Imputaciones múltiples, herramienta para la estimación de datos faltantes en la modelación de regresión

Multiple Imputations, tool for the estimation of missing data in regression modeling

Luis Miguel Mejía-Giraldo[1*], Luis Fernando Restrepo-Betancur[2]

Recibido para publicación: Septiembre 7 de 2018 - Aceptado para publicación: Noviembre 23 de 2018

RESUMEN

En los últimos años se ha apreciado un incremento en la investigación sobre problemas de datos faltantes, siendo la imputación múltiple una fundamental alternativa; donde los conjuntos de datos a menudo presentan complejidades que son actualmente difíciles de manejar de manera apropiada en el marco de probabilidad, pero relativamente simples de tratar con imputación; por esto, el presente artículo describe una serie de aspectos prácticos para aplicar dicha metodología en el caso de la modelación de captura de carbono para Colombia, con base en las bases de datos del Banco Mundial incluyendo datos faltantes alcanzando R^2 de 79,30%, resaltándose que al momento de estimar dichos datos y recalcular el modelo respectivo se evidencia un mayor R^2, siendo del 94,79%, lo cual evidencia una mejora sustancial del respectivo modelo de regresión lineal múltiple como tal.

Palabras clave: Análisis de Varianza, Coeficiente de Determinación, Ecuación, Sumas de Cuadrados, Validación.

ABSTRACT

In recent years there has been an increase in research on missing data problems, with multiple imputation being a fundamental alternative; where data sets often present complexities that are currently difficult to manage appropriately in the probability framework, but relatively simple to deal with imputation; For this reason, this article describes a series of practical aspects to apply this methodology in the case of carbon capture modeling for Colombia, based on the World Bank databases including missing data reaching R^2 of 79.2988%, highlighting that when estimating said data and recalculating the respective model, a greater R^2 is evidenced, being of 94.76901%, which evidences a substantial improvement of the respective multiple linear regression model as such.

Key words: Analysis of Variance, Coefficient of determination, Equation, Sum of Squares, Validation.

[1*] Universidad La Gran Colombia Facultad de Ingenierías Grupo de Investigación –GIDA-, Campus La Santa María, Armenia, mejiagluismiguel@miugca.edu.co, código postal 630008, Colombia.
[2] Universidad de Antioquia, Facultad de Ciencias Agrarias, Grupo de Investigación – GRICA-, Calle 67 53-108, Medellín, frbstatistical@yahoo.es, código postal 050034, Colombia.

INTRODUCCIÓN

El abordaje del manejo de datos faltantes es un aspecto que ha implicado diversos debates con respecto al tratamiento de los mismos, siendo Rubin (1976) quien estableció un referente teórico para los problemas de datos faltantes. Desde entonces, ha habido un aumento sustancial en la investigación de datos faltantes, y la mayoría de las aplicaciones de software ahora implementan una o más rutinas sofisticadas de manejo de estos.

A pesar del aumento en la investigación metodológica y la publicación concurrente de varios textos asociados al tema, como los plateados por Allison (2002), Carpenter & Kenward (2013), Enders (2010), Graham (2012), y Van Buuren (2006), quienes resaltan que la migración hacia una mejora en las prácticas analíticas ha sido lenta, sumándose que ha habido un enfoque hacia la eliminación de los casos con datos faltantes como lo mencionan Jelicic, Phelps y Lerner (2009), Peugh y Enders (2004), y Wood, White y Thompson (2004), quienes a su vez resaltan que dicha práctica gravita en lo llamado "los peores métodos disponibles para aplicaciones prácticas", lo cual implica una serie de imprecisiones que afectan la modelación de fenómenos y su posterior simulación.

