COMPARACIÓN EMPÍRICA DE LA EFICIENCIA DE ALGUNAS TÉCNICAS DE TRATAMIENTO DE DATOS FALTANTES APLICADAS AL ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE

Claudia J. Polo Yepes

Analista Estadístico. Coomeva Financiera, Gerencia De Riesgo, Cali.
claupol1@yahoo.com

Roberto Behar Gutiérrez

Profesor Titular. Universidad del Valle, Escuela de Ingeniería Industrial y Estadística, Cali
olaya@univalle.edu.co

Javier Olaya Ochoa

Profesor Titular. Universidad del Valle, Escuela de Ingeniería Industrial y Estadística, Cali
robehar@pino.univalle.edu.co

RESUMEN

En este trabajo se caracterizan algunas de las más comunes técnicas estadísticas de tratamiento de datos faltantes y se comparan empíricamente a través de una simulación para determinar cuál es la más eficiente en la estimación de los coeficientes de regresión y de determinación de un modelo lineal de regresión múltiple con dos variables explicativas y un patrón univariado de datos faltantes sobre una de las variables. Se midieron la eficiencia relativa a través del error cuadrático medio y con base en las estimaciones por intervalos de confianza de los coeficientes de regresión a través de su cubierta y amplitud. Los resultados sugieren que análisis de casos completos, debe ser usado cuando el porcentaje de faltantes es pequeño y bajo mecanismos completamente al azar. En general, para todas las técnicas cuando el porcentaje aumenta, las estimaciones de los coeficientes de determinación y regresión se vuelven ineficientes alterando la cubierta y amplitud de los intervalos de confianza de los coeficientes de regresión. El análisis de casos disponibles y la imputación de la media no condicional y condicional no son recomendables porque producen en muchos casos estimaciones ineficientes de los coeficientes de determinación y de regresión. El algoritmo EM es una técnica eficiente y menos sensible a mecanismos que no son completamente al azar.

Palabras clave: Algoritmo EM, Análisis de casos completos, Análisis de casos disponibles, Coeficientes de regresión, Simulación, Imputación de media condicional, Imputación de media no condicional.

ABSTRACT

In this paper we have characterized some of the most common statistical tools for handling missing data. We make an empirical comparison through simulation of these tools in order to determine which one is the most efficient on the multiple linear regression parameter estimation process. We fitted a linear model with two independent variables and a univariate missing data pattern on X1. We measured the relative efficiency using the mean square error and based on confidence intervals estimates of the regression coefficients through their cover and amplitude. Results suggest that that complete cases analysis should be used whenever the percentage of missing data is low and under completely random mechanisms. Generally speaking, all techniques become inefficient as the percentage of missing data grows. This fact introduces some alterations on cover and amplitude of regression coefficients interval estimates. We do not recommend the available cases analysis and imputation of non-conditional and conditional mean because they produce inefficient estimates of determination and regression coefficients very often. To the contrary, the EM algorithm is a more efficient and less sensitive technique to not completely random mechanisms.

Keywords: EM Algorithm, Complete cases analysis, Available cases analysis, Regression coefficients, Simulation, Conditional mean imputation, Non-conditional mean imputation.

1. Introducción

Cuando el objetivo de un investigador es estudiar la relación existente entre una variable dependiente y una o más independientes a través de un modelo de regresión lineal, se puede aplicar el método de mínimos cuadrados lineales para encontrar los estimados de los parámetros poblacionales β del modelo en la situación ideal en la cual ninguna variable presenta valores faltantes.

Sin embargo cuando esta situación se presenta, la principal dificultad, es cómo encontrar los estimados de los coeficientes β y su precisión asociada, teniendo en cuenta los problemas de eficiencia, complicación en el tratamiento de los datos y sesgos ocasionados por la presencia de valores faltantes. El investigador debe elegir un método estadístico para solucionar el problema, teniendo en cuenta el efecto que tal metodología tendrá en la estimación del modelo y que una elección incorrecta puede llevarlo a conclusiones erróneas sobre su objeto de estudio.

Desde esta perspectiva, este trabajo tuvo como propósito caracterizar de forma general algunos de los métodos existentes para remediar el problema y determinar de forma empírica cuales de estos métodos son más eficientes en la práctica para la estimación de los coeficientes de regresión y de determinación de un modelo de regresión lineal, teniendo en cuenta el caso de los Ítem de no respuesta sobre las variables independientes.

2. Datos faltantes

Un conjunto de datos rectangular[1] contiene valores faltantes, cuando tal conjunto presenta: a) Información perdida o sin respuesta, b) Información errónea e inconsistente o c) información atípica. No se considera un dato faltante aquel dato que es faltante por diseño; por ejemplo, una pregunta de un cuestionario que solo es respondida por un individuo que cumple ciertas características, en este caso, los individuos que no cumplan la característica no responderán la pregunta y por tanto se tendrán valores faltantes por diseño.

Se han identificado los siguientes tipos de datos faltantes: 1) Registros que tienen todas las variables sin respuesta. En muestreo se le denomina No-respuesta unitaria ó Ausencia de respuesta por unidad; 2) Registros que tienen algunas variables con valores faltantes. En muestreo se denomina como ítem de no-respuesta ó Ausencia de respuesta por elemento. En el primer caso, estos registros son eliminados y se utilizan procedimientos de ponderación sobre los casos observados para compensar la falta de los datos. En el segundo caso, el analista de los datos tiene varias opciones: a) omitir los valores faltantes del análisis, b) "llenar" los valores faltantes con valores plausibles o c) utilizar técnicas basadas en modelos estadísticos para encontrar estimaciones para los datos faltantes.

En la opción a), se ignora la presencia de valores faltantes y estos son omitidos del análisis. Esta es la opción por defecto de muchos paquetes estadísticos. La opción b) corresponde a técnicas de imputación simple, en las cuales el analista utiliza los datos resultantes como si los valores faltantes no hubieran existido. A excepción de algunos casos donde se puedan mantener los supuestos de estas técnicas, en general su uso no es una buena estrategia. La opción c), corresponde a un grupo de técnicas fundamentadas estadísticamente que sirven para completar los datos faltantes asumiendo un modelo o distribución de probabilidad.

Los principales problemas que surgen cuando se tienen datos faltantes son a) Pérdida de Eficiencia, b) Sesgos y, c) Complicación en el tratamiento de los datos y su análisis.

La pérdida de eficiencia se produce en dos sentidos: 1) Cuando el análisis se basa en los registros que tienen todas las variables completas y estos difieren de aquellos casos incompletos que han sido descartados, se produce un cambio en la varianza dado que disminuye el tamaño de muestra analítica y se afecta la eficiencia en la precisión de las estimaciones. 2) Schafer (1997) considera que: en conjuntos de datos multivariados, omitir registros incompletos puede ser ineficiente debido a las grandes cantidades de información descartada de las variables que están completas.

[1] El cual contiene en las filas los individuos o unidades y en las columnas las variables

Los sesgos aparecen cuando el análisis se realiza solamente sobre todos los casos observados completamente. En este caso se está haciendo el supuesto que las unidades con elementos faltantes son similares a las unidades que tienen las respuestas de todas las variables, ignorando (si existe) cualquier relación o factor que esté incidiendo en que estos valores se produzcan; este supuesto generalmente es erróneo y si lo es, puede conducir a sesgos en las estimaciones. Lohr (2000) considera que si el supuesto no es erróneo, los casos completos se consideran como representativos de la muestra seleccionada y por tanto, también las estimaciones que se hagan con ellos; el problema de este supuesto, es que como no se tiene acceso a los datos faltantes, no puede probarse en la realidad práctica.

Y la complicación en el tratamiento de los datos y su análisis ocurre cuando se tienen bases de datos incompletas que deben ser tratadas con herramientas estadísticas diseñadas para bases de datos completas y el analista debe decidir qué hacer con los valores faltantes.

Adicional a estos problemas, Rubin (1988) considera que también está la consistencia en los resultados a través de diferentes usuarios, ya que si el constructor de la base de datos la entrega al público conteniendo estos valores, todos los usuarios, pueden utilizar diferentes herramientas para solucionar el problema y hallar resultados diferentes, perdiéndose la consistencia, uniformidad y confiabilidad de la información.

2.1 Patrones de datos faltantes

Los patrones de datos faltantes se refieren en general, a la forma como aparecen o como pueden reorganizarse estos valores dentro de todo el conjunto de datos. Según Schafer (1997), un patrón de faltantes corresponde a una única combinación de estatus de respuesta, es decir observada o faltante, para cada variable. De acuerdo con Little (1992) los patrones son: Patrón univariado. En este patrón los valores faltantes se hallan en solo una variable para cualquier registro.
Patrón monótono. Un patrón monótono se caracteriza porque la matriz de datos puede ser reorganizada de tal forma que haya una jerarquía entre los valores faltantes; Según Schafer y Graham (2002) esta jerarquía es de tal forma que si se tienen p variables, pueden ser ordenadas de tal forma que si X_j es faltante para una unidad, entonces $X_{j+1},...,X_p$ son también faltantes, para j=1,2,..,p.

Patrones especiales, en los cuales dos variables no son nunca observadas simultáneamente. Estos patrones ocurren cuando dos muestras que contienen datos cada una sobre cada variable, son unidas dentro de una sola base de datos.

Patrón general. Este patrón no presenta ninguna estructura especial.

2.2 Distribución o mecanismo de datos faltantes

Schafer (1997), indica que para un conjunto de datos rectangular de dimensión nxp denominado Z, donde $Z=(Z_{obs},Z_{mis})$ con (Z_{obs}) valores observados y (Z_{mis}) valores faltantes, se tiene una matriz indicadora de datos faltantes R de dimensión nxp, donde R_{ij} =1 si Z_{ij} es observado y R_{ij} =0 si Z_{ij} es faltante. De esta forma, Schafer (1997) plantea que en general, "no se espera que la distribución de R sea no relacionada a Z, entonces se impone un modelo de probabilidad para R, $f(R|Z,\varphi)$" y este modelo es lo que se denomina Mecanismo de datos faltantes.

La matriz R es tomada, por tanto, como un conjunto de variables aleatorias con una distribución de probabilidad conjunta; según Schafer y Graham (2002), no se puede tener una distribución particular específica para R, pero se está de acuerdo en que R tiene una distribución.

Schafer y Graham (2002) enuncian que la forma de R depende del patrón de datos faltantes:

- Patrón univariado, R puede ser un ítem binario simple para cada unidad, indicando si z es observado (R=1) o faltante (R=0);
- Si el patrón es monótono, R puede ser una variable entera (1,2,..,p) indicando el más alto j para el cual Z_j es observado;
- Para el patrón general, R corresponde a una matriz de la misma dimensión de la matriz de datos, compuesta por ceros y unos.

La naturaleza de las relaciones entre el ser faltante y los valores de los ítem faltantes permite clasificar la distribución de R en: Al azar (MAR), Completamente al azar (MCAR) ó No ignorable (NI). Los datos faltantes son faltantes al Azar (MAR) Si la distribución de R depende (o puede depender) solamente de los valores observados de Z, es decir de Z_{obs}, pero no de los valores faltantes Zmis: $f(R|Z,\varphi)=f(R|Z_{obs},Z_{mis},\varphi)=f(R|Z_{obs},\varphi)$, para todo Zmis. Los datos faltantes son Completamente al Azar (MCAR) si la distribución de R no depende de los valores observados o faltantes de Z: $f(R|Z,\varphi)=f(R|Z_{obs},Z_{mis},\varphi)=f(R|\varphi)$.

Este mecanismo está enunciando que los datos faltantes son faltantes al azar (MAR) y los datos observados son observados al azar (OAR), es decir, que el hecho que una observación sea faltante no depende de ningún valor observado ni faltante en los datos de las variables.

Cuando los datos faltantes son MAR o MCAR el mecanismo de datos faltantes es ignorable para inferencias basadas en verosimilitud. Cuando los datos faltantes son MCAR este mecanismo es ignorable tanto para las inferencias basadas en muestreo como para las inferencias basadas en verosimilitud. Cuando los datos faltantes no son ni MAR ni MCAR, el mecanismo es No ignorable (Little y Rubin, 1987). No ignorable significa por tanto que el mecanismo está relacionado a los valores faltantes (Grace, 2001) y no es producto del azar. Cuando esto sucede, el problema es modelar tal mecanismo imponiéndole una distribución que debe ser cercana a la realidad.

2.3 Relaciones entre patrones y mecanismos

Schafer y Graham (2002) plantean las siguientes relaciones entre patrones y mecanismos:

Patrón univariado. Supóngase que se tienen p variables $X_1,X_2,...,X_p$ que han sido observadas completamente y sea Y una variable que es faltante para algunos individuos, entonces: El mecanismo MCAR significa que la probabilidad que Y sea faltante para un individuo en particular no depende de su o sus valores de X y Y respectivos (y, por independencia, no depende de los X o Y de otros individuos); el mecanismo MAR significa que la probabilidad que Y sea faltante puede depender de las variables X pero no de Y; Un mecanismo No ignorable, significa que la probabilidad que Y sea faltante depende sobre Y.

Patrón monótono. Supóngase que se tienen p variables $Y_1,Y_2,...,Y_p$ tales que si Y_j es faltante para una unidad, entonces $Y_j+1,...,Y_p$ son también faltantes, para j=1,2,..,p, entonces: El mecanismo MCAR significa que la probabilidad que Y_j sea faltante, es no relacionada a cualquier variable presente en el sistema; El mecanismo MAR significa que esta probabilidad puede estar solamente relacionada a $Y_1,...,Y_{j-1}$; Un mecanismo No ignorable significa que esta probabilidad está relacionada a $Y_j,...,Y_p$.

Patrón general. El mecanismo MCAR al igual que en los otros patrones, requiere independencia entre las variables y el ser faltante; El mecanismo MAR significa que las probabilidades de los individuos de responder pueden depender de su propio conjunto de ítems observados, conjunto que puede variar de un individuo a otro.

