)	0,46 (0,04)
λ^2	1,95 (0,15)	1,51 (0,1)	1,5 (0,1)	1,57 (0,11)	1,81 (0,13)
δ	0,00	0,50	0,48	0,39	0,27

Parámetro	reg	regx	rege	cc	locf
β_1	9,92 (0,18)	9,93 (0,18)	9,9 (0,2)	10,11 (0,23)	10,12 (0,2)
β_2	1,73 (0,27)	1,73 (0,27)	1,7 (0,3)	1,58 (0,37)	1,56 (0,29)
β_3	1,79 (0,23)	1,78 (0,23)	1,86 (0,26)	1,76 (0,3)	1,05 (0,23)
β_4	-1,6 (0,34)	-1,62 (0,34)	-1,57 (0,39)	-1,6 (0,47)	-1,01 (0,34)
α	0,48 (0,04)	0,48 (0,04)	0,42 (0,04)	0,45 (0,06)	0,69 (0,03)
λ^2	1,55 (0,11)	1,56 (0,11)	1,83 (0,12)	1,7 (0,15)	2,2 (0,18)
δ	0,40	0,39	0,18	0,36	1,03

Dado un método M de tratamiento de pérdidas, el indicador δ se estima por la distancia euclidiana entre los vectores $\hat{\theta}_M$ y $\hat{\theta}_{ML}$, compuestos por las estimaciones respectivas del vector de parámetros θ bajo el método M y bajo el método de máxima verosimilitud.

6. CONCLUSIONES

El estudio de simulación realizado permite extraer un grupo de conclusiones importantes.

1. Los métodos de imputación por cero y última observación llevada adelante son totalmente incompatibles con un modelo longitudinal. Producen estimadores muy sesgados.
2. En imputación por la media deben promediarse individuos similares; o sea, debe promediarse condicionando a valores dados de toda covariable que se considere influyente. En tal caso, el método funciona sólo bajo MCAR y sólo para los coeficientes correspondientes a la estructura de medias del modelo longitudinal.
3. El método de media condicional corrige a imputación por la media, en el sentido de que tiende a producir estimadores consistentes para los coeficientes de la estructura de medias bajo un mecanismo MAR. Sin embargo no debe emplearse para estimar coeficientes de la estructura de covarianzas del modelo.
4. Si se añade una componente aleatoria al método de media condicional y el porciento de pérdidas es bajo (hasta un 20%) se logran muy buenos estimadores tanto para los coeficientes de la estructura de medias como para los de la estructura de covarianzas. Si la proporción de pérdidas es moderada (35%) el método no se recomienda.
5. Si se emplea el método de imputación por regresión, es preciso incluir todas las covariables que se consideren influyentes. En tal caso el método proporciona buenos estimadores bajo MAR sólo para los coeficientes de la estructura de medias.
6. El método de regresión estocástica corrige el defecto de regresión no estocástica, ya que produce buenos estimadores para todos los coeficientes bajo MAR cuando los porcientos de pérdidas son bajos (20%). Aunque no se reporta en este trabajo, el método se muestra robusto ante porcientos de pérdidas de 35% y para 50%, se mantiene aceptable sólo para los coeficientes de la estructura de medias.
7. El método de casos completos es totalmente aceptable, tanto para los coeficientes de la estructura de medias como para los de covarianza, bajo MCAR o MAR si los porcientos de pérdidas son bajos (20%). No se recomienda para porcientos de pérdidas moderados o altos.
8. Máxima verosimilitud es el método favorito. Funciona de manera excelente, tanto para coeficientes de medias como de covarianzas, bajo un mecanismo MAR con proporciones de pérdidas que pueden llegar al 50%. Es robusto ante desviaciones del supuesto MAR de tipo monótono que responden al mecanismo de regresión logística descrito por los autores Diggle and Kenward (1994). Sin embargo, ante mecanismos MNAR monótonos definidos a partir de un punto de corte, el método falla. En este caso se precisa modelar el mecanismo generador de las pérdidas.
9. El modelo de transición se muestra más noble que el de interceptos aleatorios, porque los métodos de tratamiento de datos faltantes tienen un mejor comportamiento.

En Verbeke and Molenberghs (2000) y Verbeke (2005) se critica el método LOCF mediante un contraejemplo teórico. Nuestra simulación contribuye a esclarecer la debilidad del método desde otra perspectiva. En Buuren (2012) se describen los métodos de imputación por la media, regresión y regresión estocástica; se realizan afirmaciones rápidas acerca de cuándo utilizar cada uno, limitándose a ejemplificar mediante unos pocos casos de estudio. Se dice que imputación por la media "debe emplearse como un rápido y simple método reparador de datos faltantes cuando se tiene un mínimo de pérdidas, y debe evitarse en general". Nada se dice acerca del hallazgo en nuestro estudio de la necesidad e importancia de promediar *sólo individuos similares*. Ni del buen desenvolvimiento del método para porcientos de pérdidas que pueden llegar hasta un 20%. Buuren (2012) comenta la fortaleza de regresión estocástica sobre regresión al ser capaz de preservar las correlaciones; en este sentido, la conclusión 6 precisa la afirmación de Buuren.

