Revista Colombiana de Estadística
Volumen 30 No. 2. pp. 265 a 285. Diciembre 2007

Estimación de datos faltantes en medidas repetidas con respuesta binaria

Estimation of Missing Data in Repeated Measurements with Binary Response

Yolima Ayala[1,a], Óscar Orlando Melo[2,b]

[1]Departamento de Matemáticas y Estadística, Universidad Pedagógica y Tecnológica de Colombia, Tunja, Colombia

[2]Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

Resumen

Se propone una metodología para la estimación de datos faltantes en condiciones longitudinales con respuesta binaria, desde una perspectiva univariada, basada en máxima verosimilitud. Suponiendo que las respuestas son faltantes de forma aleatoria (FFA), en cada una de las ocasiones se emplea el algoritmo EM de dos formas distintas: en la primera, el paso E se expresa como una log-verosimilitud ponderada de la respuesta, condicionada a las anteriores ocasiones tomadas como covariables adicionales, con base en el método de Ibrahim (1990) para covariables categóricas faltantes, obteniendo de esta forma estimadores máximo verosímiles. En la segunda, en el paso E se realiza la estimación e imputación de datos faltantes basada en el método Ancova de Bartlett (1937). La metodología propuesta es aplicada en un caso de estudio relacionado con factores de riesgo coronario, presentado en Fitzmaurice et al. (1994).

***Palabras clave*:** datos longitudinales, regresión logística, máxima verosimilitud, algoritmo EM.

Abstract

A maximum likelihood method is proposed to provide estimates for models with binary response in longitudinal data based on an univariate model. Under a missing at random (MAR) mechanism, the EM algorithm is used in two different forms: in the first, the E step can be expressed as a weighted log-likelihood responses given the previous times, based in the method of weights proposed by Ibrahim (1990), for partially missing covariates. In the second, on the E step the estimation and imputation for missing data

[a]Profesora auxiliar. E-mail: yayalas@unal.edu.co
[b]Profesor asistente. E-mail: oomelom@unal.edu.co

is based in Ancova method proposed by Bartlett (1937). Finally, we apply our method to the data from the Muscatine Coronary Risk Factor Study, employed in Fitzmaurice et al. (1994).

***Key words*:** Longitudinal data, Logistic regression, Maximum likelihood, EM algorithm.

1. Introducción

En cualquier tipo de análisis estadístico se desea hacer inferencias válidas sobre una población de interés. La presencia de información faltante en una matriz de datos lleva consigo ciertos inconvenientes, dentro de los cuales, según Horton & Lipsitz (2001), se encuentran: pérdida de eficiencia, complicaciones en el análisis de datos faltantes, y además estimadores sesgados que ponen en riesgo la validez del proceso.

Yates (1933), uno de los precursores en el manejo de datos faltantes, señaló que si éstos fueran remplazados por sus estimadores de mínimos cuadrados aplicados a datos completos, se produciría un estimador de mínimos cuadrados correcto, teoría que no fue muy acogida por la desventaja de producir un estimador de la varianza menor de la real. Bartlett (1937), en cambio, propuso un método de estimación de datos faltantes basado en indicadores faltantes, tomados a través de covariables, el cual tiene la ventaja de obtener estimadores y errores estándares correctos. Posteriormente, varios autores, entre ellos Healy & Wesmacott (1956), proponen la estimación iterativa de información faltante, teoría profundizada más adelante por Dempster et al. (1977), con el algoritmo EM (Esperanza, Maximización). Srivastava & Carter (1986) presentan un método de estimación e imputación a través del análisis de máxima veromisilitud en datos continuos. Little & Rubin (2002) hacen un gran aporte al análisis estadístico con la recopilación de algunas metodologías relacionadas con la estimación y el análisis de información con datos faltantes.

En algunos diseños de experimentos es más común la presencia de información faltante que en otros, como es el caso de los relacionados con medidas repetidas; estos diseños en el campo de la experimentación son frecuentes, especialmente en las ciencias médicas, biológicas y agronómicas; no obstante, presentan ciertas dificultades con su manejo, debido no solo a la ocurrencia de observaciones faltantes, sino a la característica de dependencia entre las observaciones repetidas hechas sobre la misma unidad experimental, pues este tipo de diseños son raramente balanceados y completos. Estas dificultades llevan consigo complicaciones en el modelamiento, pérdida de precisión en estimaciones, y además la inferencia se ve afectada ya que se pueden presentar funciones no estimables.

Entre los aportes más recientes en estudios con medidas repetidas con respuesta binaria, teniendo en cuenta información faltante, se tiene el de Ibrahim (1990), quien propone un método por ponderaciones para modelos lineales generalizados cuando las covariables son faltantes de forma aleatoria y las respuestas son completamente observadas; Park & Davis (1993) tratan mecanismos de datos faltantes en diseños de medidas repetidas con respuesta categórica. De igual forma, Lipsitz

et al. (1999), en datos longitudinales aplican métodos de verosimilitud ponderada para modelos de medidas repetidas incompletos (respuestas y covariables parcialmente observadas). Yang et al. (2005) proponen un método para el manejo de datos faltantes creando un conjunto de estrategias de imputación, analizadas a través de la imputación múltiple.

En este artículo se presenta un método alternativo para el manejo de información faltante en medidas repetidas con respuesta binaria, desde una perspectiva univariada, asumiendo que el mecanismo faltante es de forma aleatoria (FFA). Según Lipsitz et al. (1999), un mecanismo FFA se tiene cuando dados los datos observados, la probabilidad que los datos sean faltantes es condicionalmente independiente de los datos no observados, lo cual permite la estimación de información basada únicamente en los datos observados.

El problema de los datos faltantes merece especial atención en el contexto de modelos que hacen uso de las ecuaciones de estimación generalizada (EEG), debido a la inconveniencia de su aplicación en información bajo el mecanismo FFA. Ya que los datos bajo un mecanismo de faltante de forma completamente aleatoria (FFCA) no tienen problema, se han desarrollado pruebas que aseguran este mecanismo de datos faltantes (Diggle et al. 1994), dentro de las cuales se encuentran la de Chen & Little (1999) basado en patrones de datos faltantes y la de Park & Lee (1997), quienes prueban la significancia del indicador faltante. Según Zorn (2001), si los datos no son FFCA hay varias opciones para los investigadores: una de ellas usa correcciones basadas en imputación, donde los datos faltantes son imputados a partir de los datos disponibles y el análisis es conducido sobre datos completados (empleando métodos de datos completos).

En este artículo se asume que los datos son FFA, suposición que es empleada en el momento de la imputación, ya que está basada en la información observada (modelos condicionales). Para el análisis de la información con datos completados, como lo describen algunos autores mencionados anteriormente, no se tiene en cuenta el mecanismo de información faltante.

En la metodología propuesta en este artículo, en cada una de las ocasiones, iniciando por la primera, se emplea el algoritmo EM en dos formas: en la primera se encuentra un estimador máximo verosímil, en el cual el paso E del algoritmo se expresa como una log-verosimilitud de datos completos ponderada, similar al propuesto por Ibrahim (1990) para covariables faltantes. En el segundo ciclo, el estimador obtenido en la fase anterior se emplea como estimador inicial para determinar la estimación de datos faltantes en el paso E, teniendo en cuenta el modelo con una covariable adicional relacionada con el indicador faltante, como el propuesto por Bartlett (1937); se realiza la imputación de datos faltantes estimados y se efectúa la maximización (paso M) con datos imputados.

La revisión de la notación y los supuestos para Y matriz de respuestas binarias, y X matriz de covariables categóricas se contemplan en la segunda sección; la especificación paso a paso de la metodología propuesta de estimación e imputación de datos faltantes está dada en la tercera sección. En la cuarta, se muestra una aplicación de la metodología propuesta en un caso de estudio relacionado con factores de riesgo coronario, presentado en Fitzmaurice et al. (1994).