Aunque las prácticas de presentación de informes han mejorado en los últimos años, la aplicación de técnicas modernas de tratamiento de datos faltantes está lejos de ser uniforme en diversas disciplinas. En términos generales, la reciente literatura de datos faltantes apoya el uso de estimación de máxima verosimilitud e imputación múltiple como lo resaltan Schafer y Graham (2002), quienes además resaltan que la estimación de máxima verosimilitud emplea un algoritmo de optimización iterativo que identifica parámetros estimaciones que maximizan el ajuste a los datos observados. Ejemplo de ello se puede aplicar en el análisis de regresión, donde las estimaciones de máxima verosimilitud son coeficientes que pueden minimizar la suma de las distancias estandarizadas al cuadrado entre los datos observados y los estimados en la regresión.

No obstante, la imputación múltiple crea varias versiones de un conjunto de datos, cada uno de las cuales puede contener diferentes estimaciones de los valores faltantes, donde el modelo de regresión se convierte en herramienta para completar los datos, tratando las variables incompletas como resultados y completando las variables como predictores, para –posteriormente- realizar uno o más análisis estadísticos en cada conjunto de datos completo para obtener estimaciones estándar de imputaciones y errores estándar, finalizando con agrupación de las estimaciones y los errores estándar en un único conjunto de resultados; sumado a lo anterior, autores como Gelman et al. (2014), Meng (1994) y Schafer (2003) plantearon que, cuando los datos se distribuyen normalmente, el conjunto común de variables de entrada y soportados en un tamaño de muestra suficientemente grande, denota que no hay ninguna razón teórica para esperar diferencias entre la estimación de máxima verosimilitud y la imputación múltiple, lo cual es corroborado por Collins, Schafer y Kam (2001) quienes afirman que los estudios empíricos sugieren que los dos métodos suelen arrojar estimaciones similares y errores estándar.

En este espacio, adicionar un párrafo que describa el objetivo de la investigación, haciendo énfasis sobre el caso donde se va a aplicar el método de estimación.

MATERIALES Y MÉTODOS

El presente estudio se realizó a partir de un análisis estocástico multivariante de emisión de CO_2 como factor de captura de carbono

y desarrollo sostenible para Colombia, plateado por Jimenez y Mejía (2014), donde los datos fueron obtenidos a partir de los registros de banco mundial (https://datos.bancomundial.org/indicador/IE.PPI.ENGY.CD?locations=CO), siendo el enfoque principal del trabajo de dichos autores una serie de medidas de análisis de componentes principales y regresión múltiple (siendo esta última el de interés para el presente trabajo) aplicados con el fin de determinar las variables que mayor correlación presentaban, así como los modelos de mayor significancia asociados a la captura de carbono como tal y cuyo conjunto de datos también incluye una serie de variables de fondo (como son los aspectos económicos, sociales y ambientales) y medidas de emisión CO_2, consumo de combustible líquido dependiendo principalmente de variables de producción de electricidad en kWh, producción de electricidada partir de fuentes renovables, el consumo de energía eléctrica en kWh, la población total y la inversión en energía con participación privada, respectivamente (Jiménez y Mejía, 2014).

Para efectos de determinar la dinámica de las respectivas imputaciones se aplicó la estimación para la variable de respuesta, formándose dos grupos de observaciones (ausentes y presentes en la misma) y se aplicó contraste de comparación bajo la metodología de Tukey para dos muestras y determinar así si existen diferencias significativas entre los dos grupos sobre las variables de interés, teniéndose en cuenta que si se encuentran diferencias significativas, dicho proceso de datos ausentes no es aleatorio, lo cual conlleva complejidad para el análisis de información. Posteriormente, se llevaron a cabo modelos de regresión lineal simple del tipo $y_i=\beta_0+\beta^*x_i$ para cada variable que reportó datos faltantes, para luego realizar las respectivas estimaciones de datos faltantes con base en los modelos obtenidos; es de agregar que para el presente estudio se utilizó el software Statgraphics Stratus.

RESULTADOS Y DISCUSIÓN

Para efectos del análisis del presente estudio de imputaciones múltiples, se estableció –inicialmente- el porcentaje de valores observados para un subconjunto de variables que utilizo a lo largo del trabajo de Jiménez y Mejía, desde que se reportaron los datos desde 1971 a 2014 (Tabla 1).

Tabla 1. Porcentaje de valores observados por variable.