En Verbeke and Molenberghs (2000) y Buuren (2012) se comenta el método de casos completos, pero no se arriba a las recomendaciones precisas que reporta el presente estudio de simulación. Aún cuando cabe esperar que el método produzca estimadores en general sesgados bajo MAR, el hecho es que en nuestro estudio, para bajas proporciones de pérdidas (20%), resultó ser el método favorecido después de máxima verosimilitud, superando a media condicional y regresión estocástica.

El método de media condicional de Buck se describe en Buck (1960) y Verbeke and Molenberghs (2000). Allí se habla de la pobre estimación de los parámetros de precisión según este método. En nuestro estudio esa característica se recoge mediante las coberturas. Como señala la conclusión 3, el método se comporta de manera muy buena bajo MAR, bajos porcientos de pérdidas, y referido a los parámetros de la estructura de medias. Cierto es que las coberturas son algo bajas (pierden la categoría de Excelente al rozar el 80%); esto confirma la afirmación hecha en Verbeke and Molenberghs (2000) pero sin llegar a invalidar el método.

Es un hallazgo del presente estudio la adición de una componente aleatoria al método de media condicional de Buck (1960). Tal propuesta no se encuentra documentada en la literatura. Nuestro estudio permite conjeturar que el método modificado supera al original porque es capaz de reconocer los coeficientes de la estructura de covarianzas (producir estimadores consistentes). Dicho resultado debe estar respaldado por una formulación teóricamente demostrable.

Con respecto al método de máxima verosimilitud, las afirmaciones (teóricamente probadas) halladas en la literatura estadística acerca de su funcionamiento eficaz ante el supuesto MAR (Rubin 1976, Verbeke and Molenberghs 2000) han sido exitosamente confirmadas. Se ha detectado además que el mecanismo MNAR logístico de pérdidas monótonas de Diggle and Kenward (1994) es noble, porque ML es robusto ante él. Sin embargo un mecanismo MNAR tan simple como generar pérdidas monótonas cuando la variable dependiente no excede un valor umbral o punto de corte, resulta ser altamente nocivo para ML, llevándolo a producir estimadores muy sesgados y por tanto, invalidando el supuesto de ignorabilidad.

La aplicación al caso de estudio confirma la jerarquía en la bondad de los métodos prevista en el estudio de simulación. Debe notarse sin embargo que para que ello ocurra, primero que todo los datos deben ser compatibles con el supuesto de normalidad. Esto motivó la búsqueda de un estudio donde se procesara una variable "noble" como hemoglobina. Aunque no se muestra en el trabajo, se hicieron los gráficos y tests correspondientes. Segundo, la comparación de los métodos, tal como se ha presentado aquí, exige que se asuma el supuesto de pérdidas aleatorias (MAR), para que la elección de ML como patrón de referencia sea legítima. Existen criterios alternativos de comparación; por ejemplo puede pensarse en una adaptación del criterio de información de Akaike (Akaike 1974, Verbeke and Molenberghs 2000), o bien en definir un error cuadrático como la distancia de las respuestas observadas a sus valores estimados por el método a evaluar. Esto es motivo de próximos trabajos.

En resumen, la simulación resulta ser una técnica experimental que permite realizar la comparación concreta de diversos métodos de tratamiento de datos faltantes. Los resultados sugieren la superioridad del método de verosimilitud directa sobre los restantes, con el hallazgo de la conveniencia de métodos más simples como casos completos, media condicional con componente aleatoria incluida y regresión estocástica cuando las proporciones de pérdidas son bajas (20%). El trabajo contiene aún aspectos por desarrollar. Las investigaciones pueden extenderse a casos más generales, no necesariamente datos longitudinales, y estudiarse métodos no tratados aquí.