2. Notación y supuestos

Sea $Y = (Y_1, Y_2, \ldots, Y_T)$ la matriz de respuestas, observadas parcialmente en tiempos igualmente espaciados $t = 1, 2, \ldots, T$, en donde $Y_t = (y_{1t}, y_{2t}, \ldots, y_{Nt})'$ hace referencia al vector de observaciones en el tiempo t de los N individuos.

Se considera además que el vector de N observaciones en el tiempo t, Y_t, se puede escribir como una partición de n_t datos completamente observados y $N - n_t$ datos faltantes, así

$$Y_t = (y_{1t}, y_{2t}, \ldots, y_{n_t t}, y_{(n_t+1)t}, \ldots, y_{Nt})' = (Y_{obs,t}, Y_{fal,t})'$$

donde $y_{i(obs),t}$, componente de $Y_{obs,t}$, hace referencia al valor del i-ésimo individuo en el tiempo t, cuando $i = 1, 2, \ldots, n_t$, y $y_{i(fal),t}$, componente de $Y_{fal,t}$, hace referencia al valor del i-ésimo individuo en el tiempo t, cuando $i = n_t + 1, \ldots, N$.

Adicionalmente, se tiene una matriz $X_t = (x_{1t}, x_{2t}, \ldots, x_{jt}, \ldots, x_{pt})$ de p covariables completamente observadas, fijas en el tiempo asociadas a Y_t. Particularmente, cada elemento x_{ijt} de la matriz corresponde al valor de la j-ésima covariable ($j = 1, 2, \ldots, p$), del i-ésimo individuo ($i = 1, 2, \ldots, N$) en el t-ésimo tiempo ($t = 1, 2, \ldots, T$) y x_{it} corresponde al vector de covariables del i-ésimo individuo en el tiempo t. Esta matriz se puede escribir también de forma particionada

$$X_t = \begin{bmatrix} X_{obs,t} \\ X_{fal,t} \end{bmatrix}$$

donde $X_{obs,t}$ de tamaño $n_t \times p$ hace referencia a los valores de las p covariables correspondientes a los n_t individuos de las $Y_{obs,t}$ completamente observados en el tiempo t y $X_{fal,t}$ de tamaño $(N - n_t) \times p$ hace referencia a los valores de las p covariables correspondientes a los $N - n_t$ individuos de las $Y_{fal,t}$ en el tiempo t.

Además, se define la matriz C_{t-1} de covariables adicionales relacionada con los tiempos previamente imputados, con $c_{i(t-1)}$ correspondiente al vector asociado al i-ésimo individuo. Esta matriz se puede escribir en forma particionada como

$$C_{t-1} = \begin{bmatrix} C_{obs,t-1} \\ C_{fal,t-1} \end{bmatrix}$$

donde $C_{obs,t-1}$ hace referencia a los valores de las covariables que relacionan los tiempos anteriores, correspondientes a los n_t individuos de las $Y_{obs,t}$ en el tiempo t y $C_{fal,t-1}$ hace referencia a los valores correspondientes a los $N - n_t$ individuos de las $Y_{fal,t}$ en el tiempo t.

3. Estimación de datos faltantes con respuesta binaria

Se considera el vector de observaciones en el tiempo t de los N individuos, $Y_t = (y_{1t}, y_{2t}, \ldots, y_{Nt})'$ donde la variable aleatoria binaria $y_{it} = 1$ si el i-ésimo

sujeto en el tiempo t tiene respuesta 1 y $y_{it} = 0$ si el i-ésimo sujeto en el tiempo t tiene respuesta 0.

La distribución marginal de y_{it} dado el vector de covariables x_{it} con vector de parámetros β es Bernoulli y se modela en términos del log-odds, tomando la forma (Ayala 2006)

$$\begin{aligned} f(y_{it}|x_{it},\beta) =& [p(y_{it}=1|x_{it})]^{y_{it}}[1-p(y_{it}=1|x_{it})]^{1-y_{it}} \\ =& \frac{\exp(y_{it}x'_{it}\beta)}{1+\exp(x'_{it}\beta)}, \qquad y_{it}=0,1 \end{aligned} \tag{1}$$

La expresión (1) se tiene en cuenta tanto para la estimación de parámetros como en la estimación de los datos faltantes.

3.1. Estimación del vector de parámetros vía algoritmo EM

Partiendo del conjunto de observaciones para el tiempo t, teniendo en cuenta tanto observados como faltantes $Y_t = (y_{1t}, y_{2t}, \ldots, y_{n_t,t}, y_{(n_t+1),t}, \ldots, y_{Nt})$, con el empleo del algoritmo EM se realiza la estimación del vector de parámetros, $\theta_{(0)t} = (\beta_{(0)t}, \delta_{(0)t})$, con $\beta_{(0)t}$ y $\delta_{(0)t}$ correspondientes a los parámetros relacionados con X_t y C_t, respectivamente. El paso E del algoritmo se expresa como una log-verosimilitud de datos completos ponderada, como lo muestra Ibrahim (1990).

Para la t-ésima ocasión, con X_t y $Y_{t-1}, Y_{t-2}, \ldots, Y_2, Y_1$ (completadas y reordenadas convenientemente) como covariables relacionadas, el modelo elemento a elemento es

$$\log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = x'_{it}\beta_{(0)t} + c'_{i(t-1)}\delta_{(0)t} \tag{2}$$

o equivalentemente:

$$\log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \begin{bmatrix} x'_{it} & c'_{i(t-1)} \end{bmatrix} \begin{bmatrix} \beta_{(0)t} \\ \delta_{(0)t} \end{bmatrix} = \begin{bmatrix} x'_{it} & c'_{i(t-1)} \end{bmatrix} \theta_{(0)t}$$

con $\pi_{it} = E(y_{it})$.

Siguiendo el proceso para el algoritmo EM introducido por Dempster et al. (1977), a continuación se especifica el paso E de esperanza y el paso M de maximización, los cuales conducen a la estimación del vector de parámetros para el t-ésimo tiempo.

Paso E: Para la iteración $(m+1)$ del algoritmo en el t-ésimo tiempo, con $\theta'_{(0)t} = (\beta_{(0)t}, \delta_{(0)t})$ la log-verosimilitud esperada dados los datos observados se escribe como:

$$\begin{aligned} Q_t\left(\theta_{(0)t}\middle|\ \theta^{(m)}_{(0)t}\right) &= \sum_{i=1}^{N} E\left[l(\theta_{(0)t}; y_{it})\middle|\ x_{it}, c_{it}, \theta_{(0)t} = \theta^{(m)}_{(0)t}\right] \\ &= \sum_{i=1}^{N} \sum_{y_{i(fal)}(k)} l\left(\theta_{(0)t}; y_{it}) p(y_{i(fal),t}(k)\middle|\ x_{it}, c_{it}, \theta^{(m)}_{(0)t}\right) \end{aligned} \tag{3}$$

Esta suma se extiende sobre todos los posibles valores de las componentes faltantes de los vectores respuesta, con $k = 0, 1$ indicando los dos posibles patrones de respuesta que el sujeto i podría tener dadas las covariables. Por ejemplo, si la observación para el tercer individuo y_{3t} es faltante, la cual está relacionada con el vector de covariables $x_{3t} = (1, 0, 0)$, los dos patrones posibles para (y_{3t}, x_{3t}) están dados por $(0, 1, 0, 0)$ y $(1, 1, 0, 0)$, manteniendo fijos los valores de las covariables para cada patrón.