2.4 Algunas estrategias para el tratamiento de datos faltantes

2.4.1. Procedimientos basados en la omisión de valores faltantes

Análisis de Casos Completos (ACC). Este procedimiento es conocido como "Listwise deletion" o "Eliminación por lista". Se basa en el supuesto que los datos son completamente al azar (MCAR). Es el procedimiento que realizan automáticamente los principales paquetes de software estadístico, tal como SPSS y consiste en el descarte u omisión de todos los casos o registros con alguna información faltante en cualquier variable. Esta técnica es de fácil implementación en el sentido que se pueden utilizar los análisis estadísticos estándar para datos completos sin necesidad de hacer modificaciones y permite comparar estadísticos univariados porque todos son calculados sobre un mismo tamaño de muestra (el obtenido después de ignorar los casos con valores faltantes). Sin embargo tiene como desventajas la perdida de una gran cantidad de información en las variables

que están completas, siendo aun más crítico cuando el número de variables es grande. Adicionalmente existe un cambio en la varianza debido a la disminución del tamaño de muestra y usualmente, se producen sesgos en las estimaciones.

Análisis de Casos Disponibles (ACD). Emplea todas las observaciones que tienen valores válidos para la variable de interés en cada momento (Puerta, 2002), intentando no descartar información tal como lo hace ACC. La principal ventaja de este método es que usa todos los valores de la variable dependiendo del análisis involucrado; sin embargo, como mencionan Little y Rubin (1987), la principal desventaja de este método es que la base muestral cambia de variable a variable de acuerdo al patrón de faltantes. Por ejemplo, el cálculo de las medias y las varianzas, se hará con base en los valores observados (ó valores disponibles) de cada variable, mientras que en el caso de las covarianzas o correlaciones, este método utiliza solamente los pares de valores disponibles. Los estimados son no sesgados solamente si el supuesto MCAR se cumple (Grace, 2001). Esta técnica puede dejar matrices de covarianzas o correlaciones que no son positivas definidas.

2.4.2. Procedimientos basados en imputación simple

Estos procedimientos completan cada valor faltante con un solo valor plausible y con la base de datos resultante realizan análisis estadísticos por métodos estándar. Según Schafer (1997), después que se han imputado los valores faltantes no se involucra ninguna medida adicional de incertidumbre de las predicciones.

Las principales ventajas son (Grace, 2001):

- El problema es "solucionado" al principio
- No descarta información
- Le permite al usuario proceder usando análisis y software de datos completos
- Si el constructor de la base de datos es quien realiza la imputación de los valores faltantes, entonces se tendría consistencia en los análisis que puedan hacer los futuros usuarios (Rubin, 1988).

Las principales desventajas son:

- Una vez que se han imputado los valores, estos en general son tomados como si fueran los valores reales, esta situación usualmente subestima sistemáticamente la incertidumbre asociada a los datos faltantes. Entonces, como la fuente extra de error es ignorada, se obtienen errores estándar (e intervalos de confianza) y p-valores demasiado pequeños (Grace, 2001), así como tasas de error tipo I más altas que los niveles nominales (Schafer y Olsen, 1998).
- Estos métodos pueden causar sesgos sistemáticos (Smith, 2005)
- Según Pérez (2000), son métodos ad hoc, es decir, tienen una aplicación para cada caso particular y pueden necesitar ajustes.

Estas desventajas se vuelven más graves cuando la tasa de información faltante y el número de parámetros aumenta (Schafer, http://www.stat.psu.edu/~jls/session1.pdf).

2.4.3. Imputación de la Media No Condicional (IMNC)

Este procedimiento consiste en que cada valor faltante para una variable es reemplazado por el promedio calculado con los casos completos de esa misma variable. Según Schafer (www.stat.psu.edu/~jls/session1.pdf) este procedimiento tiene como desventajas que cambia la distribución marginal de la variable imputada, así como produce covarianzas y correlaciones con otras variables alteradas, presenta errores estándar de los datos imputados demasiado pequeños y existe incertidumbre acerca de la población. Además, este procedimiento se basa sobre un supuesto MCAR. La imputación de la Media es consistente en los primeros momentos, pero se obtienen estimaciones sesgadas de la matriz de varianzas y covarianzas (Σ).

2.4.4. Imputación de la Media Condicional (IMC).

Con este método, se busca crear un modelo de regresión de cada variable que tenga valores faltantes (variable dependiente) sobre las variables con información completa (explicativas). Según DIAZ (2005), este procedimiento se puede iterar hasta que haya convergencia en las estimaciones. Como ejemplo, se tienen tres variables Y, X_1 y X_2 donde X_2 está sujeta a valores faltantes. Para estimar los valores faltantes de X_2, se realiza una regresión lineal de X_2 sobre X_1 (con los casos completos) y con el modelo resultante se estima el valor de X_2 para cada registro u observación. Es decir,

$$\hat{X}_{i2} = E(X_{i2} \mid X_{i1}) = \beta_0 + \beta_1 X_{i1}$$

La imputación de la media condicional puede hacerse condicionando solamente sobre las variables X o condicionando sobre las variables X y Y (Little, 1992); este trabajo solo tendrá en cuenta la primera opción en la cual la imputación se hace condicionando sobre las variables X, ya que según este mismo autor, parece tramposo utilizar información de Y para completar los X faltantes cuando el objetivo es la regresión de Y sobre las X; al hacer esto, se obtienen estimados de la regresión sujetos a sesgos.

Schafer y Graham (2002) consideran que esta técnica no es apropiada cuando el objetivo es analizar las covarianzas o correlaciones entre las variables, ya que exagera la intensidad de la relación entre las variables imputadas y observadas; además el coeficiente de determinación entre los valores imputados es igual a uno ya que se asume una relación lineal perfecta en la cual todos los valores que han sido imputados caen sobre la recta de regresión. Es importante anotar, que según Laaksonen (1999) esta técnica sobreestima la varianza, en cuyo caso es usual adicionar un término de ruido para predecir los valores y evitar el problema.

2.4.5. Procedimientos basados en modelos

Los procedimientos basados en modelos: máxima verosimilitud, imputación múltiple y los métodos bayesianos, "asumen una distribución para los datos completos, es decir, los datos faltantes y observados. Intuitivamente, este modelo describe las relaciones entre las variables, y cuando son combinados con los datos observados, pueden ser usados para "completar" los datos" (Insightful Corporation, 2001); las principales ventajas es que son más eficientes y aún más robustos frente al no cumplimiento de supuestos, permiten evaluar supuestos del modelo y estimar la varianza de los parámetros estimados. Este estudio se enfocó en el algoritmo EM, se presentan sus características siguiendo principalmente los desarrollos de Little y Rubin (1987).

2.4.6. Algoritmo EM

Cuando se debe realizar estimación máximo verosímil (ML) para datos incompletos, Little y Rubin (1987) plantean que en un sentido formal no hay diferencias entre la estimación ML para datos incompletos y la estimación ML para datos completos: se halla la verosimilitud para los parámetros basada sobre los datos incompletos y los estimados ML son encontrados al resolver la ecuación de verosimilitud. De esta forma, el método sirve en dos sentidos: proporcionando estimaciones adecuadas de los parámetros y proporcionando estimaciones para los valores faltantes.

Sin embargo, se sabe que los faltantes presentan un mecanismo que debe ser tenido en cuenta para formar la distribución conjunta de los datos y del mecanismo $f(Z,R \mid \theta,\varphi) = f(Z \mid \theta) f(R \mid Z,\varphi)$ donde $Z=(Z_{ij})$ es la matriz de n observaciones por K variables (n x K), y R es la variable indicadora de Respuestas $R= (R_{ij})$, tal que,

$$R_{ij} = \begin{cases} 1, & Z_{ij}\, obs \\ 0, & Z_{ij}\, falt \end{cases}$$

Así la estimación máximo verosímil depende de este mecanismo $f(R \mid Z,\varphi)$ de tal forma que si se puede asumir que es ignorable se facilita la estimación dado que la verosimilitud podría calcularse sin tener en cuenta el modelo para R, situación que no comparten los mecanismos no ignorables.

Cuando se tienen patrones de datos generales, la tendencia utilizada es el algoritmo EM (Expectation Maximization), técnica general para encontrar estimados máximo verosímiles para modelos paramétricos cuando los datos no han

sido completamente observados (Schafer, 1997) y cuando los estimados ML no son explícitos. Frente al uso de algoritmos como el de Newton Raphson, el algoritmo EM es una mejor alternativa para la maximización de la verosimilitud logarítmica de los datos.

Entre las principales ventajas se tienen: facilidad para programarlo, interpretación estadística directa, estimadores insesgados de mínima varianza cuando el tamaño de la muestra incrementa, estimaciones consistentes y eficientes cuando los datos son MAR. Las principales desventajas son: la tasa de convergencia es lenta si hay muchos faltantes; no provee automáticamente errores estándar; y no siempre converge a un máximo único.

En general los pasos del algoritmo EM (Esperanza-Maximización) descritos por Little y Rubin (1987) son:

1. Reemplaza cada valor faltante por valores estimados, a partir de valor inicial de los parámetros (Paso E: Esperanza)
2. Estima los parámetros del modelo con los datos resultantes en el paso anterior (Paso M: Maximización)
3. Re estima los valores faltantes asumiendo que los nuevos parámetros estimados son correctos (Paso E: Esperanza)
4. Re estima los parámetros, iterando sucesivamente hasta la convergencia (Paso M: Maximización)

Cuando los datos se comportan según una familia de distribuciones exponenciales[2] (como la distribución normal), Little y Rubin (1987) plantean que el algoritmo EM adopta una interpretación relativamente simple y útil. En este caso, el paso E consiste en estimar los estadísticos suficientes de datos completos *s(Z)* a través de $s^{(t+1)} = E(s(Z) \mid Z_{obs}, \theta^{(t)})$ y el paso M calcula los nuevos estimados $\theta^{(t+1)}$ de θ como la solución de las ecuaciones de verosimilitud $E(s(Z) \mid \theta^{(t)}) = s^{(t)}$, donde *s(Z)* es reemplazado por $s^{(t)}$.

Al aplicar el algoritmo a una muestra aleatoria procedente de una distribución Normal multivariada, López (1995) plantea que se tendría el siguiente esquema. Sea $Z = [Z_1, Z_2, ..., Z_K]$ un conjunto de K variables aleatorias distribuidas Normal multivariante con medias $\mu=(\mu_1, \mu_2, ..., \mu_k)$ y matriz de varianzas covarianzas dada por:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2k} \\ \vdots & \vdots & & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk} \end{bmatrix}$$

Con información faltante, Z puede ser escrito como: $Z = [Z_{obs}, Z_{mis}]$ con $Z_{obs} = [z_{obs,1}; z_{obs,2}; \cdots; z_{obs,n}]$ y $z_{obs,i}$ es el conjunto de variables observadas para la observación *i* con *i=1,2,...,n*

Sean $\sum_{i=1}^{n} z_{ij}$ con j=1,2,...,K y $\sum_{i=1}^{n} z_{ij} z_{ik}$ con j,k=1,2,...,K los estadísticos suficientes para μ_j y $(\sigma_{jk})'$. Con el algoritmo EM en este caso se busca estimar los parámetros $\theta = (\mu, \Sigma)$ y las observaciones faltantes Z_{mis}. Por tanto en la iteración t, $\theta^{(t)} = (\mu^{(t)}, \Sigma^{(t)})$ corresponde a los estimados actuales de los parámetros.

El paso E de EM sería:

$$E\left(\sum Z_{ij} \mid Z_{obs}; \theta^{(t)}\right) = \sum_{i=1}^{n} Z_{ij}^{(t)} \quad \text{j=1,..., K}$$

donde

$$Z_{ij}^{(t)} = \begin{cases} z_{ij} & \text{si inf obs.} \\ E(Z_{ij} \mid Z_{obs,i}, \theta^{(t)}) & \text{si inf falt.} \end{cases} \quad \text{y}$$

$$E\left(\sum Z_{ij} Z_{ik} \mid Z_{obs}; \theta^{(t)}\right) = \sum_{i=1}^{n} Z_{ij}^{(t)} Z_{ik}^{(t)} + C_{jki}^{(t)} \quad \text{j,k=1,...,K}$$

donde

[2] La familia exponencial es: $f(Z \mid \theta) = b(Z) e(s(Z)\theta) / a(\theta)$, donde θ es un vector de parámetros de dimensión (px1), s(Z) denota un vector de estadísticos suficientes de datos completos de dimensión (1xp), a y b son funciones de θ y Z respectivamente

$$C_{jki}^{(t)} = \begin{cases} 0, & \text{si } Z_{ij} \text{ o } Z_{ik} \text{ obs.} \\ Cov(Z_{ij}Z_{ik} \mid Z_{obs,i}, \theta^{(t)}, & \text{si } Z_{ij} \text{ o } Z_{ik} \text{ obs.} \end{cases}$$

Los valores faltantes z_{ij} son por tanto reemplazados por la media condicional de z_{ij} dados el conjunto de valores $z_{obs,i}$ para esa observación (Little y Rubin, 1987).
El paso M, consiste en estimar los parámetros a través de las esperanzas calculadas en el paso E, es decir, $\mu_j^{(t+1)} = n^{-1}\sum_{i=1}^{n} z_{ij}^{(t)}$ con j=1,2,...,K.

$$\sigma_j^{(t+1)} = n^{-1}E\left(\sum_{i=1}^{n} z_{ij}z_{ik} \mid Z_{obs}\right) - \mu_j^{(t+1)}\mu_k^{(t+1)},$$

$$\sigma_j^{(t+1)} = n^{-1}\sum_{i=1}^{n}\left[\left(z_{ij}^{(t)} - \mu_j^{(t+1)}\right)\left(z_{ik}^{(t)} - \mu_k^{(t+1)}\right) + c_{jki}^{(t)}\right]$$

con j,k=1,2,..,K. En este caso $\theta^{(t)} = \left(\mu^{(t)}; \Sigma^{(t)}\right)$ y el proceso continúa hasta que $\left|\theta^{(t+1)} - \theta^{(t)}\right| < \delta$ con δ tendiendo a cero.