VARIABLE	PORCENTAJE DE VALORES OBSERVADOS
Emisiones de CO_2 del consumo de combustible líquido (% del total)	*100%*
Producción de electricidad a partir de fuentes renovables, excluida la hidroeléctrica (kilovatio-hora)	*80%*
Producción de electricidad a partir de fuentes renovables, excluida la hidroeléctrica (% del total)	*80%*
Población total)	*100%*
Inversión en energía con participación privada (US$ a precios actuales)	*40%*

Fuente: Los autores, 2017

Es de agregar que los proyectos de esta índole giran en torno al uso de herramientas estadísticas de alto nivel de complejidad como son aquellas del tipo multivariante, como son los modelos de regresión lineal múltiple, tendientes a determinar los parámetros significativos como lo denotan Jiménez y Mejía (2014) que orientaron su trabajo hacia la comprensión del concepto de captura de carbono, con enfoque hacia el análisis de la dinámica de los factores de relevancia para Colombia y la comprensión de dicha dinámica de captura de carbono en el país con base en su impacto económico y ambiental. Es de resaltar que la literatura de datos faltantes a veces recomienda un procedimiento de manejo de datos faltantes a gran escala que da cuenta de docenas de variables (Rubin, 1996). Sin embargo, la complejidad de los conjuntos de datos a partir de metadatos generalmente excluye esta estrategia, y los

tamaños de muestra que son típicos de dichos estudios también son un factor limitante (por ejemplo, el número de variables utilizadas para imputar los datos no puede exceder el número de casos y generalmente ser mucho más bajo), lo cual conlleva a centrarse en un análisis específico o una familia de análisis porque es más fácil implementar un procedimiento de manejo de datos faltantes que respeta las características importantes de los datos como en el presente caso y cuyos datos faltantes se pueden apreciar en la tabla 2:

Para ilustrar esta estrategia enfocada, se considera un análisis de regresión que modela la influencia de producción de electricidad a partir de fuentes renovables, excluida la hidroeléctrica en kilovatio-hora (PeFRKwH), producción de electricidad a partir de fuentes renovables, excluida la hidroeléctrica como porcentaje del total (PeFRPorc), consumo de energía eléctrica en kWh per cápita (CeE), población total PT) e inversión en energía con participación privada en US $ a precios actuales (IePP) sobre emisiones de

Tabla 2. Datos faltantes en la base de datos.