Para la estimación del vector de parámetros inicial del algoritmo, bajo el modelo de covarianza (2), empleando mínimos cuadrados se tienen las siguientes ecuaciones normales:

$$X'_{obs,t}X_{obs,t}\widehat{\beta}^{(0)}_{(0)t} + X'C_{t-1,obs}\widehat{\delta}^{(0)}_{(0)t} = X'_{obs,t}Y_{obs,t}$$
$$C_{obs,t-1}X\widehat{\beta}^{(0)}_{(0)t} + C'_{obs,t-1}C_{obs,t-1}\widehat{\delta}^{(0)}_{(0)t} = C'_{obs,t-1}Y_{obs,t}$$

Con las anteriores ecuaciones se obtiene la estimación del parámetro inicial $\theta^{(0)}_{(0)t}$ para el tiempo t en el modelo de covarianza con datos observados:

$$\widehat{\beta}^{(0)}_{(0)t} = \left(X'_{obs,t}X_{obs,t}\right)^{-1}X'_{obs,t}Y_{obs,t} - \left(X'_{obs,t}X_{obs,t}\right)^{-1}X'C_{obs,t-1}\widehat{\delta}^{(0)}_{(0)t} \quad (4)$$
$$\widehat{\delta}^{(0)}_{(0)t} = \left[C'_{obs,t-1}(I - P_x)C_{obs,t-1}\right]^{-1}C'_{obs,t-1}(I - P_x)Y_{obs,t} \quad (5)$$

donde $P_x = X_{obs,t}\left(X'_{obs,t}X_{obs,t}\right)^{-1}X'_{obs,t}$.

Dado este estimador inicial, se especifica la función Q teniendo en cuenta los dos posibles patrones, con ponderación para el i-ésimo individuo en el t-ésimo tiempo, determinada por:

$$w^{(m)}_{itk} = p\Big(y_{i(fal),t}(k)\Big|\ c_{i(t-1)}, x_{it}; \theta^{(m)}_{(0)t}\Big) \quad (6)$$

De esta forma, la expresión (3) está dada por

$$Q_t\Big(\theta_{(0)t}\Big|\ \theta^{(m)}_{(0)t}\Big) = \sum_{i=1}^{N}\sum_{y_{i(fal),t}(k)} l\big(\theta_{(0)t}; y_{it}\big)w^{(m)}_{itk} \quad (7)$$

expresión correspondiente a una log-verosimilitud de datos completos ponderada, basada en un nuevo conjunto de datos que tiene en cuenta, para cada faltante, las dos posibles respuestas en este caso binario. De esta forma, el nuevo número de observaciones está dado por $N_t = n_t + 2(N - n_t) = 2N - n_t$, donde la ponderación $w^{(m)}_{itk}$ para la i-ésima observación, en el t-ésimo tiempo, del k-ésimo patrón de respuesta, se especifica de la siguiente forma (Ayala 2006):

$$w^{(m)}_{itk} = \begin{cases} 1, & \text{si } 1 \le i \le n_t; \\ \widetilde{\pi}^{(m)}_{it}, & \text{si } n_t < i < N_t \text{ para } k = 0; \\ 1 - \widetilde{\pi}^{(m)}_{it}, & \text{si } n_t < i < N_t \text{ para } k = 1. \end{cases} \quad (8)$$

con $\widetilde{\pi}_{it}^{(m)} = \dfrac{\exp\left(x'_{i(fal)t}\theta_{(0)t}^{(m)}\right)}{1+\exp\left(x'_{i(fal)t}\theta_{(0)t}^{(m)}\right)}$.

Paso M: Maximiza la función (7), lo cual es equivalente a aplicar máxima verosimilitud a un conjunto de datos completos, con cada observación incompleta remplazada por un conjunto de observaciones ponderadas (Ibrahim 1990).

Con $\nabla Q_t\left(\theta_{(0)t}|\ \theta_{(0)t}^{(m)}\right)$ correspondiente al vector gradiente de $Q_t\left(\theta_{(0)t}|\ \theta_{(0)t}^{(m)}\right)$ con respecto a $\theta_{(0)t}$, el paso M encuentra el valor de $\theta_{(0)t}$ que satisface $\nabla Q_t\left(\theta_{(0)t}|\ \theta_{(0)t}^{(m)}\right) = 0$, el cual se denota por $\theta_{(0)t}^{(m+1)}$:

$$\nabla Q_t\left(\theta_{(0)t}|\ \theta_{(0)t}^{(m)}\right) = \sum_{i=1}^{N}\sum_{y_{i(fal)}(k)} \frac{\partial l\left(\beta_{(0)tj},\delta_{(0)t};y_{it}\right)}{\partial\left(\beta_{(0)tj},\delta_{(0)t}\right)} w_{itk}^{(m)} \tag{9}$$

con $j = 1,\ldots,p$.

Partiendo de (1) la ecuación de log-verosimilitud, teniendo en cuenta datos binarios, con $x'_{it}\beta = \eta_i = \log\left(\frac{\pi_{it}}{1-\pi_{it}}\right)$, es:

$$l(\pi_i; y_{it},\ldots,y_{Nt}) = \sum_{i=1}^{N}\left[y_{it}\log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) + \log(1-\pi_{it})\right]$$

Derivando la función de log-verosimilitud con respecto a cada uno de los parámetros $\theta'_{(0)t} = \left(\beta_{(0)t},\delta_{(0)t}\right)$, por regla de la cadena y teniendo en cuenta que l está en función de π_{it}, se obtiene (Ayala 2006):

$$\begin{aligned}\frac{\partial l_i}{\partial\beta_{(0)tj}} &= \sum_{i=1}^{N}\frac{(y_{it}-\pi_{it})\partial\pi_{it}}{\pi_{it}(1-\pi_{it})\partial\eta_i}x_{ijt}\\ \frac{\partial l_i}{\partial\delta_{(0)t}} &= \sum_{i=1}^{N}\frac{(y_{it}-\pi_{it})\partial\pi_{it}}{\pi_{it}(1-\pi_{it})\partial\eta_i}c_{i(t-1)}\end{aligned}$$

lo que conlleva a ecuaciones no lineales que deben ser resueltas por métodos iterativos. Siguiendo el proceso de estimación en modelos lineales generalizados de McCullagh & Nelder (1989) y teniendo presente el método por ponderaciones de Ibrahim (1990), en la m-ésima iteración del algoritmo EM y la s-ésima iteración del algoritmo de Scoring, la ecuación iterativa de estimación para $\theta_{(0)t}$ toma la forma (Ayala 2006):

$$\theta_{(0)t}^{(s)} = \left[\left(\widetilde{X}_t\widetilde{C}_{t-1}\right)'W^{(m)}M_t\left(\widetilde{X}_t\widetilde{C}_{t-1}\right)\right]^{-1}\left(\widetilde{X}_t\widetilde{C}_{t-1}\right)W^{(m)}M_t z \tag{10}$$

donde $W^{(m)} = diag\left(w_{itk}^{(m)}\right)$ es una matriz de tamaño $N_t \times N_t$, $M_t = diag(\pi_{it}(1-\pi_{it}))$ y $z = \widetilde{X}_t\beta_{(0)}^{(s-1)} + \nu^{(s-1)}$, con $\nu = \left((y_{1t}-\pi_{1t})\frac{\partial\eta_1}{\partial\pi_{1t}},\ldots,(y_{N_1t}-\pi_{1t})\frac{\partial\eta_{N_t}}{\partial\pi_{N_1t}}\right)'$.

$\widetilde{X}_t$ y $\widetilde{C}_{t-1}$ son matrices aumentadas de tamaño $N_t \times p$ y $N_t \times (t-1)$ respectivamente, conformadas cada una por dos submatrices:

$$\widetilde{X}_t = \begin{bmatrix} X_t \\ X_{fal,t} \end{bmatrix}, \qquad \widetilde{C}_{t-1} = \begin{bmatrix} C_{t-1} \\ C^*_{t-1} \end{bmatrix}$$

con X_t y C_{t-1} correspondientes a Y_t, anteriormente definidas, $X_{fal,t}$ como se especificó en la sección 2 y C^*_{t-1} matriz conformada por las últimas $N - n_t$ filas de C_{t-1}.