2.5 Criterio de comparación de las técnicas

Schafer y Graham (2002) plantean que con o sin datos faltantes el objetivo de un procedimiento estadístico debe ser el hacer inferencias válidas y eficientes acerca de una población de interés; los objetivos, según estos autores, no deben ser estimar, predecir o recobrar valores faltantes ni obtener los mismos resultados que se deberían obtener con datos completos. En este sentido, consideran que un tratamiento de datos faltantes no puede ser evaluado apropiadamente sin tener en cuenta la modelación, la estimación o los procedimientos de prueba en los cuales está inmerso.
Plantean usar el Error Cuadrático Medio (ECM) dado que esta medida combina el sesgo y la varianza de las estimaciones sobre muestras repetidas; se busca que ambas, el sesgo y la varianza, sean pequeñas, dado que los estimadores se vuelven más precisos y por tanto menos sensibles a errores en las mediciones de las variables.

Sin embargo, los autores en mención, plantean que también es necesario validar la honestidad de las estimaciones a través de pruebas de hipótesis; un procedimiento por intervalos de confianza debería cubrir el verdadero valor de Q con una probabilidad cerrada (igual) a la tasa nominal. Si la tasa de cubierta es segura la probabilidad de cometer error tipo I también será segura. Sujetos a la cubierta correcta, se busca que los intervalos sean angostos, porque entre más angostos sean los intervalos, menor es el error de tipo II y por tanto se produce un incremento en la potencia.

2.6 Algunas características de los métodos al aplicarlas en análisis de regresión

2.6.1. Análisis de Casos Completos

Una de las principales características de ACC en regresión es que cuando los faltantes dependen de las variables independientes X y no de la variable Y, se obtienen inferencias válidas para los coeficientes de regresión, es decir no sujetas a sesgos; propiedad que no conservan las demás técnicas de tratamiento de datos faltantes (Little, 1992). Sin embargo, Schafer y Graham (2002), plantean que esta propiedad no puede ser extendida a las medidas de asociación entre las variables, tal como el coeficiente de correlación, ni se puede extender a parámetros de la distribución marginal de Y; esto quiere decir que cuando el supuesto MCAR no se cumple, ACC puede dejar estimados sesgados de estos últimos parámetros.

2.6.2. Análisis de casos disponibles

Cuando se aplica análisis de casos disponibles en regresión, las propiedades del método se mantienen; sin embargo, el principal inconveniente de esta técnica es que la matriz de varianza covarianza estimada de las X no necesariamente es positiva definida y cuando no lo es, las pendientes estimadas resultantes son indeterminadas. Según Little (1992) este problema es aún más severo cuando las variables X son altamente correlacionadas.

2.6.3. Imputación de la Media

Cuando se desea calcular un análisis de regresión, se imputan primero los valores faltantes a través de la imputación de la media no condicional o condicional con los casos completos y después la ecuación de regresión es calculada sobre los datos completados, utilizando mínimos cuadrados ordinarios (OLS) o mínimos cuadrados ponderados (WLS). Sin embargo, Little (1992) plantea que sobre los datos imputados es recomendable realizar mínimos cuadrados ponderados.

Este método produce sesgos en la matriz de varianzas covarianzas; Little (1992), plantea, que con este método, las inferencias a través de pruebas e intervalos de confianza son distorsionadas por los sesgos y la exagerada precisión.

2.6.4. Algoritmo EM

Bajo supuestos de normalidad, este método es consistente cuando los faltantes dependen sobre Y. Además, los estimados máximo verosímil permanecen válidos cuando los datos son MAR y en particular cuando los faltantes dependen de las covariables observadas completamente y cuando también dependen de Y – asumiendo que los valores faltantes están solo en las variables independientes (Little, 1992). Por tanto, los estimadores son consistentes y eficientes si el modelo de base es correcto; Más aún, hay evidencia estadística que estos métodos funcionan bien aún cuando el modelo no es totalmente correcto (Pérez, 2000)

2.7 Software

Los grandes avances computacionales han hecho que en la actualidad, se pueda contar con sofisticadas herramientas estadísticas para el tratamiento de datos faltantes. Estas herramientas ya se encuentran disponibles en los principales paquetes de software estadístico tales como SAS V.8.1, SPSS (desde la versión 11.5), SPLUS V.6.0, entre otros, también existe software especializado en el tratamiento de datos faltantes, como el caso de SOLAS V.8.1 y NORM. Entre los paquetes de distribución gratuita se destacan R V.2.3.1, NORM, AMELIA, IVEWARE y Mx.

3. Metodología

Se generó con el software EASYREG una población de N=1.000 observaciones bajo el modelo de regresión $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, donde se asumió $\beta_0 = \beta_1 = \beta_2 = 1$, con X_1 y X_2 bajo una distribución conjuntamente normal estándar y una covarianza de 0.5; el error aleatorio del modelo ε se distribuyó Normal estándar, manteniéndose bajo los supuestos del modelo de regresión, con media cero y varianza constante, así como independiente de las variables X's.. Después de generada la población se realizaron pruebas con el software SPSS para verificar que cumplía con los supuestos bajo los cuales fue generada y se procedió a ajustar el modelo de regresión encontrando los parámetros poblacionales.

Tabla 1: Estadísticas descriptivas de la población generada

	Media	Desv. Est.	Correlaciones X1	Correlaciones X2	Covarianza X1	Covarianza X2
Y	1,1563	2,0001	0,7506	0,7399	1,4842	1,4642
X1	0,0820	0,9887		0,4968		0,4859
X2	0,0416	0,9894				

Tabla 2: Modelo de regresión calculado de la población

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	1,031	,032		31,962	,000	,968	1,094
	X1	1,029	,037	,509	27,449	,000	,955	1,102
	X2	,985	,037	,487	26,300	,000	,911	1,058

Tabla 3: Estadísticos de Colinealidad del Modelo de regresión

Model		Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)		
	X1	,753	1,328
	X2	,753	1,328

Para la simulación se utilizó un diseño experimental no estadístico de 5x4x3x3 para un total de 180 tratamientos.

Las variables consideradas fueron:

1. Técnicas de tratamiento de datos faltantes: Análisis de casos completos (ACC), Análisis de casos disponibles (ACD), Imputación de la media no condicional (IMNC), Imputación de la media condicional (IMC) y Algoritmo EM (EM)
2. Mecanismos de datos faltantes: MCAR (MCAR), MAR sobre X (MAR X), MAR sobre Y (MAR Y) y Mecanismo No Ignorable (NI).
3. Porcentaje de datos faltantes sobre X_1: 0%, 10%, 30%, 60%, el nivel 0%, corresponde al caso muestral en el cual los datos están completos y sirvió de comparación con los resultados obtenidos después de aplicar las técnicas de tratamiento de datos faltantes.
4. Tamaño de muestra: 50, 100 y 200.

Se generaron por cada uno de los 180 tratamientos resultantes, 200 muestras aleatorias simples sin reemplazamiento, las cuales corresponden a las réplicas del experimento.

Básicamente, el proceso que se llevó a cabo fue: 1) Generar cada uno de los 180 conjuntos de datos y calcular para cada uno el modelo de regresión correspondiente, 2) Producir pérdida artificial de datos sobre la variable X_1 en cada uno de las muestras resultantes utilizando las variables porcentaje y mecanismo de datos faltantes, 3) Utilizar algunas técnicas de tratamiento de datos faltantes para estimar los valores perdidos; 4) Sobre los conjuntos de datos resultantes, ajustar nuevamente el modelo de regresión y comparar los resultados con los obtenidos antes de eliminar datos.

El análisis de la información se hizo con base en dos medidas: 1) Error cuadrático medio, 2) intervalos de confianza para las estimaciones de los coeficientes de la regresión. Las variables dependientes a analizar fueron los coeficientes de regresión (β_0, β_1 y β_2) y el coeficiente de determinación (ρ) de cada modelo.

- Error cuadrático Medio

Se calculó el Error cuadrático medio (ECM) de las estimaciones de los parámetros ρ^2, β_0, β_1, β_2 de cada uno de los 180 tratamientos respecto a: 1) los parámetros del modelo poblacional y 2) las estimaciones de la muestra con 0% de datos incompletos. En el primer caso se denominó Error cuadrático medio del tratamiento (ECM_{tto}) y en el caso de las muestras completas Error cuadrático medio del referente (ECM_{ref}).

Si el muestreo aleatorio simple está bien implementado, se espera que el ECM_{ref}, sea siempre menor que el ECM_{tto}, dado que la pérdida de datos y su posterior estimación introduce un sesgo y un cambio en la variabilidad de las estimaciones; La mejor técnica será aquella que presente por tanto el más bajo ECM_{tto} respecto al ECM_{ref}.

Para tener un rango de cuantificación de que tan eficientes fueron las estimaciones, se creó una medida de eficiencia relativa de cada estimación, calculando el cociente entre el ECM_{ref} y el ECM_{tto}. Si al aplicar las técnicas de tratamiento de datos faltantes se obtienen inferencias eficientes (con varianza y sesgo pequeño) entonces el valor de la medida estará cercana a 1, entre más se acerque a cero las inferencias serán menos eficientes.

$$Eficiencia = \frac{ECM_{ref}}{ECM_{tto}} \text{ ,donde}$$

$$0 \leq Eficiencia \leq 1$$

- Intervalos de Confianza

Para cada técnica, se calcularon los intervalos de confianza de los coeficientes de regresión, así como la cubierta y amplitud de los mismos, antes y después de eliminar datos sobre la muestra.
La cubierta corresponde al porcentaje de intervalos que cubren o contienen el verdadero valor del parámetro; la amplitud o ancho se calcula como el promedio de la diferencia entre el límite superior y el límite inferior de cada intervalo. Se buscan intervalos con una cubierta igual o cercana a la tasa nominal y que además sean angostos.

Si se asume que se quiere probar la significancia de los coeficientes de la regresión a un 95% de confianza, Schafer y Graham (2002) establecen como una regla empírica el considerar la cubierta por estar seriamente afectada si esta se ubica por debajo del 90% cifra que corresponde al doble de la tasa de error nominal (2α, con α=0,5)

Para la simulación de mecanismos y porcentajes de datos faltantes, artificialmente se eliminaron datos de la variable X_1, de la siguiente forma:

Mecanismo MCAR

Para la simulación de este mecanismo, se creó una variable artificial U, basándose en la metodología propuesta por Little (1992), a partir de la siguiente fórmula: Sea $U = \alpha_1 X_1 + \alpha_2 X_2 + D$, donde X_1 y X_2, corresponden a las variables del modelo poblacional y D es una variable aleatoria independiente normal estándar.

Se hizo $\alpha_1 = \alpha_2 = 0$, dando como resultado: $U = D$ y por tanto $U \sim N(0,1)$. Para cada observación (y, x_1, x_2), se generó un valor de U de la distribución Normal Estándar y se utilizó como criterio de eliminación los puntos porcentuales de la función de distribución normal estándar acumulada, para cada porcentaje de datos faltantes considerado y cada tamaño de muestra. Se utilizaron los puntos porcentuales de la distribución Normal Estándar para garantizar que se pierden exactamente los porcentajes de datos faltantes establecidos (10%, 30% y 60%). Los criterios utilizados se resumen en la Tabla 4.

Mecanismo MAR X

Para eliminar las observaciones dependiendo de la variable X_2, el mecanismo se simuló utilizando

la variable U creada para el mecanismo MCAR, es decir $U = \alpha_1 X_1 + \alpha_2 X_2 + D$, igualando los coeficientes de las variable X_1 y X_2 a $\alpha_1 = 0$ y $\alpha_2 = 1$; Por tanto, $U = X_2 + D$.

El criterio para eliminar las observaciones sobre X_1, haciéndolas depender de X_2 fue utilizando la distribución normal asociada con la variable U, en este caso:

Sea $U = X_2 + D$, con $D \sim N(0,1)$. Para calcular la esperanza y la varianza de U, se debe asumir independencia entre X_2 y D y además, dado que X_1 y X_2 se distribuyen conjuntamente normales, la distribución marginal de X_2 es una distribución normal univariada con media y varianzas respectivas, en este caso con media cero y varianza uno; entonces la esperanza y la varianza de U son:

$E(U) = E(X_2 + D) = E(X_2) + E(D) = 0$ Y

$Var(U) = Var(X_2 + D) = Var(X_2) + Var(D) = 2$.

Por tanto $U \sim N(0,2)$.

Al igual que en el mecanismo MCAR, el mecanismo MAR X, se simuló utilizando como criterio de eliminación los puntos porcentuales de la función de distribución normal acumulada hallada para U, para cada porcentaje de datos faltantes considerado y cada tamaño de muestra.

MAR sobre Y

De la población generada se sabe que $Y \sim N\left[E(Y/X); \sigma^2\right]$ donde

$E(Y/X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = 1 + X_1 + X_2$ Y

$V(Y) = \sigma^2 = 1$.

Para hacer depender la pérdida de datos de X_1 sobre la variable Y, se utiliza una variable

Tabla 4. Criterios utilizados en la simulación.

Porcentaje	$Z_{(p\%)}$ Asociado	Criterio
10%	-1,281550794	Elimine el valor de X_1 en el registro i, con i=1,...,n, si al generar un valor al azar de U_i, se cumple que $U_i < Z_{(p\%)}$; Donde n = 50 , 100 y 200, p%=10, 30 y 60.
30%	-0,524401003	
60%	0,25334657	El proceso se detiene hasta que se alcanza el p% de datos faltantes especificado y se reinicia si al terminar los n datos no se ha alcanzado el p% deseado.

artificial de la siguiente forma: Sea U=Y, entonces $U \sim N(1 + X_1 + X_2;1)$; utilizando esta distribución normal, se calcula para cada registro el valor correspondiente del promedio de la variable U y con la distribución normal hallada se establece el punto porcentual de la distribución para el porcentaje de datos faltantes (10%,30%,60%).