Emisiones de CO_2 del consumo de combustible líquido (% del total)	Producción de electricidad a partir de fuentes renovables, excluida la hidroeléctrica (kilovatio-hora)	Producción de electricidad a partir de fuentes renovables, excluida la hidroeléctrica (% del total)	Consumo de energía eléctrica (kWh per cápita)	Población total	Inversión en energía con participación privada (US$ a precios actuales)
Emisión CO_2	PeFRKwH	PeFRPorc	CeE	PY	IePP
7939,055				16480383	
9204,17				16982315	
9515,865				17500171	
10476,619				18033550	
10967,997				18581974	
11210,019				19144223	
12306,452				19721462	
12335,788				20311371	
14297,633				20905059	
14554,323				21490945	
17377,913				22061215	
18122,314	0	0	403,6354878	22611986	
18602,691	0	0	389,4706323	23146803	
19585,447	0	0	420,8324702	23674504	
21726,975	0	0	459,2692645	24208021	
20531,533	189000000	1,461603898	441,6533475	24756973	
21836,985	202000000	1,473807092	456,1787621	25323406	
22519,047	192000000	1,258438749	498,2411397	25905127	
22555,717	206000000	1,186430916	551,6907561	26502166	
23633,815	216000000	1,136543015	589,3740361	27113512	
23923,508	227000000	1,110241612	618,0352514	27737900	
23032,427	212000000	1,020359051	614,6040855	28375991	
23553,141	231000000	1,041384907	638,3331584	29027162	
25250,962	234000000	1,021611002	642,2993103	29687094	74639729,9
24660,575	255000000	1,008144224	696,1759515	30350086	42417056,87
25008,94	267000000	0,995265963	729,3701652	31011688	61430435,37
26156,711	266000000	0,903839619	781,2811811	31669776	124074514
26970,785	268000000	0,850874686	757,6956363	32324325	168505777,4
28301,906	274000000	0,830001212	776,4847485	32975535	241156897,4
29427,675	274000000	0,783528739	821,9318065	33624444	302600131,1
31136,497	274000000	0,753637539	841,8640935	34271565	395865756,1
32023,911	269000000	0,727223574	842,2887733	34916766	444299836,1
35679,91	340000000	1,032054395	753,8243403	35558682	643840239,9
36064,945	389000000	1,00884359	846,1902981	36195168	282000000
39592,599	444000000	1,072230674	883,9726613	36823537	258000000
31818,559	477000000	1,116677592	895,3854119	37441977	1273200000
32603,297	496000000	1,143858678	887,8016837	38049038	1495200000
34825,499	483000000	1,082838247	895,8631595	38645411	3382926000
35386,55	532000000	1,184934405	874,444252	39234062	597100000
30417,765	457000000	1,067208444	810,6877073	39819279	263600000
29350,668	496000000	1,150144928	829,3989416	40403958	81900000
28565,93	496000000	1,141752221	850,4251723	40988909	53000000
28745,613	491000000	1,09009369	858,6206682	41572491	51300000
29669,697	495000000	1,06357835	862,5182615	42152151	50700000
29548,686	519000000	1,04386653	898,3441057	42724163	53000000
30964,148	552000000	1,096608856	895,8168431	43285634	61700000
34000,424	593000000	1,102948015	947,1270942	43835722	147600000
34275,449	581000000	1,052078806	974,3868628	44374572	499100000
34081,098	591000000	1,056526869	974,2203965	44901544	196300000
35133,527	599000000	1,048137325	1049,978201	45416181	142100000
34862,169	2456000000	4,133010232	1077,984569	45918097	235016000
34972,179	2040000000	3,343988198	1121,391104	46406646	684000000
38708,852	2008000000	3,221149219	1149,963818	46881475	45000000
42405,188	2107000000	3,03344419	1287,202426	47342981	1400000000
39808,952	2201000000	3,147883295	1289,569693	47791911	774600000
55	44	44	44	55	32

Fuente: Los autores, 2017

CO_2 del consumo de combustible líquido como porcentaje del total (EmisiónCO_2), obteniéndose la siguiente ecuación: Emisión CO_2 =

$$21926{,}1 - 0{,}00000875614 * PeFRKwH + 4315{,}85 * PeFRPorc + 54{,}1536 * CeE - 0{,}000921251 * PT - 6{,}31801E\text{-}7 * IePP$$

Se eligió este modelo porque presenta complejidades que son comunes en la investigación de esta índole (como son mezclas de variables categóricas y continuas, puntajes compuestos) y complejidades que, a menudo favorecen la imputación múltiple como solución y donde la verosimilitud de los diferentes mecanismos de datos faltantes suele ser la principal preocupación, a pesar que el modelo es altamente significativo (valor p=0,0001), como se aprecia en la tabla de análisis de varianza (Tabla 3) y en sus pruebas complementarias de coeficiente de determinación que fue del 79,30%.

Tabla 3. Análisis de varianza del modelo de regresión lineal múltiple sin imputaciones Múltiples

Fuente	Gl	Suma de cuadrados	Cuadrado medio	Razón F	Valor p
Modelo	5	2,49E+08	4,97E+07	12,26	0,0001
Residuo	16	6,49E+07	4,06E+06		
Total (corregido)	21	3,14E+08			
R-cuadrado	79.3				

La teoría de datos ausentes de Rubin (1976) define un conjunto de datos hipotéticos sin valores perdidos, y divide los datos realizados en componentes observados y faltantes y es útil ver las partes faltantes como puntajes variables latentes, cuyos valores residen solo en la matriz de datos hipotéticamente completa.