En cada una de las iteraciones se tiene a $\widetilde{X}_t, \widetilde{C}_{t-1}, M_t$ y $W^{(m)}$ matrices fijas, mientras que las demás cambian con cada iteración. De esta forma, se obtiene el estimador máximo verosímil de $\theta_{(0)t}$ en la s-ésima iteración del paso M, de la m-ésima iteración del algoritmo EM.

En las anteriores condiciones se itera entre paso E y M hasta lograr un nivel de convergencia deseado en las ponderaciones y en las estimaciones de los parámetros.

3.2. Estimación de datos faltantes vía algoritmo EM

Se hace uso del algoritmo EM partiendo nuevamente del conjunto de observaciones $Y_t = (y_{it})_i = (y_{1t}, y_{2t}, \ldots, y_{(n_t+1),t}, \ldots, y_{Nt})$, con el fin de estimar e imputar los datos faltantes. En el paso E del algoritmo se realiza la estimación e imputación de los datos faltantes basada en el método ANCOVA de Bartlett (1937), tomando como estimador inicial $\theta_{(0)t}$ descrito en la sección 3.1. Luego en el paso M se hace el proceso de maximización con los datos completados por medio del algoritmo de estimación de Scoring.

Paso E: El paso de esperanza imputa los valores de los datos faltantes bajo el supuesto FFA, es decir éstos son remplazados por sus esperanzas condicionales dados los observados y unos parámetros iniciales (Fitzmaurice et al. 1994). Para dicho fin, se adecúa el modelo propuesto por Bartlett (1937), teniendo en cuenta el modelo de regresión logística de la forma:

$$Logit(\pi_t) = X_t\beta_t + C_{t-1}\delta_t + Z_t\gamma_t \tag{11}$$

donde $Logit(\pi_t)$ es un vector de tamaño N cuyo elemento i-ésimo es $logit(\pi_{it}) = \log\left(\frac{\pi_{it}}{1-\pi_{it}}\right)$, Z_t es una matriz de tamaño $N \times (N - n_t)$ correspondiente a $N - n_t$ covariables de valor faltante en el t-ésimo tiempo y γ_t es el vector columna de los $N - n_t$ coeficientes de regresión para las covariables de valor faltante. Particionando las matrices entre observados y faltantes en (11), se obtiene

$$\begin{bmatrix} Logit(\pi_{obs,t}) \\ Logit(\pi_{fal,t}) \end{bmatrix} = \begin{bmatrix} X_{obs,t} \\ X_{fal,t} \end{bmatrix} \beta_t + \begin{bmatrix} C_{obs,t-1} \\ C_{fal,t-1} \end{bmatrix} \delta_t + \begin{bmatrix} 0_{obs,t} \\ -I_{fal,t} \end{bmatrix} \gamma_t$$

donde $Logit(\pi_{obs,t})$ hace referencia al vector logit correspondiente a los n_t individuos de las $Y_{obs,t}$ en el tiempo t, y $Logit(\pi_{fal,t})$ hace referencia al vector logit correspondiente a los $N - n_t$ individuos de las $Y_{fal,t}$ en el tiempo t; $0_{obs,t}$ es una matriz de ceros de tamaño $n_t \times (N - n_t)$ relacionada con la información observada

y $-I_{fal,1}$ es una identidad negativa de tamaño $(N - n_t) \times (N - n_t)$ relacionada con los datos faltantes.

Con base en lo anterior, para el t-ésimo tiempo la *log*-verosimilitud esperada dados los datos observados se escribe como

$$Q_{it}\big(\theta_t \big|\ \theta_t^{(m)}\big) = E\Big[l(\theta_i; y_{it})\big|\ x_{it}, c_{i(t-1)}, y_{i(obs)t}, \theta_t = \theta_t^{(m)}\Big] \tag{12}$$

donde $\theta_t = (\beta_t, \delta_t, \gamma_t)$.

De acuerdo con la descripción original del método de Bartlett y teniendo en cuenta las características de un modelo lineal generalizado, la suma de los cuadrados de los errores a ser minimizada sobre γ_t dados $\widehat{\beta}_t$ y $\widehat{\delta}_t$ es

$$SCE\big(\gamma_t/\widehat{\beta}_t, \widehat{\delta}_t\big) = \sum_{i=1}^{n_t} \big(logit(\pi_{it}) - x_{it}\widehat{\beta}_t - c_{i(t-1)}\widehat{\delta}_t\big)^2 + \sum_{i=n_t+1}^{N} \big(logit(\pi_{it}) - x_{it}\widehat{\beta}_t - c_{i(t-1)}\widehat{\delta}_t + \gamma_t\big)^2$$

Al suponer que el resultado sobre la respuesta inicial en los datos faltantes es equiprobable, es decir $\pi_{it} = 0.5$, se encuentra que $logit(\pi_{it}) = 0$, entonces el estimador de mínimos cuadrados de γ_t para la m-ésima iteración se escribe como

$$\log\left(\frac{\widehat{\pi}_{i(fal),t}^{(m)}}{1 - \widehat{\pi}_{i(fal),t}^{(m)}}\right) = \widehat{\gamma}_t^{(m)} = x'_{i(fal),t}\widehat{\beta}_t^{(m)} + \big(c_{i(fal)(t-1)} - \overline{c}_{t-1}\big)'\widehat{\delta}_t^{(m)} \tag{13}$$

que por las características propias de las respuestas faltantes, es transformado a partir del log-odds, obteniendo

$$\widehat{\pi}_{i(fal),t}^{(m)} = \frac{\exp\left[x'_{i(fal),t}\widehat{\beta}_t^{(m)} + \big(c_{i(fal)(t-1)} - \overline{c}_{t-1}\big)'\widehat{\delta}_t^{(m)}\right]}{1 + \exp\left[x'_{i(fal),t}\widehat{\beta}_t^{(m)} + \big(c_{i(fal)(t-1)} - \overline{c}_{t-1}\big)'\widehat{\delta}_t^{(m)}\right]} \tag{14}$$

donde $\widehat{\pi}_{i(fal),t}^{(m)}$ hace referencia al valor de la media correspondiente al individuo i en el tiempo t con $n_t + 1 \leq i \leq N$, dada en la m-ésima iteración.

Teniendo en cuenta el siguiente criterio

$$\widehat{y}_{i(fal),t} = \begin{cases} 0, & \text{si } \widehat{\pi}_{i(fal),t} < p_0; \\ 1, & \text{si } \widehat{\pi}_{i(fal),t} \geq p_0. \end{cases} \tag{15}$$

se obtiene la imputación de los datos faltantes dados los observados y el estimador $\beta_t^{(m)}$, donde p_0 hace referencia a un valor particular que va de acuerdo con las condiciones del experimento o con el juicio del experto. Sería necesario la realización de simulaciones para identificar cómo el valor de p_0 afecta el proceso de estimación de los parámetros, pero este proceso está fuera del contexto de este artículo.

Paso M: imputadas las observaciones, se procede a la maximización partiendo del conjunto de datos completos, utilizando para ello algún método de maximización: Scoring o Newton Raphson.

Se itera entre paso E y M hasta lograr convergencia en las estimaciones de los datos faltantes y parámetros.

3.3. Procedimiento

Para ser aún más explícito el procedimiento en la estimación e imputación de información faltante, se establecieron los siguientes pasos que muestran aspectos específicos para cada una de las ocasiones con base en las dos secciones anteriores. Se trata de la descripción para las tres primeras ocasiones (paso 1 al paso 3) y una descripción generalizada para la t-ésima ocasión (paso 4).