Una vez calculado el punto porcentual, se genera un valor aleatorio de la función de distribución de U, si el valor de U es menor que el punto porcentual hallado entonces se elimina el valor de X_1, sino se pasa al siguiente registro hasta completar el porcentaje especificado de datos faltantes; el proceso se reinicia si al terminar los n datos no se ha alcanzado el p% deseado.

No Ignorable – NI

Para simular la alternativa no Ignorable, se hizo depender la ausencia de valores faltantes de la misma variable X_1; para hacer esto, se utilizó la variable U, creada para el caso MCAR, tomando $\alpha_1 = 1$ y $\alpha_2 = 0$, por tanto, $U = X_1 + D$.

Siguiendo la misma metodología planteada para el caso MAR sobre X_2, se asumió independencia entre X_1 y D, de tal forma que la esperanza y la varianza de U son:

$E(U) = E(X_1 + D) = E(X_1) + E(D) = 0$ y $Var(U) = Var(X_1 + D) = Var(X_1) + Var(D) = 2$. De esta forma, $U \sim N(0,2)$. El criterio para eliminar observaciones está dado por los puntos porcentuales de la función de distribución normal acumulada hallada para U, para cada uno de los porcentajes de datos faltantes, al igual que en el caso MAR X_2.

4. Resultados

La presentación de resultados se realizará utilizando para el coeficiente de determinación, una tabla resumen con los errores cuadráticos medios y los ER resultantes para cada tratamiento considerado. En el caso de los Coeficientes de regresión β_0, β_1 y β_2, se presentarán tablas resumen (ver apéndice) con el ER para los tres coeficientes y tratamientos considerados, así como tablas resumen con las Cubiertas y amplitudes de los intervalos de confianza de cada coeficiente (muestras completas y muestras con faltantes)

Efectos sobre la eficiencia de estimación del Coeficiente de Determinación

Los resultados para el coeficiente de determinación se presentan en la tabla 5. Para porcentajes de datos faltantes pequeños (10%), la variabilidad de los estimados del coeficiente de determinación es baja produciendo estimaciones cercanas a las estimaciones de la condición de datos completos para las técnicas ACC y EM y para los mecanismos MCAR y MARy y se refleja en las medidas ER respectivas que se ubican por encima del 80% para ambos mecanismos; sin embargo a medida que aumenta el porcentaje de faltantes al 30% y 60% se va reduciendo la eficiencia en la estimación del coeficiente hasta llegar a niveles en donde ninguna técnica realiza una estimación eficiente para ningún tamaño de muestra. Los mecanismos MARx y NI aún con porcentajes de faltantes pequeños y muestras grandes, tienden a sub-estimar los coeficientes de determinación respecto a la condición de datos completos, situación que se agudiza al aumentar el porcentaje de faltantes. La falta de eficiencia es mayor en los faltantes bajo NI.

La técnica ACD subestimó los coeficientes de determinación resultando en una eficiencia relativa menor a 0.8; solo en dos casos para el 10% de faltantes bajo MCAR y muestras de tamaño 50 y 100 se obtuvo un ER de 0.846 y 0.811 respectivamente. Comparada con ACC esta técnica es mucho menos eficiente en la estimación del coeficiente de determinación aún para porcentajes de datos faltantes pequeños (10%) y muestras grandes.

Al igual que con ACD, imputar los valores con la técnica IMNC produce estimados del coeficiente de determinación que subestiman el verdadero valor poblacional, aún para el mecanismo MCAR y un 10% de información faltante y se empeora a medida que crece el porcentaje de faltantes. Al igual que las otras técnicas, IMNC presenta la mejor estimación con el 10% de faltantes bajo MCAR y MARy, donde la eficiencia relativa fue de 0.847. En general esta técnica no es eficiente para estimar el coeficiente de determinación.

La IMC mejora la estimación del coeficiente de determinación respecto a la técnica IMNC haciéndola un poco más eficiente, en especial

cuando el mecanismo es MCAR y se tiene el 10% de faltantes; sin embargo, a nivel general tanto IMC como IMNC no producen buenos resultados en la estimación del coeficiente de determinación, más aún cuando el porcentaje de datos faltantes aumenta a un 30% o 60%; a pesar que con esta técnica las estimaciones se acercan un poco más al valor real comparados con los obtenidos con IMNC, los resultados presentan sesgos.

Se observó que con la técnica EM para porcentajes de faltantes pequeños (del 10%) y para todos los mecanismos de faltantes, las estimaciones del coeficiente de determinación presentan una distribución similar a la condición de datos completos; sin embargo, la mayor eficiencia relativa se presenta en las muestras que tienen un 10% de faltantes y bajo MCAR y MARy. Cuando el porcentaje de faltantes se incrementa a un 30%, las muestras de tamaño 50 y 100 presentan una variabilidad un poco mayor que la condición de datos completos, cuando el tamaño de muestra se incrementa a 200, las estimaciones son más cercanas a las estimaciones de muestras de datos completos especialmente para los mecanismos MCAR y MARy. En general, se observó que para las muestras de tamaño 200 estos dos mecanismos producen mejores estimaciones que cuando los faltantes siguen otras distribuciones. En general, se observa que a pesar que las estimaciones son influenciadas por los mayores niveles de datos faltantes y por mecanismos como el NI, esta técnica es superior a todas las demás en la estimación del coeficiente de determinación.

Efectos sobre la eficiencia de estimación de los Coeficientes de Regresión y sus intervalos de confianza

Los resultados de la eficiencia relativa de la estimación de los coeficientes de regresión se presentan en la tabla 6. Las tablas 7, 8 y 9 presentan los resultados de cubiertas y amplitudes de intervalos de confianza para los coeficientes β_0, β_1 y β_2 respectivamente.

Usando ACC, en general se observa que para los tres coeficientes de regresión, en las muestras de tamaños 50, 100 y 200 con el 10% de faltantes bajo un mecanismo MCAR o MAR (selección sobre X o sobre Y) la medida de eficiencia relativa osciló entre 0.8 y 1.0 con cubiertas de intervalos que superaron en todos los casos el criterio del 90% y con amplitudes cercanas a las halladas en la condición de datos completos (Tabla 5). Para ACC con el mecanismo No ignorable y todos los tamaños de muestra, se observó un comportamiento generalizado en los coeficientes de regresión: 1) en el caso del intercepto β_0 las estimaciones son sobreestimadas, 2) en el caso de β_1 y β_2 las estimaciones son subestimadas. Los sesgos aumentan a medida que aumentan el porcentaje de faltantes y el tamaño de muestra (Tablas 7, 8 y 9).

En cuanto a ACD, en general, se halló una pobre estimación de los coeficientes de regresión; salvo algunas excepciones, los estimados presentan baja eficiencia producto de sesgos positivos o negativos y una alta variabilidad que como es de esperarse es mucho mayor en las muestras pequeñas (Tabla 6). Respecto a las estimaciones del intercepto (β_0) se encontró que los intervalos de este coeficiente presentan cubiertas del verdadero valor superiores al 90% bajo los mecanismos MCAR y MAR para todos los tamaños y porcentajes de faltantes; el único caso donde la estimación bajo el mecanismo NI presentó una cubierta superior al 90% fue en el caso de las muestras de tamaño 50 y 10% de faltantes. Pese a tener cubiertas superiores, los únicos casos donde la eficiencia relativa de estas estimaciones fue superior a 0.8 fue en las muestras con el 10% de faltantes sujetos a un mecanismo MCAR y MARx. En los mecanismos NI y MARy se observaron subestimaciones de este coeficiente que se acentúan al aumentar el porcentaje de faltantes.

Con ACD para las estimaciones del coeficiente β_1 en el caso de las muestras de tamaño 50, todas las estimaciones por intervalos tuvieron una cubierta superior al 90%, de estas estimaciones solo aquellas correspondientes a porcentajes de faltantes del 10% bajo mecanismos MCAR, MAR (selección sobre X ó Y) tuvieron una eficiencia relativa entre 0.8 y 0.9, el resto de casos presenta ER inferiores a 0.7, mostrando intervalos que se van ampliando a medida que aumenta el porcentaje de faltantes. Cuando el tamaño de muestra aumenta a 100 y 200 las estimaciones pierden eficiencia y la cubierta de los intervalos disminuye para algunos casos

cuando falta el 10% y 30% de información con un mecanismo NI y para porcentajes del 30% y 60% con mecanismos MCAR y MARy. Para las estimaciones del coeficiente β2 en todos los casos la eficiencia relativa no superó el valor de 0.8 como consecuencia de la sobreestimación del coeficiente y la cubierta de los intervalos fue seriamente afectada; solo bajo los mecanismos MCAR y MARy en el caso de muestras de tamaño 50 y 100 con el 10% de faltantes se obtuvieron cubiertas superiores al 90% (Tablas 7, 8 y 9).

En cuanto a IMNC, los únicos casos donde presenta eficiencias relativas superiores a 0.8 para el intercepto fueron para aquellas muestras con el 10% de faltantes bajo los mecanismos MCAR y MARy. Respecto a los intervalos de confianza generados, se observa que para todos los porcentajes de faltantes bajo esos mecanismos la cubierta de los intervalos es superior al 90%. El mecanismo MARx no presenta estimaciones seguras del coeficiente β_0 para el 30% y el 60% de información faltante y en todos estos casos subestima el verdadero valor poblacional y genera intervalos con una cubierta inferior al 90%. En general, se observó que la técnica de imputación de la media no condicional presentó estimados eficientes del Coeficiente de regresión de β_1 solamente cuando se tiene un 10% de información faltante de X_1 y bajo los mecanismos MCAR y MARy. En estas condiciones, los intervalos de confianza resultantes tuvieron una cubierta superior al 90%.

Para el mecanismo NI y todos los porcentajes de faltantes, IMNC produce intervalos con una cubierta inferior al 90% que subestiman el verdadero valor del parámetro; en las muestras de tamaño 200 el impacto sobre la cubierta del intervalo del coeficiente β_1 es muy severo, llegando a valores del 24% para el 60% de faltantes. En general, para el Coeficiente de regresión β_2 se observó que este procedimiento no es eficiente en su estimación y sobreestima el verdadero valor del mismo, generando a la vez intervalos de confianza con una baja cubierta. Las únicas muestras donde IMNC produjo intervalos de confianza con una cubierta superior al 90% fueron aquellas que presentaban un 10% de faltantes bajo MCAR o MARy y de tamaño 50 y 100.

Si se compara IMC con IMNC, la primera técnica presenta un comportamiento similar en la estimación de los coeficientes de regresión, aunque las estimaciones son un poco más eficientes y las amplitudes de los intervalos de confianza un poco más pequeñas. Para los tres coeficientes de regresión, se obtuvieron estimaciones eficientes cuando las muestras tuvieron el 10% de faltantes bajo MCAR y MARy, con cubiertas de intervalos de confianza superiores al 90% (Tabla 5). Con IMC, se observa un deterioro en las cubiertas de algunos intervalos: 1) Para β_0, disminuye la cubierta de aquellas muestras con el 30% y 60% de información faltante bajo NI y MARx. 2) Para β_1, disminuye la cubierta de aquellas muestras de tamaño 200 con el 30% y 60% de faltantes bajo un mecanismo NI. Igual sucede con las muestras de tamaño 100 y el 60% de faltantes. 3) Para β_2, disminuye la cubierta de las muestras de tamaño 200 con el 30% y 60% de faltantes bajo NI y MARx. Para las muestras de tamaño 50 y faltantes del 30% y 60% bajo NI (Tablas 6, 7 y 8).

Para los tres coeficientes de la regresión, se observó que entre todas las técnicas, la única que preserva la amplitud de los intervalos entre la condición de muestras completas y muestras con datos faltantes es el algoritmo EM. También se observó que es la técnica menos sensible en la estimación de los tres coeficientes de regresión a la ocurrencia de valores faltantes sobre X_1 que siguen un mecanismo NI y con porcentajes del 10% y el 30% de información faltante. A pesar que este mecanismo a un 60% de faltantes disminuye la cubierta de los intervalos de confianza, esta disminución aunque severa no es tan extrema como las otras técnicas de tratamiento de faltantes.

En el caso de las muestras con el 10% de datos faltantes, en general el algoritmo EM presenta cubiertas del verdadero valor del intercepto superiores al 90%. Esta técnica presenta estimaciones eficientes del coeficiente de regresión de β_1 para los tres tamaños de muestra estudiados y todos los mecanismos de datos faltantes solamente cuando falta el 10% de datos sobre X_1; de igual forma produce en estos casos intervalos de confianza del coeficiente β_1 con una cubierta superior al 90% y con amplitudes que se

incrementan poco respecto a la condición de datos completos. El algoritmo EM presenta estimaciones eficientes del coeficiente de regresión β_2 para los tres tamaños de muestra estudiados y todos los mecanismos de datos faltantes solamente cuando falta el 10% de datos sobre X_1. En el caso de las muestras de tamaño 100 y 200, EM produce estimaciones eficientes de β_2 para los mecanismos MCAR y MARy bajo un 10% de datos faltantes sobre X_1.

5. Conclusiones y futuras investigaciones

La ocurrencia de valores faltantes en los conjuntos de datos es un problema que debe ser direccionado desde una perspectiva más amplia que abarque no solo la aplicación de cualquier técnica de datos faltantes sino también el análisis del impacto que tal técnica produce sobre la eficiencia y la honestidad de los estimados de los coeficientes de determinación y de regresión, aún para pequeños porcentajes de datos faltantes.