No obstante, el fundamento de la teoría de Rubin es que los indicadores de falta de respuesta pueden no estar completamente relacionados con los datos, o pueden estar sistemáticamente relacionados con los puntajes observados o latentes (o ambos). Volviendo a las variables en la ecuación (1), se pueden crear cuatro indicadores de datos faltantes, para producción de electricidad a partir de fuentes renovables, excluida la hidroeléctrica (kilovatio-hora), producción de electricidad a partir de fuentes renovables, excluida la hidroeléctrica (% del total) consumo de energía eléctrica (kWh per cápita) e Inversión en energía con participación privada (US$ a precios actuales).

Para el presente caso, se aprecia una relación entre los indicadores de falta de respuesta y los datos, conocido como tipo de mecanismo NMAR (que no falta al azar), dada la diferencia significativa entre registros con datos faltantes (código 0) en contraste con aquellos con datos no faltantes (código 1), las cuales definen diferentes tipos de faltantes sistemáticos y para el presente caso, todas las variables con datos faltantes poseen variación significativa (Valor p <0,05) con respecto a los registros de datos faltantes codificadas con ceros, frente a aquellos que están completos, codificados con unos, lo cual se evidencia en la evaluación de la emisión de CO_2 en función de las demás variables explicatorias PeFRKwH, CeE, PeFRPorc e IePP, respectivamente (Figura 1)

En la práctica, es difícil determinar qué mecanismo se aplica mejor a un análisis particular porque las condiciones de Rubin implican que NMAR permite un enlace. Por supuesto, sin acceso a los puntajes latentes, es imposible saber si predicen la desaparición, por lo que finalmente conlleva a adoptar una suposición no comprobable sobre el proceso que causó la pérdida de datos tal como lo resalta Raykov (2011); además, la obtención de estimaciones precisas de un mecanismo de NMAR requiere enfoques de modelado complejos que introducen los indicadores de una forma u otra (Enders, 2010; Muthen, Asparouhov, Hunter y Leuchter, (2010), para el presente caso se estimó el valor estimado faltante para cada variable con datos faltantes en función de la variable emisión. Obteniéndose

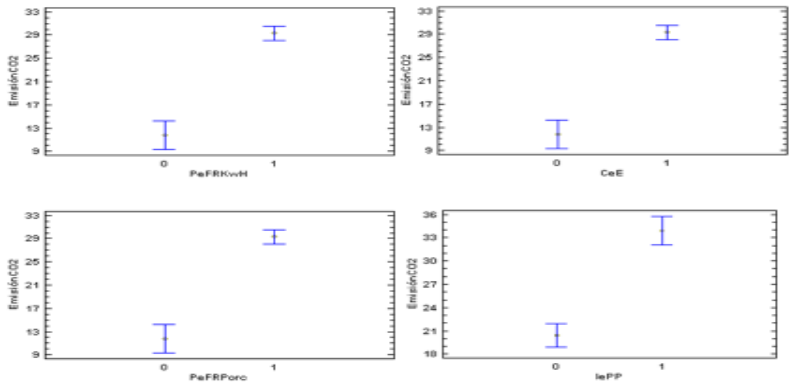

Figura 1. Comparaciones de datos faltantes versus completos con respecto a la variable de respuesta emisión de CO_2.