1. *Estimación e imputación de información faltante en la primera ocasión* ($t = 1$). Inicialmente se encuentra la estimación del vector de parámetros basada en la información observada, siguiendo el proceso desarrollado en la sección 3.1, para de esta forma obtener el estimador inicial $\theta_{(0)1} = \beta_{(0)1}$. Con (2), el modelo propuesto para $t = 1$ es

$$\log\left(\frac{\pi_{i1}}{1-\pi_{i1}}\right) = x'_{i1}\beta_{(0)1}$$

por lo cual siguiendo a (3) y (6), la log-verosimilitud esperada dados los datos observados se escribe como

$$\begin{aligned} Q_{i1}\left(\beta_{(0)1}\middle|\ \beta^{(m)}_{(0)1}\right) &= \sum_{y_{i(fal)}(k)} l(\beta_{(0)1}; y_{i1}) p\left(y_{i(fal),1}(k)\middle|\ x_{i1}, \beta^{(m)}_{(0)1}\right) \\ &= \sum_{y_{i(fal)}(k)} l(\beta_{(0)1}; y_{i1}) w^{(m)}_{ik} \end{aligned}$$

En general, se puede escribir el paso E del algoritmo EM para todas las observaciones en el primer tiempo como:

$$Q_1\left(\beta_{(0)1}\middle|\ \beta^{(m)}_{(0)1}\right) = \sum_{i=1}^{N} \sum_{y_{i(fal)}(k)} l(\beta_{(0)1}; y_{i1}) w^{(m)}_{ik} \tag{16}$$

El estimador inicial sugerido para este caso $\beta^{(0)}_{(0)1}$ es el estimador de mínimos cuadrados para un modelo lineal general, o cualquier otra estimación que tenga sentido.

Para el paso M, la ecuación iterativa de $\beta_{(0)1}$ en la m-ésima iteración del algoritmo EM y la s-ésima iteración del algoritmo de Scoring toma la forma

$$\beta^{(s)}_{(0)} = \left(\widetilde{X}_1 W^{(m)} M_1 \widetilde{X}_1\right)^{-1} \widetilde{X}'_1 W^{(m)} M_1 z \tag{17}$$

En (17) las matrices relacionadas son como se definen en la sección 3.1. Con estas especificaciones para el primer tiempo, se realiza el proceso mostrado en la sección 3.1 para $t = 1$.

Siguiendo el proceso de la sección 3.2, relacionado con la estimación e imputación de datos faltantes en la primera ocasión, para el paso E del algoritmo el modelo propuesto es

$$Logit(\pi_1) = X_1\beta_1 + Z_1\gamma_1 \tag{18}$$

y la log-verosimilitud esperada dados los datos observados se escribe como

$$Q_{i1}\left(\theta_1 \middle|\ \theta_1^{(m)}\right) = E\left[l(\theta_1; y_{i1}) \mid x_{i1}, y_{i(obs),1}, \theta_1 = \theta_1^{(m)}\right]$$

donde $\theta_1 = (\beta_1, \gamma_1)$.

Análogo a (14), la estimación de los datos faltantes está dada por

$$\widehat{\pi}_{i(fal),1}^{(m)} = \frac{\exp\left(x'_{i(fal),1}\beta_1^{(m)}\right)}{1 + \exp\left(x'_{i(fal),1}\beta_1^{(m)}\right)} \tag{19}$$

Según el mismo criterio opcional dado en (15), se obtiene la imputación de los datos faltantes para $t = 1$.

2. *Estimación e imputación de datos faltantes en la segunda ocasión* ($t = 2$). En particular, se considera el siguiente modelo de regresión logística

$$\log\left(\frac{\pi_{i2}}{1 - \pi_{i2}}\right) = x'_{i2}\beta_2 + c_{i1}\delta_2 \tag{20}$$

donde $\pi_{i2} = E(y_{i2})$, $c_{i1} = y_{i1}$ completado y δ_2 es el parámetro correspondiente a la covariable relacionada con el primer tiempo. Lo demás es como se establece en la sección 3.1.

Se continúa con el proceso dado en la sección 3.2 para la estimación e imputación de datos faltantes en la segunda ocasión.

3. *Estimación e imputación de datos faltantes en la tercera ocasión* ($t = 3$). Como se describe en la sección 3.1, dados los datos observados e imputados en el primer y segundo tiempo, se estiman los parámetros bajo el modelo (2) de covarianza para $t = 3$, donde Y_3 es la variable respuesta, X_3 las variables explicativas, y Y_1 y Y_2 las covariables. Por la presencia de multicolinealidad debida a la correlación entre Y_1 y Y_2, es necesario hacer una descomposición de las variables de tal forma que se ortogonalicen. Debido a la característica categórica de la respuesta se emplea un análisis de correspondencias entre las dos variables Y_1 y Y_2, el cual permite obtener una nueva covariable, C_2, que retiene la máxima variabilidad contenida en Y_1 y Y_2.

   En el análisis de correspondencias simples se especifica una matriz F de densidades o frecuencias relativas (f_{ij}) de 2×2 cuyas filas corresponden a las dos categorías de Y_1 y las columnas corresponden a las dos categorías de Y_2. De acuerdo con este criterio, la tabla de contingencia está dada por

		Y_2		
		1	0	Total
Y_1	1	f_{11}	f_{10}	$f_{1\bullet}$
	0	f_{01}	f_{00}	$f_{0\bullet}$
Total		$f_{\bullet 1}$	$f_{\bullet 0}$	$f_{\bullet\bullet}$

donde $f_{\bullet j} = \sum_{i=0}^{1} f_{ij}$, $f_{i\bullet} = \sum_{j=0}^{1} f_{ij}$ y $f_{\bullet\bullet} = \sum_{i=0}^{1}\sum_{j=0}^{1} f_{ij}$, con matriz de densidades

$$F = \frac{1}{f_{\bullet\bullet}} \begin{bmatrix} f_{11} & f_{10} \\ f_{01} & f_{00} \end{bmatrix}$$

Siguiendo a Peña (2002), se especifica la matriz R de frecuencias relativas condicionadas al total de la fila dada por

$$R = D_f^{-1} F$$

donde D_f es una matriz diagonal de 2×2 con los términos $f_{i\bullet}$.

Adicionalmente, se especifica la matriz G de frecuencias relativas condicionadas al total de la fila estandarizadas por su variabilidad

$$G = RD_c^{-1/2}$$

con D_c matriz diagonal de 2×2 con los términos $f_{\bullet j}$, cuyo elemento ij-ésimo está dado por $g_{ij} = \left(\frac{f_{ij}}{f_{i\bullet} f_{\bullet j}^{1/2}}\right)$ con $i = 0, 1$ y $j = 0, 1$.

Ahora, se obtiene un vector a de norma la unidad, tal que el vector de puntos proyectados sobre esta dirección Ga tenga variabilidad máxima (Peña 2002). Al proyectar los puntos sobre las direcciones de máxima variabilidad, de forma similar que en componentes principales, el vector a es un vector propio de la matriz $G'G$ ponderada, es decir de $G'D_fG$. Con lo anterior, la nueva covariable que retiene la máxima variabilidad es

$$C_2 = [Y_1 \;\; Y_2]a$$

donde $a' = (a_1 \;\; a_2)$. Con C_2 se realiza el proceso de estimación (sección 3.1) partiendo del modelo

$$\log\left(\frac{\pi_{i3}}{1-\pi_{i3}}\right) = x'_{i3}\beta_3 + c_{i2}\delta_3 \tag{21}$$

se continua el proceso de la sección 3.2 para la estimación e imputación de datos faltantes en la tercera ocasión.

4. *Estimación del vector de parámetros en la t-ésima ocasión.* Dados los datos completados en $Y_1, Y_2, \ldots, Y_{t-1}$, se estiman los parámetros en Y_t, según un modelo de covarianza, en donde Y_t corresponde al vector de respuestas, X_t y

C_{t-1} son las covariables. Como en el paso 3 para Y_3, en Y_t también hay presencia de multicolinealidad por la correlación dada por $Y_1, Y_2, \ldots, Y_{t-1}$, y por ello, en este caso de varias variables se aplica el análisis de correspondencias múltiples.