Efectuar Análisis de casos completos aunque es una alternativa fácil de implementar y disponible en todos los paquetes de software estadístico debe ser usada solamente cuando el porcentaje de faltantes es pequeño, ya que cómo se encontró en este estudio, a medida que aumenta el porcentaje de valores faltantes las estimaciones de los coeficientes de determinación y de regresión pierden eficiencia y en muchas ocasiones se altera la cubierta y la amplitud de los intervalos de confianza de los coeficientes de regresión.

A pesar que las técnicas de Análisis de casos disponibles, Imputación de la media no condicional e Imputación de la media condicional intentan evitar el descartar información de los casos con algún valor faltante, no son recomendables para el tratamiento de estos valores dado que producen en muchos casos estimaciones ineficientes de los coeficientes de determinación y de regresión más aún cuando el porcentaje de faltantes es alto.

En cuanto a las técnicas de imputación de la media no condicional y condicional, se halló que no presentan en general estimaciones eficientes de los coeficientes de determinación y de regresión, aunque se verificó que la eficiencia de las estimaciones a través de la imputación de la media no condicional es inferior a la hallada a través de la imputación de la media condicional.

Se verificó de acuerdo a la teoría estadística, que la estimación máximo verosímil a través del algoritmo EM es una técnica en general eficiente y menos sensible a mecanismos que no son MCAR; tanto para la estimación del coeficiente de determinación como para los coeficientes de regresión, el algoritmo EM fue en general más eficiente que el resto de técnicas bajo el 10% de datos faltantes y mecanismos MCAR y MAR.

Debe tenerse en cuenta que a pesar que la estimación máximo verosímil a través del algoritmo EM es mucho más atractiva que la técnicas de ACC e imputación simple, está basada sobre el supuesto que la muestra es lo suficientemente grande para que las estimaciones sean aproximadamente insesgadas y normalmente distribuidas, cuando los datos no son normales los estimados no son necesariamente eficientes. Sin embargo, en este estudio, el muestreo fue obtenido de una población bajo una distribución normal, lo cual asegura la normalidad de las estimaciones de la muestra.

En investigaciones aplicadas, la principal recomendación para el tratamiento de los valores faltantes es diseñar mecanismos adecuados que permitan minimizar la falta de respuesta al momento de recolectar los datos y no esperar al momento de la edición de la información para determinar qué hacer con los valores faltantes tanto de ausencia de respuesta como de errores en la recolección de información. Una vez que han ocurrido estos valores faltantes, es necesario antes de proceder a aplicar alguna técnica, hacer un análisis exploratorio de las variables sujetas a faltantes y de las características de los individuos que presentan tales valores, con el fin de descartar patrones en los datos que sugieran comportamientos no aleatorios. Se debe recordar, que la aplicación de las técnicas como Análisis de casos disponibles e imputación de la media no condicional y condicional se basan en general sobre un mecanismo MCAR y presentan estimaciones que no son eficientes cuando la distribución de los faltantes se aleja de este mecanismo.

Debido al auge de diversas herramientas computacionales tanto de distribución gratuita como no gratuita, incluidas en paquetes estándar

como SPLUS, SPSS, SAS, así como paquetes como NORM y SOLAS especializados en el tema de datos faltantes, la estimación máximo verosímil a través del algoritmo EM se torna más accesible al investigador aplicado, quien puede de esta forma obtener estimaciones válidas y eficientes. El software R utilizado en este estudio para simular los mecanismos de datos faltantes y el experimento en general, es una herramienta potente y fácil de manipular y programar para hacer estudios de este tipo; además, R cuenta con diversos paquetes como NORM que son especializados en el tratamiento de datos faltantes.

Entre las futuras investigaciones, se hace necesario abordar temas como los siguientes:

a) La imputación múltiple basada en inferencia bayesiana y cadenas de Markov de Monte Carlo, la cual según Rubin (1988), mantiene las principales ventajas de la imputación simple y rectifica su principal desventaja de no tener en cuenta la incertidumbre de las predicciones realizadas cuando se han imputado los valores faltantes.

b) Según la literatura estadística, la eficiencia de la imputación de la media condicional y no condicional puede ser mejorada si se adiciona a las estimaciones un error aleatorio proveniente de alguna distribución, en cuyo caso se tendría la versión estocástica de estas técnicas.

c) La convergencia del algoritmo EM en diversas situaciones.

d) El efecto de los outliers o medidas de influencia en el análisis de regresión para diversas técnicas de tratamiento de datos faltantes.

e) El efecto de diversos niveles de colinealidad sobre la eficiencia de las estimaciones para diversas técnicas de tratamiento de datos faltantes.

6. Referencias bibliográficas

1. AFIFI, A.A., ELASHOFF, R.M. (1966) Missing observations in multivariate statistics. I. Review of the literature. Journal of the American Statistical Association JASA, Vol. 61, 595 – 604.
2. ALLISON, Paul. Multiple imputation for missing data : A Cautionary tale. (http://www.ssc.upenn.edu/~allison/MultInt99.pdf)
3. BAUTISTA, Leonardo. (2000) Técnica de Diseño de Encuestas. Memorias Simposio de Estadística Universidad Nacional de Colombia
4. BEHAR G, Roberto. (2002) Validación de Supuestos en el modelo de regresión. Universidad del Valle
5. BERNARDO, José M. Bayesian Statistics. Página Web http://matheron.uv.es/pub/personal/bernardo/BayesStat2.pdf
6. BROCKMEIER, Lantry L., KROMREY, Jeffre D., HINES, Constance V. (1998) Systematically Missing Data and Multiple Regression Analysis: An Empirical Comparison of Deletion and Imputation Techniques. Multiple Linear Regression Viewpoints, Vol. 25.
7. Correa M, Juan Carlos y González, Nelfi. Introducción al R. Posgrado en Estadística. Universidad Nacional-Sede Medellín. 2002
8. DIAZ M, Luis G. Estadística Multivariada: Inferencia y Métodos. Notas de clase. Universidad Nacional de Colombia. Departamento de Matemáticas y Estadística. Santafe de Bogotá
9. DRAPER, Norman R., SMITH, Harry. (1998) Applied Regression Analysis. John Wiley. United States of America
10. FICHMAN, M., CUMMINGS, J. Multiple imputation for missing data: Making the most of what you know. (http://www.gsia.cmu.edu/andrew/mf4f/work/misspres.pdf)
11. FRONGILLO, E. (2002) Office Of Statistical Consulting. StatNews #50: What is Maximum Likelihood. Universidad de Cornell. http://www.he.cornell.edu/admin/statcons/statnews/statnews50.pdf
12. GLASSER, M. (1964) Linear regressions analysis with missing observations among the independent variables. Journal of the American Statistical Association JASA, Vol. 59, 834 – 844.
13. GRACE, Kelly. (2001) StatNews # 47 Limitations of common solutions to missing data. Office Of Statistical Consulting

(http://www.human.cornell.edu/admin/statcons/Statnews/stnews47.pdf)
14. GRACE, Kelly. (2001) StatNews #46: Missing Data Mechanisms. Office Of Statistical Consulting. Universidad de Cornell. http://www.human.cornell.edu/admin/statcons/Statnews/stnews46.pdf
15. GRAYBILL, Franklin A. (1976) Theory and Application of the linear models. North Scituate – Massachusetts. Duxbury Press
16. HORTON, Nicholas J., LIPSITZ, Stuart R. (2001) Múltiple imputation in practice: Comparison of software packages for regression models with missing variables. The American Statistician, Vol. 55, 244 – 254.
17. INSIGHTFUL CORPORATION. (2001) Analyzing Data with Missing Values in S-PLUS (doc pdf) Seattle, Washington
18. LAAKSONEN, Seppo (1999) How to Find the Best Imputation Technique? Tests with three methods. Resumen para "International Conference on Nonresponse 1999", Portland, Oregon, November 28, 1999.
19. LITTLE, Roderick J.A. (1992) Regression with missing X's: A review. Journal of the American Statistical Association JASA, Vol. 87, No.420, 1227 – 1237.
20. LITTLE, Roderick J.A., RUBIN, Donald B. (1987) Statistical analysis with missing data, New York: John Wiley.
21. LOHR, Sharon (2000) Muestreo: Diseño y Análisis.
22. LÓPEZ, Luis Alberto. (1995) Comparación de métodos para estimar observaciones faltantes en modelos de clasificación y regresión. Memorias Simposio internacional de estadística en agricultura y medio ambiente. Departamento de Matemáticas y Estadística – Universidad Nacional de Colombia
23. MANLY, Bryan F.J. (1997) Randomization, Bootstrap and Monte Carlo Methods in Biology. Texts in Statistical Science, 2a edición. Chapman & Hall, Londres.
24. MENDEZ, C.E. (1995) Metodología Guía para elaborar diseños de investigación en ciencias económicas, contables y administrativas. 2ª edición. Mc GrawHill. Colombia.
25. MOOB, Alexander M., GRAYBILL, Franklin A., BOES, Duane C (1974) Introduction to the Theory of Statistics. 3a edición. McGraw Hill. Estados Unidos
26. PEREZ, Adriana. (2000) Tratamiento de datos faltantes en encuestas. Memorias Simposio de Estadística Universidad Nacional de Colombia.
27. PEREZ, Jaime. (2003) Notas de Clase, Modelos Lineales. Universidad del Valle, Cali.
28. PUERTA Goicochea, Aitor. (2002) Imputación basada en árboles de clasificación. Eustat. Página Web : (http://www.eustat.es/spanish/general/info/CTAitor.pdf)
29. Página Web del paquete estadístico R: http : //cran.r _ project.org
30. RIAL, Antonio., VARELA, Jesús., ROJAS, Antonio J. (2001) Depuración y Análisis Preliminares de Datos en SPSS. Edit. Ra-ma. Madrid.
31. RUBIN, Donald B. (1976) Inference and missing data. Biometrika, Vol. 63, 581 – 92.
32. RUBIN, Donald B. (1988) An Overview Of Multiple Imputation. Harvard University. One Oxford Street, Cambridge. Documento pdf.
33. RUBIN, Donald B. (1996) Multiply Imputation after 18+ years. Journal of the American Statistical Association JASA, Vol. 91, 473 – 489.
34. RUBIN, Donald B., SCHAFER, Joseph L. (1998) Multiple imputation for missing data problems. (http://www.stat.psu.edu/~jls/aug98.pdf)
35. RUBINSTEIN, Reuven Y. (1981) Simulation and the Monte Carlo Method. Wiley series in probability and mathematical statistics, Estados Unidos.
36. SAURABH . Methodologies for dealing missing data.
37. SCHAFER, Joseph L. (1997) Analysis of Incomplete Multivariate Data. Series Monographs on Statistics and Applied Probability. Chapman & Hall, Londres.
38. SCHAFER, Joseph L. (1998) The practice of multiple imputation (http://www.stat.psu.edu/~jls/hhdev.pdf)
39. SCHAFER, Joseph L. The multiple imputation FAQ page Joe Schafer (http://www.stat.psu.edu/~jls/mifaq.html#cc)

40. SCHAFER, Joseph L. http://www.stat.psu.edu/~jls/jls.html#res
41. SCHAFER, Joseph L. Imputation Procedures for missing data. Penn State University. (http://www.stat.psu.edu/~jls/session1.pdf)
42. SCHAFER, Joseph L. (1999) Imputation Procedures for missing data. Penn State University. (http://www.stat.psu.edu/~jls/session2.pdf)
43. SCHAFER, Joseph L., GRAHAM, John W. (2002) Missing data: Our view of the state of art. Pennsylvania State University. Psychological Methods. Vol. 7, No.2, 147–17 Página Web: (http://www.nyu.edu/classes/shrout/G89-2229/Schafer&Graham2002.pdf)
44. SCHAFER, Joseph L., OLSEN, M. (1998) Multiple imputation for multivariate missing data problems: a data analysis perspective. Universidad del estado de Pennsylvania. (http://www.stat.psu.edu/~jls/mbr.pdf)
45. SCHAFER, Joseph L., SCHENKER, N. (1997) Inference with Imputed Conditional Means (http://www.stat.psu.edu/~jls/impvar.pdf)
46. SMITH, M. (2002) What Can I Do about Missing Data? (http://www.herc.research.med.va.gov/FAQ_19.htm)
47. WERNER WOTHKE, SMALLWATERS CORP. Longitudinal and multi-group modeling with missing data (documento pdf en CD-R)
48. YUAN, Y.C. SAS Institute Inc., Rockville, MD. Multiple Imputation for Missing Data: Concepts and New Development (documento pdf en CD-R)