Tabla 4. Base de datos con estimación de datos bajo imputación múltiple

Emisiones de CO_2 del consumo de combustible líquido (% del total)	Producción de electricidad a partir de fuentes renovables, excluida la hidroeléctrica (kilovatio-hora)	Producción de electricidad a partir de fuentes renovables, excluida la hidroeléctrica (% del total)	Consumo de energía eléctrica (kWh per cápita)	Población total	Inversión en energía con participación privada (US$ a precios actuales)
Emisión CO_2	PeFRKwH	PeFRPorc	CeE	PY	IePP
7939,055	-893090484	-0,478233018	78,73455168	16480383	-870225111,1
9204,17	-808104233,2	-0,37731821	120,9911638	16982315	-801176526
9515,865	-787165591,7	-0,352455142	131,4022132	17500171	-784164555,8
10476,619	-722625308,4	-0,275818389	163,4927419	18033550	-731727659,3
10967,997	-689616155,9	-0,236622493	179,905455	18581974	-704908788,5
11210,019	-673357916,6	-0,217317051	187,9893286	19144223	-691699494
12306,452	-599703165,9	-0,129857551	224,6117258	19721462	-631857386,9
12335,788	-597732470,3	-0,127517498	225,5915893	20311371	-630256260,3
14297,633	-465942197,2	0,028973581	291,1199589	20905059	-523180918,3
14554,323	-448698610,1	0,049449049	299,6937643	21490945	-509171060,4
17377,913	-259019151,8	0,2746792	394,0056233	22061215	-355062624,2
18122,314	0	0	403,6354878	22611986	-314434036,4
18602,691	0	0	389,4706323	23146803	-288215588,2
19585,447	0	0	420,8324702	23674504	-234577846,7
21726,975	0	0	459,2692645	24208021	-117695604,2
20531,533	189000000	1,461603898	441,6533475	24756973	-182941513,5
21836,985	202000000	1,473807092	456,1787621	25323406	-111691379,4
22519,047	192000000	1,258438749	498,2411397	25905127	-74465185,69
22555,717	206000000	1,186430916	551,6907561	26502166	-72463777,43
23633,815	216000000	1,136543015	589,3740361	27113512	-13622374,5
23923,508	227000000	1,110241612	618,0352514	27737900	2188750,781
23032,427	212000000	1,020359051	614,6040855	28375991	-46445470,01
23553,141	231000000	1,041384907	638,3331584	29027162	-18025472,68
25250,962	234000000	1,021611002	642,2993103	29687094	74639729,9
24660,575	255000000	1,008144224	696,1759515	30350086	42417056,87
25008,94	267000000	0,995265963	729,3701652	31011688	61430435,37
26156,711	266000000	0,903839619	781,2811811	31669776	124074514
26970,785	268000000	0,850874686	757,6956363	32324325	168505777,4
28301,906	274000000	0,830001212	776,4847485	32975535	241156897,4
29427,675	274000000	0,783528739	821,9318065	33624444	302600131,1
31136,497	274000000	0,753637539	841,8640935	34271565	395865756,1
32023,911	269000000	0,727223574	842,2887733	34916766	444299836,1
35679,91	340000000	1,032054395	753,8243403	35558682	643840239,9
36064,945	389000000	1,00884359	846,1902981	36195168	282000000
39592,599	444000000	1,072230674	883,9726613	36823537	258000000
31818,559	477000000	1,116677592	895,3854119	37441977	1273200000
32603,297	496000000	1,143858678	887,8016837	38049038	1495200000
34825,499	483000000	1,082838247	895,8631595	38645411	3382926000
35386,55	532000000	1,184934405	874,444252	39234062	597100000
30417,765	457000000	1,067208444	810,6877073	39819279	263600000
29350,668	496000000	1,150144928	829,3989416	40403958	81900000
28565,93	496000000	1,141752221	850,4251723	40988909	53000000
28745,613	491000000	1,09009369	858,6206682	41572491	51300000
29669,697	495000000	1,06357835	862,5182615	42152151	50700000
29548,686	519000000	1,04386653	898,3441057	42724163	53000000
30964,148	552000000	1,096608856	895,8168431	43285634	61700000
34000,424	593000000	1,102948015	947,1270942	43835722	147600000
34275,449	581000000	1,052078806	974,3868628	44374572	499100000
34081,098	591000000	1,056526869	974,2203965	44901544	196300000
35133,527	599000000	1,048137325	1049,978201	45416181	142100000
34862,169	2456000000	4,133010232	1077,984569	45918097	235016000
34972,179	2040000000	3,343988198	1121,391104	46406646	684000000
38708,852	2008000000	3,221149219	1149,963818	46881475	45000000
42405,188	2107000000	3,03344419	1287,202426	47342981	1400000000
39808,952	2201000000	3,147883295	1289,569693	47791911	774600000
55	55	55	55	55	55

Fuente: Los autores, 2017

una serie de modelos de regresión para poder realizar la respectiva estimación, apreciándose que después de realizar la imputación múltiple con base en estimación de casos y/o variables con una alta proporción de datos ausentes, se aprecia una estimación de datos faltantes (Tabla 4).