La matriz de covariables está dada por

$$C_{t-1} = (Y_1\ Y_2\ \cdots\ Y_{t-1})A; \qquad t = 2, 3, \ldots, T$$

donde $A = (a_1\ a_2\ \cdots\ a_{t-2})$ es la matriz de vectores propios correspondientes a valores propios diferentes de 1, obtenidos de

$$S = \frac{1}{k}B'BD$$

donde B es la matriz conformada por los elementos de la tabla disyuntiva completa que comprende N filas y $t-1$ columnas, las cuales describen las dos posibles respuestas de los N individuos a través de un código binario (0 o 1) y, D es una matriz diagonal, cuyos elementos de la diagonal están asociados a los de $B'B$. Con esta matriz de covariables para el modelo dado en (2) se realiza la estimación de parámetros descrita para el t-ésimo tiempo en la sección 3.1 y luego se hace la estimación e imputación de datos faltantes en la t-ésima ocasión, siguiendo el proceso de la sección 3.2.

4. Aplicación

En esta sección se presenta un ejemplo para ilustrar el método propuesto en las anteriores secciones. En 1970, investigadores de la Universidad de Iowa iniciaron un estudio sobre la relación entre factores de riesgo coronario en jóvenes y enfermedades coronarias en adultos. Para tal fin se comenzó con un estudio sobre un grupo de niños con el objeto de examinar el desarrollo y la persistencia de los factores de riesgo de enfermedades coronarias en jóvenes (Fitzmaurice et al. 2004).

El estudio contiene registros de 1014 niños, 493 hombres y 521 mujeres a quienes se les midió la altura y el peso en tres ocasiones: 1977, 1979 y 1981. Se calculó el peso relativo (índice de masa corporal) como medida de obesidad, teniendo en cuenta la razón del peso observado de cada niño y el peso mediano con respecto a edad, género y altura. Los niños con un peso relativo mayor que el 110 % del peso mediano fueron clasificados como obesos (Wolson & Clarke 1984), obteniendo de esta forma respuestas binarias que describen si el niño es obeso o no (1 si es obeso y 0 si no lo es) en cada ocasión. La tabla de observaciones para todos los niños que participaron en este estudio está dada en Fitzmaurice et al. (2004).

Los datos incompletos corresponden únicamente a la variable peso relativo de los niños, quienes no participaron en todos los años de observación. Como lo menciona Fitzmaurice et al. (2004), más del 50 % de los niños tuvieron por lo menos una respuesta faltante. Además, comparando la cantidad de información completa, es decir, la relacionada con los niños que se midieron en las tres ocasiones, con la

cantidad de información total correspondiente a completos y faltantes, se tiene que de 1014 niños tan solo 460 fueron medidos en las tres ocasiones.

En las condiciones anteriores, $Y = (Y_1, Y_2, Y_3)$ es la matriz de respuestas medidas en tres ocasiones (1977, 1979 y 1981), la cual es parcialmente faltante. Cada uno de los elementos de dicha matriz está dado por

$$y_{it} = \begin{cases} 1, & \text{si el } i\text{-ésimo niño en el } t\text{-ésimo tiempo se clasifica como obeso;} \\ 0, & \text{si se clasifica como no obeso.} \end{cases}$$

para $i = 1, \ldots, 1014$ y $t = 1, 2, 3$.

La metodología propuesta se implementó en dos paquetes: MatLab y SAS (programas que se pueden descargar de la página web de la Revista Colombiana de Estadística). En el primero se desarrolla el programa encargado de hacer la imputación de la información, mientras que en el segundo se desarrolla un programa para la estimación de parámetros basada en información completada.

4.1. Estimación de datos faltantes

Siguiendo los pasos dados en la sección 3.3 se tiene:

1. Se considera el modelo de regresión logística para el primer tiempo

$$\log\left(\frac{\pi_{i1}}{1-\pi_{i1}}\right) = \beta_{(0)11} + \beta_{(0)21} g_i$$

con el cual se obtiene la matriz aumentada, teniendo en cuenta el primer tiempo como respuesta y el género como única covariable, con $g = 1$ si es mujer y $g = 0$ si es hombre. Como en el primer tiempo hay 306 datos faltantes, la matriz aumentada queda determinada por $N_1 = 1014 + 306 = 1320$ observaciones, con la cual se realiza el proceso de maximización (Paso M) para la estimación de parámetros, obteniéndose $\widehat{\theta}'_{(0)1} = \widehat{\beta}'_{(0)1} = (0.1648, 0.0234)$.

Con este estimador inicial de β, se estiman los datos faltantes de acuerdo con (19) y se realiza la imputación de datos, según el criterio

$$\widehat{y}_{i(fal),1} = \begin{cases} 0, & \text{si } \widehat{\pi}_{i(fal),1} < p_0; \\ 1, & \text{si } \widehat{\pi}_{i(fal),1} \geq p_0. \end{cases}$$

donde $p_0 = \bar{y}_{obs,1} = 0.1765$, el cual corresponde al valor medio de los datos observados.

2. Se considera el modelo de regresión logística para el segundo tiempo

$$\log\left(\frac{\pi_{i2}}{1-\pi_{i2}}\right) = \beta_{(0)12} + \beta_{(0)22} g_i + \delta_{(0)2} c_{i1}$$

con una matriz aumentada determinada por $N_2 = 1014 + 272 = 1286$ observaciones, y teniendo en cuenta el estimador inicial para θ_2 dado por

(4) y (5), se realiza el proceso de maximización (Paso M), obteniéndose $\widehat{\theta}'_{(0)2} = (-1.55, -1.1782, 2.3402)$. Con esta estimación inicial, se estiman los datos faltantes de acuerdo con (14), y se realiza la imputación de datos basada en el criterio (15) con $p_0 = \overline{y}_{obs,2} = 0.22$.

3. Con la matriz aumentada para el tercer tiempo, determinada por $N_3 = 1014 + 264 = 1278$ se realiza el proceso de maximización (Paso M), en donde se tiene en cuenta la introducción de Y_1 y Y_2 completadas como covariables adicionadas en el modelo, para lo cual se aplica el análisis de correspondencias (sección 3.3 paso 3). Después de 7 iteraciones la estimación de $\widehat{\theta}'_{(0)3} = (1.9865, -0.8739, 3.2937)$.

   Desarrollando el proceso iterativo EM, tomando como estimador inicial $\widehat{\theta}'_{(0)3}$, se obtienen las estimaciones para la imputación de los datos faltantes, teniendo de nuevo en cuenta el criterio presentado en la expresión (15) con $p_0 = \overline{y}_{obs,3} = 0.243$.

   El seguimiento de los pasos anteriores trae consigo la estimación e imputación de información faltante, como se muestra en la tabla 1, en donde los números en negrilla indican la información imputada, es necesario aclarar que estos resultados pueden cambiar si las condiciones del modelo son diferentes.

   Agrupando la información de observados e imputados, en 2^3 perfiles de respuesta por género, se obtiene la tabla 2 de datos completados.

   A continuación, se presenta el análisis longitudinal del estudio mencionado, basado en el conjunto de observaciones completadas, en donde se tiene una respuesta binaria que indica si el niño es obeso de acuerdo con ciertos parámetros de peso y edad mencionados anteriormente. El objetivo del análisis es determinar si el riesgo de obesidad se incrementa con la edad y si los patrones de cambio en la obesidad son los mismos para hombres y mujeres. La probabilidad marginal de obesidad se modela como una función logística de las covariables género, edad lineal y cuadrática, de igual forma que lo propuesto en Fitzmaurice et al. (1994), con fines de comparación de resultados.