RECEIVED: FEBRUARY 2016
REVISED: MAY 2016

REFERENCIAS

[1] AKAIKE, H. (1974): A new look at the statistical model identification. **IEEE Transactions on automatic control,** 19, 716-723.
[2] BUCK, S. F. (1960): A method of estimation of missing values in multivariate data suitable for use with an electronic computer. **Journal of the Royal Statistical Society, Series B,** 22, 302–306.
[3] BUUREN, S. (2012): **Flexible imputation of missing data.** Chapman & Hall/CRC Press, London-New York.
[4] CARPENTER, J. R. and M. G. KENWARD (2013): **Multiple imputation and its application.** John Wiley & Sons, Ltd., Chichester, United Kingdom.
[5] CENCEC (2014): Registro Cubano de Ensayos Clínicos. Available in **http://registroclinico.sld.cu, Consulted** 1-7, 2015.
[6] DÍAZ-ORDAZ, K., M. G. KENWARD, A. COHEN, C. L. COLEMAN and S. ELDRIDGE (2014): Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. **Clinical Trials,** II(5), 590-600.
[7] DIGGLE, P. and M. G. KENWARD (1994): Informative Drop-Out in Longitudinal Data Analysis. **Journal of the Royal Statistical Society. Series C (Applied Statistics),** 43, 49-93.
[8] HAND, D. J., F. DALY, A. D. LUNN, K. J. MCCONWAY and E. OSTROWSKI (1994): **A handbook of small data sets.** Chapman & Hall, London.
[9] MARRERO, M. A. (2012): Dos productos naturales en la prevención de la migraña. Ensayos clínicos fase III. **Proceedings of the Tercer Simposio Internacional de Farmacología de productos Naturales. Topes de Collantes, Sancti Spiritus, Cuba**.
[10] RUBIN, D. B. (1976): Inference and missing data. **Biometrika,** 63, 581–592.
[11] URANGA, R. (2011): Modelos de transición en presencia de pérdidas intermitentes: implementación y ejemplo de aplicación a un ensayo clínico cubano. **VacciMonitor,** 20(2), 11–16.
[12] URANGA, R. and G. MOLENBERGHS (2014): Longitudinal conditional models with intermittent missingness: SAS code and applications. **Journal of Statistical Simulation and Imputation,** 84(4), 753-780.
[13] VERBEKE, G. (2005): **Models for Discrete Longitudinal Data.** Springer-Verlag, NewYork.
[14] VERBEKE, G. and G. MOLENBERGHS (2000): **Linear Mixed Models for Longitudinal Data.** Springer-Verlag, NewYork.
[15] WOOD, A. M., I. R. WHITE and S. G. THOMPSON (2004): Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. **Clinical trials,** 1, 368-376.

ANEXO

Se muestran a continuación las salidas numéricas del presente estudio de simulación.

Tabla A1
Sesgos y coberturas de los 20 escenarios asociados al problema 1

		Escenario									
		1	2	3	4	5	6	7	8	9	10
	Valor	Sesgos									
β_1	10	1,98	-0,22	-0,91	0,00	0,02	0,01	0,01	0,07	0,01	0,01
β_2	9	1,85	-0,20	1,81	-0,02	0,00	0,00	0,00	-0,11	0,00	-0,01
β_3	4	0,84	0,79	-0,41	0,02	-0,06	-0,02	-0,02	-0,24	-0,02	-0,02
β_4	4	0,76	0,76	0,84	0,02	0,01	0,00	-0,01	0,43	-0,01	0,01
d^2	1	-1,78	-0,52	0,28	0,37	0,36	0,07	0,03	-0,53	0,16	0,06
σ^2	2	-58,58	-0,38	-5,40	0,24	0,29	0,42	-0,03	0,42	0,43	-0,02
max		58,58	0,79	5,40	0,37	0,36	0,42	0,03	0,53	0,43	0,06
		Coberturas									
β_1	10	0,96	0,96	0,81	0,88	0,89	0,89	0,91	0,94	0,89	0,92

β_2	9	0,94	0,93	0,69	0,87	0,89	0,88	0,92	0,92	0,88	0,90
β_3	4	0,99	0,79	0,91	0,89	0,82	0,82	0,83	0,85	0,81	0,85
β_4	4	0,95	0,82	0,94	0,90	0,84	0,83	0,86	0,81	0,83	0,84
d^2	1	0,50	0,88	0,81	0,65	0,64	0,80	0,84	0,85	0,76	0,82
σ^2	2	0,00	0,89	0,00	0,77	0,68	0,50	0,86	0,50	0,49	0,87
min		0,00	0,79	0,00	0,65	0,64	0,50	0,83	0,50	0,49	0,82