7. Apéndice

Tabla 5: Eficiencia Relativa del estimado del Coeficiente de Determinación

Tratamiento			ACC			ACD			IMNC			IMC			EM		
n	p %	Mec.	ECM Comp	ECM Falt	ER	ECM Comp	ECM Falt	ER	ECM Comp	ECM Falt	ER	ECM Comp	ECM Falt	ER	ECM Comp	ECM Falt	ER
50	10	MCAR	0,004	0,005	**0,88**	0,005	0,006	**0,85**	0,004	0,005	**0,85**	0,004	0,005	**0,86**	0,005	0,005	**0,95**
		MARx	0,004	0,006	0,78	0,004	0,007	0,61	0,004	0,007	0,59	0,004	0,007	0,59	0,004	0,005	**0,83**
		MARy	0,005	0,006	**0,90**	0,004	0,005	0,80	0,004	0,006	0,74	0,005	0,005	**0,91**	0,005	0,006	**0,95**
		NI	0,005	0,008	0,58	0,004	0,009	0,48	0,005	0,012	0,41	0,004	0,008	0,43	0,004	0,005	**0,84**
	30	MCAR	0,004	0,007	0,63	0,004	0,011	0,34	0,004	0,011	0,42	0,005	0,011	0,47	0,004	0,005	0,76
		MARx	0,004	0,007	0,58	0,006	0,009	0,62	0,004	0,016	0,26	0,005	0,014	0,33	0,003	0,006	0,63
		MARy	0,004	0,006	0,65	0,005	0,010	0,47	0,004	0,009	0,51	0,004	0,008	0,50	0,004	0,006	0,73
		NI	0,004	0,015	0,27	0,005	0,013	0,39	0,005	0,017	0,27	0,005	0,014	0,32	0,004	0,006	0,69
	60	MCAR	0,005	0,010	0,47	0,005	0,025	0,21	0,004	0,026	0,14	0,005	0,024	0,20	0,005	0,009	0,48
		MARx	0,005	0,009	0,48	0,004	0,016	0,25	0,005	0,027	0,19	0,004	0,023	0,17	0,005	0,009	0,57
		MARy	0,005	0,011	0,42	0,004	0,019	0,19	0,004	0,022	0,17	0,005	0,018	0,28	0,005	0,008	0,59
		NI	0,005	0,024	0,19	0,004	0,013	0,32	0,004	0,023	0,16	0,004	0,021	0,21	0,004	0,008	0,54
100	10	MCAR	0,002	0,002	**0,88**	0,002	0,003	**0,81**	0,002	0,003	0,59	0,002	0,003	0,72	0,002	0,002	**0,89**
		MARx	0,002	0,003	0,75	0,002	0,005	0,38	0,002	0,005	0,40	0,002	0,003	0,65	0,002	0,002	**0,88**
		MARy	0,002	0,002	**0,89**	0,002	0,003	0,67	0,002	0,003	0,71	0,002	0,002	**0,85**	0,002	0,002	**0,91**
		NI	0,002	0,005	0,41	0,002	0,007	0,27	0,002	0,008	0,22	0,002	0,006	0,38	0,002	0,003	0,77
	30	MCAR	0,002	0,003	0,60	0,002	0,006	0,28	0,002	0,007	0,23	0,002	0,007	0,29	0,002	0,002	0,67
		MARx	0,002	0,004	0,50	0,002	0,006	0,33	0,002	0,011	0,16	0,002	0,010	0,20	0,002	0,003	0,66
		MARy	0,002	0,003	0,75	0,002	0,008	0,25	0,002	0,006	0,29	0,002	0,006	0,34	0,002	0,003	0,71
		NI	0,002	0,011	0,16	0,002	0,009	0,19	0,002	0,019	0,13	0,002	0,011	0,19	0,002	0,003	0,64
	60	MCAR	0,002	0,005	0,36	0,002	0,017	0,10	0,002	0,020	0,09	0,002	0,018	0,10	0,002	0,004	0,49
		MARx	0,002	0,008	0,21	0,002	0,013	0,17	0,002	0,025	0,09	0,002	0,021	0,09	0,002	0,005	0,38
		MARy	0,002	0,005	0,39	0,002	0,020	0,11	0,002	0,019	0,09	0,002	0,017	0,11	0,002	0,004	0,52
		NI	0,002	0,021	0,11	0,002	0,011	0,21	0,002	0,025	0,07	0,002	0,019	0,10	0,002	0,003	0,56
200	10	MCAR	0,001	0,001	**0,82**	0,001	0,001	0,59	0,001	0,002	0,56	0,001	0,002	0,70	0,001	0,001	**0,91**
		MARx	0,001	0,001	0,63	0,001	0,003	0,27	0,001	0,004	0,18	0,001	0,003	0,31	0,001	0,001	**0,84**
		MARy	0,001	0,001	**0,91**	0,001	0,001	0,52	0,001	0,001	0,57	0,001	0,002	0,63	0,001	0,001	**0,88**
		NI	0,001	0,004	0,23	0,001	0,005	0,17	0,001	0,008	0,11	0,001	0,003	0,22	0,001	0,001	0,79
	30	MCAR	0,001	0,001	0,66	0,001	0,007	0,14	0,001	0,006	0,14	0,001	0,005	0,23	0,002	0,002	0,67
		MARx	0,001	0,003	0,35	0,001	0,006	0,15	0,001	0,011	0,08	0,001	0,007	0,11	0,001	0,002	0,63
		MARy	0,001	0,002	0,61	0,001	0,006	0,14	0,001	0,006	0,16	0,001	0,004	0,18	0,001	0,001	0,77
		NI	0,001	0,010	0,08	0,001	0,008	0,10	0,001	0,015	0,06	0,001	0,009	0,10	0,001	0,001	0,70
	60	MCAR	0,001	0,003	0,31	0,001	0,019	0,05	0,001	0,018	0,05	0,001	0,016	0,06	0,001	0,002	0,49
		MARx	0,001	0,004	0,18	0,001	0,010	0,10	0,001	0,022	0,04	0,001	0,019	0,05	0,001	0,002	0,37
		MARy	0,001	0,003	0,35	0,001	0,018	0,05	0,001	0,018	0,05	0,001	0,015	0,06	0,001	0,002	0,40
		NI	0,001	0,020	0,05	0,001	0,010	0,10	0,001	0,023	0,03	0,001	0,017	0,05	0,001	0,001	0,63

Nota.
ECM. Error Cuadrático Medio
ER. Eficiencia Relativa
Parámetro de referencia. $R^2 = 0,742$

Tabla 6: Eficiencia Relativa de la estimación de los coeficientes de regresión

Tratamiento			ACC			ACD			IMNC			IMC			EM		
n	p%	Mec.	β_0	β_1	β_2	β0	β1	β2	β0	β1	β2	β0	β1	β2	β0	β1	β2
50	10	MCAR	0,86	0,89	0,80	**0,90**	0,87	0,76	0,89	0,88	0,79	**1,00**	**0,91**	**0,88**	**0,99**	**0,91**	**0,84**
		MARx	**0,96**	0,84	**0,94**	0,57	0,88	0,64	0,59	0,76	0,53	0,66	0,73	0,74	0,78	0,83	**0,92**
		MARy	0,88	0,85	**0,95**	**0,96**	0,85	0,80	**0,94**	0,87	0,81	**0,98**	0,82	0,86	**0,98**	0,87	**0,94**
		NI	0,63	0,76	**0,98**	0,67	0,64	0,57	0,65	0,53	0,51	0,85	0,79	0,78	0,90	0,88	**0,91**
	30	MCAR	0,67	0,62	0,69	0,77	0,60	0,39	0,62	0,60	0,45	0,85	0,63	0,76	0,82	0,81	0,76
		MARx	0,54	0,48	0,62	0,29	0,63	0,40	0,18	0,51	0,28	0,21	0,64	0,71	0,26	0,64	0,70
		MARy	0,68	0,70	0,64	0,80	0,59	0,42	0,74	0,63	0,40	0,68	0,61	0,69	0,78	0,68	0,80
		NI	0,28	0,61	0,66	0,40	0,47	0,29	0,29	0,25	0,19	0,45	0,58	0,49	0,75	0,72	0,66
	60	MCAR	0,35	0,40	0,41	0,64	0,29	0,21	0,47	0,27	0,22	0,35	0,30	0,40	0,41	0,35	0,56
		MARx	0,21	0,33	0,37	0,23	0,50	0,32	0,06	0,24	0,17	0,06	0,32	0,45	0,11	0,34	0,48
		MARy	0,39	0,43	0,35	0,60	0,28	0,19	0,35	0,29	0,23	0,43	0,39	0,39	0,60	0,44	0,44
		NI	0,11	0,38	0,48	0,24	0,44	0,35	0,16	0,19	0,17	0,30	0,38	0,38	0,51	0,60	0,62
100	10	MCAR	**1,01**	0,89	0,87	**0,98**	0,87	0,58	**0,98**	0,75	0,69	**0,93**	0,87	**0,91**	**0,92**	0,89	**0,91**
		MARx	0,86	0,79	0,89	0,43	0,74	0,53	0,36	0,67	0,38	0,50	0,75	0,66	0,62	**0,91**	0,79
		MARy	0,85	0,87	**0,92**	**0,94**	0,81	0,74	**0,91**	0,81	0,64	**0,95**	0,86	0,84	0,89	**0,94**	**0,99**
		NI	0,51	0,75	0,82	0,47	0,45	0,29	0,53	0,24	0,21	0,68	0,59	0,55	**0,93**	0,89	0,79
	30	MCAR	0,66	0,65	0,72	0,80	0,49	0,26	0,66	0,48	0,25	0,69	0,64	0,76	0,77	0,73	0,75
		MARx	0,66	0,63	0,62	0,15	0,59	0,29	0,10	0,40	0,18	0,09	0,56	0,45	0,19	0,52	0,72
		MARy	0,66	0,82	0,67	0,79	0,44	0,27	0,65	0,56	0,29	0,80	0,77	0,70	0,86	0,87	0,73
		NI	0,17	0,44	0,52	0,26	0,37	0,18	0,17	0,14	0,11	0,30	0,43	0,30	0,74	0,74	0,70
	60	MCAR	0,39	0,35	0,31	0,60	0,26	0,10	0,48	0,22	0,11	0,36	0,37	0,43	0,54	0,55	0,53
		MARx	0,24	0,32	0,32	0,13	0,49	0,22	0,03	0,18	0,07	0,03	0,34	0,40	0,05	0,37	0,54
		MARy	0,37	0,38	0,38	0,66	0,19	0,13	0,41	0,21	0,10	0,48	0,39	0,48	0,51	0,54	0,43
		NI	0,04	0,31	0,30	0,15	0,48	0,18	0,11	0,09	0,09	0,14	0,27	0,19	0,43	0,65	0,52
200	10	MCAR	0,82	**0,91**	**0,90**	0,85	0,71	0,49	0,80	0,80	0,51	**0,92**	0,87	0,88	**0,92**	**0,90**	**0,97**
		MARx	0,89	0,79	0,85	0,25	0,69	0,33	0,20	0,50	0,23	0,25	0,78	0,64	0,46	0,75	0,76
		MARy	0,79	0,85	0,89	**0,91**	0,60	0,51	**0,98**	0,81	0,57	**0,99**	0,87	**1,00**	**0,93**	**0,92**	**0,91**
		NI	0,30	0,55	0,58	0,33	0,24	0,16	0,26	0,14	0,11	0,44	0,50	0,38	**0,91**	0,87	0,72
	30	MCAR	0,74	0,69	0,66	0,63	0,35	0,18	0,66	0,40	0,13	0,69	0,53	0,63	0,77	0,73	0,75
		MARx	0,61	0,60	0,72	0,07	0,55	0,15	0,05	0,26	0,08	0,06	0,54	0,35	0,09	0,60	0,70
		MARy	0,66	0,60	0,63	0,67	0,39	0,17	0,71	0,43	0,20	0,79	0,62	0,62	0,80	0,85	0,77
		NI	0,06	0,26	0,35	0,11	0,30	0,09	0,13	0,07	0,05	0,13	0,29	0,16	0,80	0,79	0,69
	60	MCAR	0,33	0,35	0,31	0,53	0,14	0,05	0,44	0,11	0,05	0,38	0,36	0,38	0,53	0,47	0,47
		MARx	0,17	0,29	0,33	0,07	0,42	0,09	0,01	0,15	0,05	0,01	0,32	0,26	0,02	0,28	0,46
		MARy	0,36	0,37	0,29	0,50	0,17	0,05	0,41	0,15	0,05	0,40	0,38	0,36	0,59	0,43	0,52
		NI	0,02	0,19	0,17	0,08	0,44	0,13	0,05	0,05	0,04	0,08	0,13	0,10	0,56	0,73	0,58

Nota.

Mecanismos. MCAR: Faltantes Completamente al azar, MARx: Faltantes al azar dependientes de X2, MARy: Faltantes al azar dependientes de Y, NI: Faltantes No Ignorables.

Técnicas. ACC: Análisis de Casos Completos,ACD: Análisis de Casos Disponibles, IMNC: Imputación de la media no condicional, IMC: Imputación de la media condicional, EM: Algoritmo EM.

Valores en negrita del ER corresponden a eficiencias superiores al 90%

Tabla 7. Cubierta y amplitud de intervalos de confianza para el Intercepto (β_0)