   El modelo considerado es

$$logit(\pi) = \beta_0 + \beta_1 g + \beta_2 E_L + \beta_3 E_C + \beta_4 g E_L + \beta_5 g E_C \qquad (22)$$

   donde $g = 1$ si es mujer y $g = 0$ si es hombre; E_L y E_C son los factores lineal y cuadrático, respectivamente, relacionados con la edad, y gE_L y gE_C son las interacciones entre los factores anteriormente mencionados. Se asume que el log-odds de obesidad cambia curvilíneamente con la edad (tendencia cuadrática) de forma diferente en niños que en niñas.

   Los coeficientes de regresión estimados con sus correspondientes errores estándares, obtenidos a través del paquete SAS usando la aproximación de ecuaciones de estimación generalizada (EEG) con datos completados (observados e imputados), empleando una matriz de correlación de trabajo no estructurada, se presentan en la tabla 3. Como se observa en los resultados,

Tabla 1: Estimación de datos faltantes en respuestas de obesidad en niños.

	Edad Hombres					Edad Mujeres			
Ítem	8	10	12	Frecuencia	Ítem	8	10	12	Frecuencia
1	1	1	1	20	27	1	1	1	21
2	1	1	0	7	28	1	1	0	6
3	1	0	1	9	29	1	0	1	6
4	1	0	0	8	30	1	0	0	2
5	0	1	1	8	31	0	1	1	19
6	0	1	0	8	32	0	1	0	13
7	0	0	1	15	33	0	0	1	14
8	0	0	0	150	34	0	0	0	154
9	**1**	1	1	13	35	**0**	1	1	8
10	**1**	1	0	3	36	**0**	1	0	1
11	**1**	0	1	2	37	**0**	0	1	4
12	**1**	0	0	42	38	**0**	0	0	47
13	1	**1**	1	3	39	1	**1**	1	4
14	1	**1**	0	1	40	1	**1**	0	0
15	0	**0**	1	6	41	0	**0**	1	3
16	0	**0**	0	16	42	0	**0**	0	16
17	1	1	**1**	11	43	1	1	**1**	11
18	1	0	**0**	1	44	1	0	**1**	1
19	0	1	**1**	3	45	0	1	**1**	3
20	0	0	**0**	38	46	0	0	**0**	25
21	**1**	**1**	1	14	47	**0**	**0**	1	13
22	**1**	**1**	0	55	48	**0**	**0**	0	39
23	**1**	1	**1**	4	49	**0**	1	**1**	5
24	**1**	0	**0**	33	50	**0**	0	**0**	23
25	**1**	**1**	**1**	7	51	1	**1**	**1**	7
26	0	**0**	**0**	45	52	0	**0**	**0**	47

1: Obeso observado; 0: No obeso observado

1: Obeso estimado; **0**: No obeso estimado

Tabla 2: Datos obtenidos después de aplicar la metodología propuesta.

	Edad Hombres					Edad Mujeres			
Ítem	8	10	12	Frecuencia	Ítem	8	10	12	Frecuencia
1	0	0	0	249	9	0	0	0	351
2	0	0	1	21	10	0	0	1	34
3	0	1	0	8	11	0	1	0	14
4	0	1	1	11	12	0	1	1	35
5	1	0	0	84	13	1	0	0	2
6	1	0	1	11	14	1	0	1	7
7	1	1	0	66	15	1	1	0	6
8	1	1	1	72	16	1	1	1	43

1: Obeso; 0: No obeso

los coeficientes de regresión asociados a las variables género, edad lineal, interacción entre género y edad lineal, y la interacción entre género y edad cuadrática son significativos en el modelo, como lo muestra el estadístico Z y los valores p correspondientes, a un nivel de significancia del 5 %.

Tabla 3: Estimaciones de los parámetros para el modelo con datos observados y estimados.

Parámetro	Estimación	Error estándar	Límites del 95 % de confianza	Z	Valor p
Intercepto	-1.5155	0.1016	$(-1.7147; -1.3163)$	-14.91	<0.0001
g	0.7412	0.1273	(0.4917; 0.9907)	5.82	<0.0001
E_L	-0.5242	0.0579	$(-0.6378; -0.4107)$	-9.05	<0.0001
E_C	0.0347	0.0247	$(-0.0137; 0.0831)$	1.40	0.1604
gE_L	0.9593	0.0844	(0.7940; 1.1247)	11.37	<0.0001
gE_C	-0.0967	0.0377	$(-0.1706; -0.0228)$	-2.57	0.0103

4.2. Comparación ilustrativa del método propuesto con otras metodologías

En esta sección se presenta una comparación de tres métodos para el manejo de datos faltantes, empleando la información del estudio de riesgos de enfermedades coronarias en niños.

1. *Método de caso completo.* Se realiza la estimación de parámetros con sus correspondientes errores estándares usando la aproximación EEG, a partir de los datos de los individuos que fueron completamente observados. En este caso en particular, de 1014 individuos de estudio, se realiza el proceso de estimación con 460, resultado de la eliminación de los individuos con por lo menos una observación faltante.

   Los resultados obtenidos, partiendo del modelo (22) con 460 observaciones, a través del programa en SAS, se muestran en la tabla 4.

Tabla 4: Estimadores de parámetros usando análisis de casos completos.

Parámetro	Estimación	Error estándar	Límites del 95 % de confianza	Z	Valor p
Intercepto	-1.353	0.113	$(-1.5745; -1.1315)$	-9.794	< 0.0001
g	-0.051	0.190	$(-0.4234; 0.3214)$	-0.268	0.7887
E_L	0.106	0.077	$(-0.0449; 0.2569)$	1.308	0.1661
E_C	0.045	0.047	$(-0.0471; 0.1371)$	0.956	0.3391
gE_L	0.230	0.119	$(-0.0032; 0.4632)$	1.937	0.0527
gE_C	-0.149	0.065	$(-0.2764; -0.0216)$	-2.314	0.0207

2. *Metodología propuesta por Fitzmaurice et al. (1994).* Esta metodología para el manejo de información faltante está fundamentada en modelos marginales. La esperanza marginal de la respuesta μ_{it}, se modela como una función logística de covariables, basada en Zhao & Prentice (1990), quienes describen un conjunto de ecuaciones *scoring* para la estimación conjunta de los parámetros marginales y los parámetros de asociación condicional, empleando máxima verosimilitud a través del algoritmo EM. La aplicación presentada en dicho artículo corresponde al mismo utilizado en este artículo, lo cual permite una comparación ilustrativa de los dos métodos. Teniendo en cuenta el modelo dado en (22), se presentan los resultados en la tabla 5.

Tabla 5: Estimadores de parámetros empleando modelos marginales de acuerdo a Fitzmaurice et al. (1994).

Parámetro	Estimación	Error estándar	Límites del 95 % de confianza	Wald	Valor p
Intercepto	-1.356	0.098	$(-1.5482; -1.1642)$	-13.848	<0.0001
g	-0.043	0.138	$(-0.3136; 0.2276)$	0.310	0.7620
E_L	0.142	0.063	$(0.0154; 0.2656)$	2.272	0.0231
E_C	0.014	0.035	$(0.0546; 0.0826)$	0.396	0.6922
gE_L	0.162	0.096	$(-0.0262; 0.3503)$	1.684	0.0923
gE_C	-0.089	0.049	$(-0.1852; 0.0071)$	-1.806	0.0709

3. *Metodología propuesta en este artículo.* Los resultados bajo esta metodología están dados en la tabla 3.

Las tablas 3 y 5 presentan los estimadores de los parámetros y los errores estándares para el modelo (22) basado en 1014 individuos, a diferencia de lo presentado en la tabla 4, la cual está basada en tan solo 460 individuos con datos completos. Esta supresión de información en el método de caso completo, por la falta de precisión en las estimaciones, puede llevar a conclusiones erradas acerca de los efectos del género y la edad en el riesgo de obesidad.