		Escenario									
		11	12	13	14	15	16	17	18	19	20
	Valor	Sesgos									
β_1	10	0,00	-0,09	0,00	0,03	0,01	0,37	-0,01	0,01	-0,03	0,00
β_2	9	-0,04	0,03	0,01	-0,01	0,00	-0,36	-0,01	-0,05	0,81	-0,01
β_3	4	0,00	0,05	0,00	-0,09	-0,04	-1,55	0,01	-0,06	-0,11	0,00
β_4	4	0,06	-0,01	-0,01	0,03	0,00	1,54	0,03	0,10	0,26	0,04
d^2	1	-0,02	0,00	0,00	0,00	0,00	0,13	-0,01	-0,01	0,02	0,00
σ^2	2	0,01	0,02	0,02	0,02	0,01	0,06	-0,26	0,02	0,05	0,02
max		0,06	0,09	0,02	0,09	0,04	1,55	0,26	0,10	0,81	0,04
		Coberturas									
β_1	10	0,94	0,95	0,95	0,95	0,96	0,88	0,86	0,94	0,90	0,95
β_2	9	0,95	0,95	0,95	0,95	0,94	0,89	0,87	0,94	0,85	0,94
β_3	4	0,95	0,96	0,96	0,96	0,96	0,58	0,76	0,95	0,91	0,94
β_4	4	0,95	0,95	0,95	0,95	0,95	0,70	0,75	0,94	0,83	0,95
d^2	1	0,93	0,92	0,93	0,93	0,93	0,89	0,71	0,89	0,75	0,92
σ^2	2	0,94	0,94	0,94	0,94	0,94	0,92	0,72	0,93	0,86	0,87
min		0,93	0,92	0,93	0,93	0,93	0,58	0,71	0,89	0,75	0,87

Tabla A2

Sesgos y coberturas de los 20 escenarios asociados al problema 2

		Escenario									
		1	2	3	4	5	6	7	8	9	10
	Valor	Sesgos									
β_1	9	1,80	-0,45	-0,38	0,03	0,02	0,01	0,01	0,09	0,01	0,01
β_2	4	0,83	-0,18	0,79	-0,03	0,00	0,00	0,00	-0,14	0,00	-0,01
β_3	8	1,60	1,55	-0,34	-0,03	-0,05	-0,02	-0,02	-0,26	-0,01	-0,01
β_4	3	0,58	0,60	0,64	0,03	-0,01	0,00	-0,01	0,45	0,00	0,02
α	0,6	0,58	0,20	0,35	0,14	0,09	-0,04	0,00	-0,09	-0,05	0,00
λ^2	2	-50,36	-2,16	-1,11	0,43	0,45	0,28	0,00	-0,09	0,32	0,03
max		50,36	2,16	1,11	0,43	0,45	0,28	0,02	0,45	0,32	0,03
		Coberturas									
β_1	9	0,94	0,96	0,88	0,89	0,90	0,92	0,93	0,93	0,91	0,92
β_2	4	0,96	0,99	0,84	0,88	0,89	0,89	0,93	0,93	0,89	0,92
β_3	8	0,98	0,51	0,90	0,90	0,82	0,81	0,84	0,85	0,81	0,82
β_4	3	0,94	0,90	0,90	0,90	0,85	0,81	0,84	0,81	0,82	0,85

α	0,6	0,00	0,39	0,04	0,63	0,85	0,76	0,86	0,57	0,74	0,84
λ^2	2	0,00	0,00	0,08	0,46	0,43	0,68	0,91	0,95	0,63	0,90
min		0,00	0,00	0,04	0,46	0,43	0,68	0,84	0,57	0,63	0,82

		Escenario									
		11	12	13	14	15	16	17	18	19	20
	Valor	Sesgos									
β_1	9	0,04	-0,03	0,01	0,02	0,01	0,26	0,01	0,05	0,05	0,01
β_2	4	-0,05	-0,05	0,00	-0,01	0,00	-0,26	-0,01	-0,10	0,26	0,01
β_3	8	-0,02	-0,01	-0,01	-0,07	-0,03	-0,89	-0,04	-0,05	-0,17	-0,03
β_4	3	0,03	0,03	0,00	0,02	0,00	0,90	0,02	0,06	0,30	-0,03
α	0,6	0,01	0,00	0,00	0,00	0,00	0,02	0,03	0,01	0,10	0,00
λ^2	2	0,03	0,03	0,01	0,01	0,01	0,06	-0,06	0,02	0,10	0,02
max		0,05	0,05	0,01	0,07	0,03	0,90	0,06	0,10	0,30	0,03
		Coberturas									
β_1	9	0,96	0,94	0,95	0,95	0,95	0,89	0,88	0,93	0,88	0,95
β_2	4	0,94	0,94	0,96	0,95	0,94	0,92	0,90	0,94	0,83	0,94
β_3	8	0,95	0,94	0,95	0,94	0,94	0,76	0,73	0,95	0,89	0,95
β_4	3	0,94	0,94	0,95	0,94	0,94	0,80	0,73	0,94	0,83	0,93
α	0,6	0,93	0,95	0,93	0,94	0,93	0,93	0,78	0,94	0,91	0,92
λ^2	2	0,90	0,94	0,94	0,93	0,94	0,90	0,82	0,91	0,80	0,92
min		0,90	0,94	0,93	0,93	0,93	0,76	0,73	0,91	0,80	0,92

Observe cómo en los escenarios 11, 12, 13, 14, 15 de las tablas A1 y A2, las coberturas se acercan al valor nominal 95%. Ver comentario en sección 0.