Tratamiento				Técnicas de tratamiento									
n	p %	Mec.	Ind. muestra	ACC		ACD		IMNC		IMC		EM	
				Cub.	Ancho	Cub.	Ancho	Cub.	Ancho	Cub.	Ancho	Cub.	Ancho
50	10	MCAR	Completa	0,95	0,59	0,96	0,59	0,94	0,57	0,94	0,58	0,96	0,59
			Faltante	0,96	0,63	0,94	0,61	0,92	0,61	0,94	0,60	0,96	0,59
		MARx	Completa	0,97	0,59	0,98	0,59	0,97	0,59	0,95	0,59	0,97	0,59
			Faltante	0,96	0,63	0,93	0,64	0,93	0,66	0,94	0,64	0,93	0,59
		MARy	Completa	0,95	0,58	0,96	0,59	0,98	0,59	0,96	0,59	0,96	0,58
			Faltante	0,95	0,61	0,96	0,62	0,98	0,62	0,96	0,61	0,95	0,58
		NI	Completa	0,97	0,59	0,94	0,59	0,97	0,59	0,96	0,58	0,99	0,59
			Faltante	0,93	0,61	0,93	0,67	0,94	0,68	0,96	0,64	0,98	0,59
	30	MCAR	Completa	0,96	0,58	0,92	0,59	0,95	0,59	0,94	0,59	0,95	0,59
			Faltante	0,96	0,70	0,94	0,66	0,95	0,67	0,95	0,66	0,95	0,60
		MARx	Completa	0,95	0,59	0,94	0,58	0,95	0,59	0,95	0,59	0,95	0,59
			Faltante	0,96	0,76	**0,78**	0,70	**0,75**	0,81	**0,78**	0,75	**0,71**	0,65
		MARy	Completa	0,96	0,59	0,98	0,59	0,96	0,58	0,96	0,58	0,95	0,59
			Faltante	0,94	0,71	0,98	0,66	0,96	0,66	0,94	0,66	**0,89**	0,60
		NI	Completa	0,97	0,59	0,95	0,59	0,95	0,58	0,96	0,59	0,93	0,59
			Faltante	**0,70**	0,69	**0,88**	0,71	**0,86**	0,77	0,95	0,71	0,92	0,61
	60	MCAR	Completa	0,95	0,58	0,96	0,59	0,97	0,59	0,95	0,59	0,96	0,59
			Faltante	0,96	0,99	0,95	0,72	0,95	0,77	0,90	0,77	**0,78**	0,57
		MARx	Completa	0,96	0,59	0,96	0,59	0,93	0,58	0,94	0,59	0,96	0,59
			Faltante	0,95	1,21	**0,76**	0,75	**0,61**	1,25	**0,61**	1,16	**0,45**	0,68
		MARy	Completa	0,96	0,59	0,98	0,59	0,97	0,58	0,96	0,58	0,93	0,59
			Faltante	0,94	0,97	0,98	0,72	0,92	0,76	0,91	0,74	**0,84**	0,57
		NI	Completa	0,95	0,59	0,97	0,59	0,99	0,59	0,94	0,59	0,97	0,58
			Faltante	**0,50**	0,84	**0,82**	0,73	**0,85**	0,94	**0,84**	0,82	**0,85**	0,60
100	10	MCAR	Completa	0,95	0,41	0,96	0,41	0,97	0,41	0,97	0,41	0,97	0,41
			Faltante	0,96	0,43	0,97	0,42	0,96	0,43	0,97	0,43	0,95	0,41
		MARx	Completa	0,97	0,40	0,96	0,41	0,96	0,41	0,98	0,40	0,96	0,41
			Faltante	0,96	0,43	**0,87**	0,45	**0,86**	0,45	0,91	0,44	0,93	0,42
		MARy	Completa	0,99	0,41	0,95	0,41	0,97	0,41	0,97	0,41	0,98	0,41
			Faltante	0,98	0,43	0,96	0,43	0,98	0,43	0,94	0,42	0,97	0,41
		NI	Completa	0,95	0,41	0,96	0,40	0,93	0,41	0,93	0,40	0,94	0,41
			Faltante	**0,89**	0,43	0,91	0,46	**0,88**	0,48	0,91	0,45	0,94	0,42
	30	MCAR	Completa	0,98	0,41	0,95	0,40	0,99	0,41	0,96	0,41	0,93	0,41
			Faltante	0,97	0,49	0,93	0,45	0,95	0,46	0,94	0,45	0,91	0,41
		MARx	Completa	0,94	0,41	0,95	0,40	0,96	0,41	0,98	0,40	0,96	0,28
			Faltante	0,96	0,53	**0,59**	0,49	**0,45**	0,55	**0,42**	0,52	0,94	0,29
		MARy	Completa	0,95	0,41	0,99	0,41	0,98	0,40	0,94	0,41	0,93	0,40
			Faltante	0,96	0,50	0,99	0,46	0,96	0,46	0,94	0,45	0,92	0,40
		NI	Completa	0,96	0,41	0,98	0,41	0,98	0,41	0,96	0,41	0,97	0,41
			Faltante	**0,56**	0,47	**0,76**	0,49	**0,78**	0,54	**0,84**	0,49	0,94	0,42
	60	MCAR	Completa	0,96	0,40	0,99	0,41	0,95	0,41	0,95	0,40	0,97	0,41
			Faltante	0,97	0,66	0,99	0,50	0,91	0,51	**0,88**	0,50	**0,89**	0,40
		MARx	Completa	0,96	0,41	0,97	0,40	0,98	0,41	0,94	0,41	0,96	0,41
			Faltante	0,95	0,86	**0,62**	0,53	**0,29**	0,89	**0,17**	0,80	**0,19**	0,48
		MARy	Completa	0,97	0,41	0,98	0,41	0,95	0,41	0,96	0,41	0,97	0,41
			Faltante	0,95	0,66	0,97	0,50	0,93	0,51	0,94	0,51	**0,88**	0,40
		NI	Completa	0,96	0,41	0,95	0,41	0,96	0,40	0,97	0,41	0,96	0,40
			Faltante	**0,16**	0,58	**0,62**	0,51	**0,67**	0,64	**0,70**	0,56	**0,82**	0,41

Tabla 7. (Continuación)

Tratamiento				Técnicas de tratamiento									
n	p %	Mec.	Ind. muestra	ACC		ACD		IMNC		IMC		EM	
				Cub.	Ancho	Cub.	Ancho	Cub.	Ancho	Cub.	Ancho	Cub.	Ancho
200	10	MCAR	Completa	0,99	0,28	0,95	0,28	0,98	0,29	0,98	0,29	0,98	0,29
			Faltante	1,00	0,30	0,96	0,30	0,98	0,30	0,97	0,30	0,97	0,29
		MARx	Completa	0,97	0,28	0,98	0,29	0,97	0,29	0,97	0,29	0,97	0,29
			Faltante	0,98	0,30	**0,75**	0,32	**0,71**	0,32	**0,78**	0,31	**0,83**	0,29
		MARy	Completa	0,99	0,29	0,96	0,29	0,98	0,29	0,98	0,29	0,98	0,29
			Faltante	0,97	0,30	0,96	0,30	0,98	0,30	0,98	0,30	0,96	0,29
		NI	Completa	0,94	0,29	0,98	0,28	1,00	0,29	0,97	0,28	0,98	0,28
			Faltante	**0,69**	0,30	**0,88**	0,33	**0,84**	0,33	0,92	0,32	0,97	0,29
	30	MCAR	Completa	0,97	0,28	0,96	0,28	1,00	0,28	0,99	0,29	0,99	0,28
			Faltante	0,97	0,34	0,95	0,32	0,97	0,32	0,98	0,32	0,91	0,29
		MARx	Completa	0,99	0,28	0,96	0,28	0,97	0,29	0,97	0,29	0,97	0,29
			Faltante	0,99	0,37	**0,23**	0,34	**0,12**	0,39	**0,14**	0,37	**0,21**	0,30
		MARy	Completa	0,97	0,29	0,99	0,28	0,98	0,29	0,96	0,28	0,96	0,28
			Faltante	0,97	0,34	0,96	0,32	0,98	0,32	0,97	0,32	0,94	0,29
		NI	Completa	0,99	0,29	0,96	0,28	0,95	0,28	0,99	0,28	0,98	0,28
			Faltante	**0,15**	0,33	**0,52**	0,35	**0,60**	0,38	**0,66**	0,35	0,97	0,29
	60	MCAR	Completa	0,98	0,29	0,98	0,29	0,96	0,29	0,96	0,28	0,98	0,29
			Faltante	0,95	0,46	0,96	0,35	0,93	0,35	0,93	0,35	**0,89**	0,28
		MARx	Completa	0,99	0,29	0,97	0,28	0,97	0,28	0,99	0,29	0,98	0,29
			Faltante	0,94	0,60	**0,22**	0,37	**0,01**	0,61	**0,01**	0,55	**0,01**	0,34
		MARy	Completa	0,98	0,29	0,97	0,29	0,98	0,29	0,98	0,29	0,97	0,29
			Faltante	0,95	0,46	0,95	0,35	0,94	0,36	0,91	0,35	0,91	0,29
		NI	Completa	0,98	0,29	0,97	0,29	0,98	0,28	0,97	0,29	0,98	0,28
			Faltante	**0,01**	0,41	**0,32**	0,36	**0,38**	0,45	**0,41**	0,40	0,90	0,29

Nota.

Mecanismos. MCAR: Faltantes Completamente al azar, MARx: Faltantes al azar dependientes de X2, MARy: Faltantes al azar dependientes de Y, NI: Faltantes No Ignorables.

Técnicas. ACC: Análisis de Casos Completos,ACD: Análisis de Casos Disponibles, IMNC: Imputación de la Media No Condicional, IMC: Imputación de la Media Condicional, EM: Algoritmo EM.

Los valores en negrita corresponden a intervalos que no alcanzan el 90% de cubrimento del parámetro

Tabla 8. Cubierta y amplitud de intervalos de confianza para el Coeficiente de X1 (β1)

Tratamiento				Técnicas de tratamiento									
n	p %	Mec.	Ind. muestra	ACC		ACD		IMNC		IMC		EM	
				Cub.	Ancho	Cub.	Ancho	Cub.	Ancho	Cub.	Ancho	Cub.	Ancho
50	10	MCAR	Completa	0,92	0,69	0,94	0,69	0,96	0,67	0,93	0,69	0,94	0,69
			Faltante	0,92	0,74	0,96	0,75	0,96	0,74	0,94	0,75	0,94	0,70
		MARx	Completa	0,95	0,70	0,92	0,69	0,93	0,69	0,95	0,70	0,92	0,68
			Faltante	0,95	0,76	0,95	0,81	0,94	0,82	0,97	0,81	0,90	0,71
		MARy	Completa	**0,89**	0,69	0,93	0,69	0,94	0,70	0,96	0,70	0,94	0,68
			Faltante	0,91	0,73	0,94	0,74	0,94	0,75	0,95	0,76	0,91	0,68
		NI	Completa	0,92	0,69	0,95	0,70	0,91	0,69	0,94	0,69	0,95	0,69
			Faltante	**0,88**	0,72	0,91	0,83	**0,85**	0,83	0,95	0,81	0,93	0,69
	30	MCAR	Completa	0,93	0,68	0,96	0,69	0,91	0,69	0,94	0,69	0,96	0,70
			Faltante	0,93	0,83	0,97	0,90	0,94	0,89	0,95	0,93	0,91	0,70
		MARx	Completa	0,97	0,68	0,95	0,68	0,95	0,70	0,91	0,70	0,97	0,69
			Faltante	0,93	0,90	0,98	0,96	0,97	1,03	0,96	1,05	0,95	0,78
		MARy	Completa	0,94	0,68	0,93	0,70	0,94	0,67	0,94	0,69	0,97	0,70
			Faltante	0,95	0,84	0,91	0,89	0,93	0,87	0,96	0,93	0,93	0,70
		NI	Completa	0,93	0,69	0,93	0,69	0,94	0,69	0,92	0,70	0,93	0,69
			Faltante	**0,89**	0,79	0,91	0,93	**0,80**	0,96	0,94	0,95	0,91	0,71
	60	MCAR	Completa	0,91	0,68	0,96	0,70	0,95	0,70	0,96	0,69	0,96	0,59
			Faltante	0,93	1,20	0,97	1,26	0,95	1,30	0,97	1,42	**0,78**	0,57
		MARx	Completa	0,91	0,69	0,93	0,70	0,96	0,69	0,96	0,69	0,96	0,59
			Faltante	0,95	1,29	0,97	1,17	0,96	1,47	0,99	1,58	**0,45**	0,68
		MARy	Completa	0,90	0,70	0,93	0,69	0,91	0,68	0,95	0,69	0,93	0,59
			Faltante	**0,90**	1,18	0,93	1,23	0,94	1,25	0,98	1,34	**0,84**	0,57
		NI	Completa	0,94	0,70	0,95	0,68	0,93	0,69	0,95	0,70	0,97	0,58
			Faltante	**0,85**	0,89	0,93	1,03	**0,81**	1,17	0,95	1,15	**0,85**	0,60
100	10	MCAR	Completa	0,95	0,48	0,93	0,47	0,96	0,48	0,95	0,48	0,93	0,48
			Faltante	0,93	0,50	0,92	0,52	0,95	0,52	0,95	0,53	0,92	0,48
		MARx	Completa	0,93	0,48	0,95	0,47	0,94	0,48	0,96	0,47	0,93	0,47
			Faltante	0,93	0,52	0,97	0,57	0,93	0,56	0,97	0,56	0,94	0,50
		MARy	Completa	0,96	0,48	0,91	0,48	0,94	0,47	0,92	0,48	0,93	0,48
			Faltante	0,95	0,50	0,91	0,52	0,92	0,51	0,91	0,52	0,91	0,47
		NI	Completa	0,93	0,48	0,96	0,47	0,96	0,48	0,96	0,47	0,95	0,48
			Faltante	0,91	0,50	**0,85**	0,57	**0,76**	0,57	0,98	0,55	0,94	0,49
	30	MCAR	Completa	0,94	0,48	0,95	0,48	0,94	0,48	0,91	0,48	0,96	0,48
			Faltante	0,96	0,58	0,92	0,61	0,91	0,62	0,94	0,64	**0,90**	0,48
		MARx	Completa	0,96	0,47	0,96	0,47	0,94	0,47	0,93	0,48	0,97	0,33
			Faltante	0,94	0,62	0,95	0,66	0,92	0,70	0,96	0,73	0,93	0,33
		MARy	Completa	0,94	0,48	0,95	0,47	0,95	0,47	0,92	0,48	0,94	0,47
			Faltante	0,94	0,58	0,91	0,61	0,92	0,61	0,95	0,63	0,91	0,47
		NI	Completa	0,93	0,48	0,94	0,47	0,95	0,48	0,94	0,48	0,96	0,48
			Faltante	**0,83**	0,54	**0,86**	0,64	**0,60**	0,67	0,93	0,65	0,93	0,49
	60	MCAR	Completa	0,98	0,47	0,91	0,48	0,94	0,48	0,93	0,47	0,95	0,48
			Faltante	0,96	0,77	**0,90**	0,85	0,90	0,87	0,99	0,92	**0,83**	0,47
		MARx	Completa	0,95	0,48	0,94	0,48	0,95	0,48	0,94	0,48	0,94	0,47
			Faltante	0,97	0,88	0,97	0,83	0,93	1,03	0,98	1,08	**0,82**	0,56
		MARy	Completa	0,95	0,48	0,97	0,48	0,94	0,48	0,91	0,48	0,93	0,47
			Faltante	0,93	0,77	**0,87**	0,84	**0,88**	0,85	0,97	0,92	**0,81**	0,47
		NI	Completa	0,96	0,48	0,95	0,48	0,95	0,47	0,97	0,48	0,95	0,47
			Faltante	**0,82**	0,60	0,94	0,73	**0,51**	0,79	**0,89**	0,80	**0,88**	0,47

* Continúa pág. Siguiente

Tabla 8. (Continuación)