Uno de los supuestos que tiene en cuenta la metodología propuesta por Fitzmaurice et al. (1994) es la ocurrencia de faltantes bajo un patrón monótono, supuesto que condiciona ciertos conjuntos de datos. En la metodología propuesta no se condiciona el patrón de datos faltantes, lo cual permite su aplicación en una más amplia gama de conjuntos de datos.

Los resultados de los anteriores análisis sugieren que hay un crecimiento lineal (sobre la escala logit) en la razón de obesidad en el tiempo, excepto en el método de caso completo. No existen diferencias estadísticas entre niños y niñas a través de las metodológicas de caso completo y de Fitzmaurice et al. (1994), pero sí en el método de imputación propuesto (ver figura 1). Esto último se debe posiblemente a la estrategia de imputación planteada, ya que en los diferentes tiempos la covariable género está presente. Sin embargo, es necesario resaltar que debido al desconocimiento del valor poblacional de los datos, sin un proceso de simulación no es posible hacer recomendaciones acerca de cuál método usar.

Para hacer un estudio comparativo más justo de la metodología propuesta con otras existentes, sería necesario recurrir a procesos de simulación que conlleven a conclusiones generales, que caractericen propiedades estadísticas del método con respecto a sesgo y error cuadrático medio, lo cual será evaluado en futuras publicaciones.

5. Conclusiones

En este artículo se propuso una metodología para la estimación de información faltante en diseños de medidas repetidas con respuesta binaria basada en máxima verosimilitud, desde un enfoque univariado, lo cual permite un manejo de la

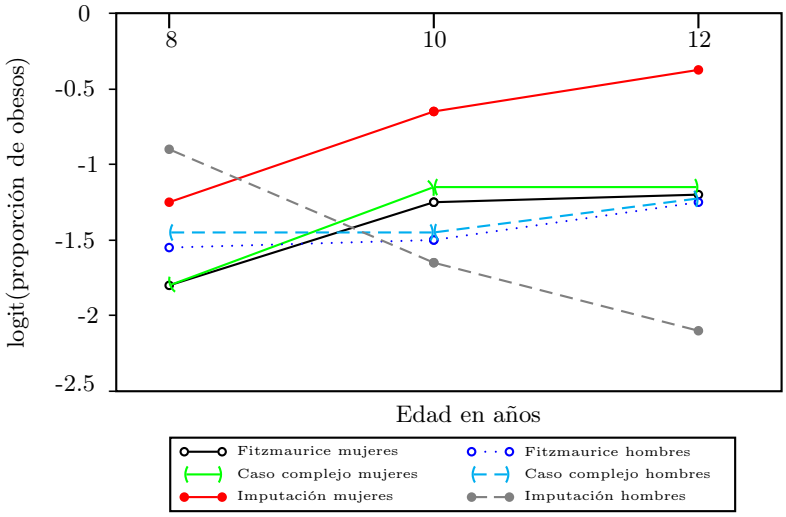

FIGURA 1: Gráfico del logit(proporción de obesos estimada) contra la edad.

información más sencillo con respecto a las desarrolladas en forma multivariada.

El método propuesto es útil en conjuntos de datos con porcentajes altos de información faltante, debido al proceso de estimación e imputación univariada, en donde los faltantes dependen de las covariables, las cuales son completamente observadas. No obstante, es necesario tener cuidado, ya que estos altos porcentajes podrían ocasionar demora en los procesos iterativos, y posiblemente se podrían tener problemas de convergencia y poca precisión en las estimaciones. Además, la metodología propuesta no está condicionada a un patrón particular de datos faltantes, permitiendo su aplicación en una más amplia gama de información faltante.

Un análisis de sensibilidad, explorando los resultados en las estimaciones usando análisis de correspondencias u omitiendo el problema de multicolinealidad, no se evaluó en este artículo, por lo cual es necesario realizar un proceso investigativo más detallado al respecto.

Agradecimientos

Este artículo se deriva de la tesis de maestría en estadística del primer autor Ayala (2006).

Agradecemos a los evaluadores por sus valiosas y oportunas observaciones que permitieron mejorar el artículo y, al licenciado en Matemáticas y Física Luis Jaime Salazar R. por su valiosa colaboración en el desarrollo del programa en MatLab. Este trabajo está enmarcado dentro del proyecto de investigación "Estadística aplicada a la investigación experimental, industria y biotecnología".

Recibido: abril de 2007
Aceptado: noviembre de 2007

Referencias

Ayala, S. Y. (2006), Estimación e Imputación de Datos Faltantes en Diseños de Medidas Repetidas con Respuesta Binaria o Poisson, Tesis de Maestría, Estadística, Universidad Nacional de Colombia, Facultad de Ciencias, Departamento de Estadística, Bogotá.

Bartlett, M. S. (1937), 'Some Examples of Statistical Methods of Research in Agricultura and Applied Botany', *Journal of Royal Statistical* **4**, 137–170.

Chen, H. Y. & Little, R. (1999), 'A Test of Missing Completely at Random for Generalised Estimating Equations with Missing Data', *Biometrika* **86**(1), 1–13.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum Likelihood from Incomplete Data Via the EM Algorithm', *Journal of the Royal Statistical* **39**, 1–38.

Diggle, P. J., Liang, K. Y. & Zeger, S. L. (1994), *Analysis of Longitudinal Data*, Oxford.

Fitzmaurice, G., Laird, N. & Lipsitz, S. (1994), 'Analysis Incomplete Longitudinal Binary Responses: A Likelihood-Based Approach', *Biometrics* **50**(3), 601–612.

Fitzmaurice, G., Laird, N. & Ware, J. (2004), *Applied Longitudinal Analysis*, Wiley Series in Probability and Statistics, New York.

Healy, M. & Wesmacott, M. (1956), 'Missing Values in Experiments Analized on Automatic Computers', *Applied Statistic* **5**, 203–206.

Horton, N. & Lipsitz, S. (2001), 'Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables', *American Statistical Association* **55**(3), 244–254.

Ibrahim, J. (1990), 'Incomplete Data in Generalized Linear Models', *Journal of American Statistical Association* **85**(411).

Lipsitz, S., Ibrahim, J. & Fitzmaurice, G. (1999), 'Likelihood Methods for Incomplete Longitudinal Binary Responses with Incomplete Categorical Covariates', *Biometrics* **55**, 214–223.

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

McCullagh, P. & Nelder, J. (1989), *Generalized Linear Models*, second edn, CRC Press, New York.

Park, T. & Davis, C. (1993), 'A Test of the Missing Data Mechanism for Repeated Categorical Data', *Biometrics* **49**(2), 631–638.

Park, T. & Lee, S. Y. (1997), 'A Test of Missing Completely at Random for Longitudinal Data with Missing Observations', *Statistics in Medicine* **16**, 1859–1871.

Peña, D. (2002), *Análisis de datos multivariantes*, McGraw-Hill, Madrid.

Srivastava, M. & Carter, E. (1986), 'The Maximum Likelihood Method for Non-Response in Sample Surveys', *Statistics Canada* **12**, 61–72.

Wolson, R. F. & Clarke, W. R. (1984), 'Analysis of Categorical Incomplete Longitudinal Data', *Royal Statistical Society* **147**, 87–99.

Yang, X., Li, J. & Shoptaw, S. (2005), 'Multiple Partial Imputation for Longitudinal Data with Missing Values in Clinical Trials'. Paper 2005010102.

Yates, F. (1933), 'The Analysis of Replicate Experiments When the Field Results are Incomplete', *Empire Journal of Experimental Agriculture* **1**, 129–142.

Zhao, L. P. & Prentice, R. L. (1990), 'Correlated Binary Regression Using a Quadratic Exponential Model', *Biometrika* **77**, 642–648.

Zorn, C. J. (2001), 'Generalized Estimation Equation Model for Correlated Data: A Review with Application', *American Journal of Political Science* **45**(2), 470–490.