Tratamiento				Técnicas de tratamiento									
n	p %	Mec.	Ind. muestra	ACC		ACD		IMNC		IMC		EM	
				Cub.	Ancho	Cub.	Ancho	Cub.	Ancho	Cub.	Ancho	Cub.	Ancho
200	10	MCAR	Completa	0,95	0,33	0,97	0,33	0,95	0,33	0,95	0,33	0,96	0,33
			Faltante	0,96	0,35	0,96	0,36	0,95	0,36	0,97	0,36	0,94	0,33
		MARx	Completa	0,96	0,33	0,97	0,33	0,95	0,33	0,94	0,33	0,96	0,33
			Faltante	0,94	0,36	0,97	0,40	**0,88**	0,40	0,97	0,40	0,93	0,35
		MARy	Completa	0,97	0,33	0,98	0,33	0,96	0,33	0,97	0,33	0,96	0,33
			Faltante	0,97	0,35	0,95	0,36	0,95	0,36	0,97	0,36	0,95	0,33
		NI	Completa	0,96	0,33	0,96	0,33	0,95	0,33	0,96	0,33	0,94	0,33
			Faltante	**0,88**	0,35	**0,76**	0,39	**0,44**	0,40	0,93	0,39	0,92	0,34
	30	MCAR	Completa	0,97	0,33	0,97	0,33	0,94	0,33	0,98	0,33	0,97	0,33
			Faltante	0,96	0,40	**0,89**	0,42	**0,87**	0,43	0,96	0,44	**0,90**	0,33
		MARx	Completa	0,96	0,33	0,97	0,33	0,96	0,33	0,95	0,33	0,95	0,33
			Faltante	0,95	0,44	0,97	0,47	0,91	0,50	0,97	0,50	**0,87**	0,37
		MARy	Completa	0,97	0,33	0,94	0,33	0,93	0,33	0,97	0,33	0,97	0,33
			Faltante	0,93	0,40	**0,89**	0,43	**0,89**	0,43	0,97	0,44	0,93	0,33
		NI	Completa	0,95	0,33	0,97	0,33	0,95	0,33	0,94	0,33	0,96	0,33
			Faltante	**0,71**	0,37	**0,87**	0,45	**0,24**	0,46	**0,87**	0,46	0,94	0,34
	60	MCAR	Completa	0,96	0,33	0,97	0,33	0,99	0,33	0,93	0,33	0,98	0,33
			Faltante	0,94	0,54	**0,84**	0,60	**0,85**	0,59	0,98	0,64	**0,84**	0,33
		MARx	Completa	0,97	0,33	0,96	0,33	0,97	0,33	0,96	0,34	0,93	0,34
			Faltante	0,95	0,61	0,97	0,56	0,91	0,70	0,99	0,74	**0,77**	0,40
		MARy	Completa	0,96	0,33	0,94	0,33	0,96	0,33	0,95	0,33	0,97	0,34
			Faltante	0,94	0,53	**0,83**	0,59	**0,84**	0,59	0,98	0,64	**0,85**	0,34
		NI	Completa	0,94	0,33	0,95	0,33	0,97	0,33	0,98	0,33	0,93	0,33
			Faltante	**0,68**	0,42	0,96	0,51	**0,24**	0,55	**0,80**	0,55	**0,89**	0,33

Nota.

Mecanismos. MCAR: Faltantes Completamente al azar, MARx: Faltantes al azar dependientes de X2, MARy: Faltantes al azar dependientes de Y, NI: Faltantes No Ignorables.

Técnicas. ACC: Análisis de Casos Completos,ACD: Análisis de Casos Disponibles, IMNC: Imputación de la Media No Condicional, IMC: Imputación de la Media Condicional, EM: Algoritmo EM.

Los valores en negrita corresponden a intervalos que no alcanzan el 90% de cubrimento del parámetro

Tabla 9. Cubierta y amplitud de intervalos de confianza para el Coeficiente de X2(β2)

Tratamiento				Técnicas de tratamiento									
n	p %	Mec.	Ind. muestra	ACC		ACD		IMNC		IMC		EM	
				Cub.	Ancho	Cub.	Ancho	Cub.	Ancho	Cub.	Ancho	Cub.	Ancho
50	10	MCAR	Completa	0,93	0,69	0,92	0,69	0,97	0,68	0,96	0,69	0,97	0,69
			Faltante	0,90	0,73	0,90	0,71	0,95	0,71	0,95	0,73	0,94	0,69
		MARx	Completa	0,94	0,69	0,97	0,69	0,92	0,69	0,97	0,70	0,93	0,69
			Faltante	0,95	0,72	**0,87**	0,72	**0,84**	0,73	0,95	0,74	0,91	0,69
		MARy	Completa	0,95	0,68	0,95	0,69	0,98	0,69	0,93	0,69	0,92	0,68
			Faltante	0,95	0,72	0,94	0,71	0,96	0,71	0,91	0,73	0,92	0,68
		NI	Completa	0,95	0,69	0,91	0,69	**0,89**	0,69	0,95	0,68	0,95	0,69
			Faltante	0,94	0,72	**0,81**	0,74	**0,75**	0,74	0,93	0,74	0,93	0,69
	30	MCAR	Completa	0,93	0,68	0,96	0,68	0,94	0,69	0,94	0,69	0,93	0,69
			Faltante	0,96	0,83	**0,81**	0,74	**0,82**	0,74	0,93	0,82	**0,89**	0,70
		MARx	Completa	0,96	0,69	0,97	0,69	0,96	0,70	0,94	0,71	0,98	0,68
			Faltante	0,94	0,82	**0,88**	0,73	**0,64**	0,77	0,95	0,83	0,93	0,70
		MARy	Completa	0,95	0,69	0,94	0,70	0,96	0,67	0,93	0,69	0,94	0,70
			Faltante	0,93	0,84	**0,79**	0,75	**0,80**	0,72	0,95	0,81	0,94	0,70
		NI	Completa	0,95	0,69	0,95	0,70	0,91	0,68	0,95	0,69	0,96	0,69
			Faltante	0,91	0,79	**0,74**	0,75	**0,45**	0,75	**0,88**	0,78	**0,89**	0,70
	60	MCAR	Completa	0,95	0,68	0,95	0,71	0,98	0,70	0,93	0,70	0,96	0,59
			Faltante	0,97	1,18	**0,54**	0,78	**0,61**	0,78	0,94	1,05	**0,78**	0,57
		MARx	Completa	0,93	0,69	0,94	0,69	0,95	0,68	0,96	0,70	0,96	0,59
			Faltante	0,94	1,15	**0,70**	0,75	**0,49**	0,78	0,94	1,02	**0,45**	0,68
		MARy	Completa	0,93	0,70	0,98	0,69	0,92	0,68	0,95	0,69	0,93	0,59
			Faltante	0,94	1,19	**0,54**	0,77	**0,59**	0,77	0,94	1,00	**0,84**	0,57
		NI	Completa	0,94	0,69	0,96	0,69	0,93	0,69	0,93	0,69	0,97	0,58
			Faltante	0,91	0,90	**0,81**	0,75	**0,39**	0,76	**0,82**	0,84	**0,85**	0,60
100	10	MCAR	Completa	0,97	0,47	0,96	0,47	0,97	0,48	0,92	0,48	0,95	0,48
			Faltante	0,97	0,50	0,91	0,49	0,93	0,49	0,94	0,50	0,94	0,48
		MARx	Completa	0,94	0,47	0,92	0,48	0,97	0,47	0,95	0,47	0,95	0,48
			Faltante	0,90	0,49	**0,83**	0,51	**0,83**	0,50	0,91	0,51	0,93	0,48
		MARy	Completa	0,96	0,48	0,94	0,48	0,96	0,48	0,93	0,48	0,96	0,48
			Faltante	0,96	0,51	0,91	0,49	0,93	0,49	0,92	0,50	0,95	0,48
		NI	Completa	0,98	0,48	0,96	0,47	0,94	0,48	0,96	0,47	0,95	0,48
			Faltante	0,97	0,50	**0,70**	0,51	**0,56**	0,51	**0,86**	0,51	0,92	0,48
	30	MCAR	Completa	0,95	0,48	0,95	0,47	0,93	0,47	0,96	0,48	0,97	0,47
			Faltante	0,95	0,58	**0,68**	0,51	**0,63**	0,51	0,95	0,56	0,93	0,47
		MARx	Completa	0,98	0,47	0,94	0,47	0,92	0,47	0,97	0,47	0,94	0,33
			Faltante	0,97	0,56	**0,65**	0,50	**0,50**	0,52	**0,89**	0,56	0,93	0,33
		MARy	Completa	0,96	0,47	0,93	0,48	0,93	0,47	0,95	0,47	0,96	0,48
			Faltante	0,97	0,58	**0,65**	0,52	**0,70**	0,51	0,96	0,55	0,90	0,47
		NI	Completa	0,96	0,47	0,96	0,47	0,95	0,48	0,96	0,48	0,97	0,48
			Faltante	**0,89**	0,53	**0,51**	0,51	**0,30**	0,52	**0,75**	0,54	0,91	0,48
	60	MCAR	Completa	0,97	0,47	0,95	0,48	0,96	0,48	0,96	0,47	0,95	0,48
			Faltante	0,91	0,78	**0,24**	0,53	**0,29**	0,53	0,91	0,67	**0,87**	0,47
		MARx	Completa	0,96	0,48	0,93	0,47	0,97	0,48	0,94	0,48	0,93	0,48
			Faltante	0,92	0,78	**0,51**	0,52	**0,21**	0,54	**0,86**	0,65	**0,84**	0,49
		MARy	Completa	0,96	0,48	0,95	0,47	0,96	0,47	0,93	0,48	0,96	0,47
			Faltante	0,94	0,77	**0,31**	0,53	**0,27**	0,53	0,93	0,68	**0,80**	0,47
		NI	Completa	0,94	0,47	0,95	0,48	0,91	0,47	0,95	0,48	0,97	0,47
			Faltante	**0,83**	0,60	**0,56**	0,52	**0,18**	0,53	**0,60**	0,58	**0,89**	0,47

* Continúa pág. Siguiente

Tabla 9. (Continuación)

Tratamiento				Técnicas de tratamiento									
n	p %	Mec.	Ind. muestra	ACC		ACD		IMNC		IMC		EM	
				Cub.	Ancho	Cub.	Ancho	Cub.	Ancho	Cub.	Ancho	Cub.	Ancho
200	10	MCAR	Completa	0,96	0,33	0,95	0,33	0,96	0,33	0,96	0,33	0,96	0,33
			Faltante	0,97	0,35	**0,87**	0,34	**0,89**	0,34	0,97	0,35	0,95	0,33
		MARx	Completa	0,96	0,33	0,96	0,33	0,98	0,33	0,94	0,33	0,97	0,33
			Faltante	0,96	0,35	**0,73**	0,35	**0,65**	0,35	**0,87**	0,36	0,94	0,33
		MARy	Completa	0,96	0,33	0,96	0,33	0,96	0,33	0,97	0,33	0,97	0,33
			Faltante	0,97	0,35	**0,85**	0,34	**0,89**	0,34	0,98	0,35	0,97	0,33
		NI	Completa	0,96	0,33	0,93	0,33	0,96	0,33	0,95	0,33	0,98	0,33
			Faltante	**0,87**	0,35	**0,41**	0,35	**0,27**	0,36	**0,78**	0,36	0,94	0,34
	30	MCAR	Completa	0,97	0,33	0,96	0,33	0,97	0,33	0,97	0,33	0,98	0,33
			Faltante	0,95	0,40	**0,56**	0,36	**0,45**	0,36	0,95	0,39	0,93	0,33
		MARx	Completa	0,96	0,33	0,97	0,33	0,98	0,33	0,95	0,33	0,95	0,33
			Faltante	0,96	0,39	**0,49**	0,35	**0,15**	0,37	**0,85**	0,39	**0,88**	0,33
		MARy	Completa	0,97	0,34	0,93	0,33	0,93	0,33	0,95	0,33	0,94	0,33
			Faltante	0,94	0,40	**0,41**	0,36	**0,52**	0,36	0,91	0,38	0,93	0,33
		NI	Completa	0,96	0,33	0,96	0,33	0,98	0,33	0,96	0,33	0,96	0,33
			Faltante	**0,87**	0,37	**0,23**	0,36	**0,04**	0,36	**0,50**	0,38	0,91	0,33
	60	MCAR	Completa	0,96	0,33	0,96	0,33	0,96	0,33	0,97	0,33	0,96	0,33
			Faltante	0,94	0,54	**0,06**	0,38	**0,06**	0,37	0,93	0,48	**0,84**	0,33
		MARx	Completa	0,98	0,33	0,96	0,33	0,95	0,33	0,96	0,33	0,95	0,33
			Faltante	0,96	0,54	**0,18**	0,36	**0,03**	0,37	**0,85**	0,46	**0,82**	0,34
		MARy	Completa	0,97	0,33	0,96	0,33	0,98	0,33	0,97	0,33	0,96	0,33
			Faltante	0,96	0,53	**0,04**	0,37	**0,02**	0,37	0,95	0,47	**0,86**	0,34
		NI	Completa	0,96	0,33	0,96	0,33	0,98	0,33	0,95	0,33	0,97	0,33
			Faltante	**0,68**	0,42	**0,33**	0,36	**0,01**	0,37	**0,37**	0,40	0,90	0,33

Nota.

Mecanismos. MCAR: Faltantes Completamente al azar, MARx: Faltantes al azar dependientes de X2, MARy: Faltantes al azar dependientes de Y, NI: Faltantes No Ignorables.

Técnicas. ACC: Análisis de Casos Completos,ACD: Análisis de Casos Disponibles, IMNC: Imputación de la Media No Condicional, IMC: Imputación de la Media Condicional, EM: Algoritmo EM.

Los valores en negrita corresponden a intervalos que no alcanzan el 90% de cubrimento del parámetro