En este caso, el modelo es altamente significativo (valor p=0,0001), como se aprecia en la tabla de análisis de varianza (Tabla 5) y en sus pruebas complementarias de coeficiente de determinación que fue del 94,79%.

Tabla 5. Análisis de varianza del modelo de regresión lineal múltiple con imputaciones múltiples.

Fuente	Gl	Suma de cuadrados	Cuadrado medio	Razón F	Valor
Modelo	5	4,19E+09	8,37E+08	178,3**	0,000
Residuo	49	2,30E+08	4,70E+06		
Total (corregido)	54	4,42E+09			
R-cuadrado		94,79			

Resaltándose que mantuvo la significancia del modelo y un incremento del 15,49% en el coeficiente de determinación R2, demostrándose que la estimación de dichos datos faltantes conllevan a la mejora en la estimación del nuevo modelo y a su bondad de ajuste, lo cual también resalta el impacto cuando se da la ausencia de datos al momento de modelar los datos y la mejora al ser estimados bajo imputación; alcanzándose para este caso, el presente modelo:

Emisión CO_2 =

$$7163{,}97 - 0{,}00000175265 * PeFRKwH + 283{,}657 * PeFRPorc + 31{,}3017 * CeE - 0{,}0000733374 * PT + 0{,}00000128655 * IePP$$

Cabe destacar que al usar un modelo de regresión para definir una distribución de valores de reemplazo plausibles para cada caso y luego usar la simulación por computadora para "trazar" un valor al azar de esta distribución, cada imputación es probada para determinar si cumple la respectiva normalidad de varianzas y es de agregar que el proceso de actualización de los parámetros de regresión imita el paso de imputación en el sentido de que las nuevas estimaciones se muestrean a partir de una distribución de valores fiables y estocásticamente dentro del contexto de la imputación múltiple (Rubin, 1987; Schafer, 1997), donde se utiliza un modelo de regresión multivariante donde las variables incompletas son resultados y las variables completas son predictores, donde todas las variables incompletas son imputadas, y cuyo soporte para la imputación se denomina imputación de ecuaciones encadenadas o especificación totalmente condicional e utiliza una serie o modelos de regresión univariante para generar imputaciones (Raghunathan, Lepkowski, Van Hoewyk y Solenberger, 2001; Van Buuren, 2012; van Buuren, Brand, Groothuis, Oudshoorn, & Rubin, 2006) como en este caso, donde las variables se imputan en forma de turno rotativo y cada una de ellas se completa convirtiéndose en un resultado en un primer paso y posteriormente en un predictor para la modelación.

CONCLUSIONES

Es posible la imputación múltiple siempre y cuando al convertir a los regresores en variables de respuesta y se estime su función, esta sea significativa, se podrá modelar la serie de datos en aras de la estimación de los datos faltantes y completar así la respectiva base de datos.

Para el presente caso, se aprecia que al estimar los datos faltantes, el efecto esperado del modelo prevalece (altamente significativo) y se aprecia un incremento en el coeficiente de determinación.

REFERENCIAS

Allison, P. 2002. Missing data. Newbury Park, CA: Sage. Asparouhov, T., & Muthen, B. (2010). Multiple imputation with Mplus. Retrieved from: http://www.statmodel.com/download/Imputations7.pdf.

Carpenter, J. and Kenward, M. 2013. Multiple imputation and its application. West Sussex, UK: Wiley.

Collins, L., Schafer, J. and Kam, C. 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6, 330e351.

Enders, C. 2010. Applied missing data analysis. New York: Guilford Press.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. 2014. Bayesian data analysis (3rd ed.). Boca Raton, FL: CRC Press.

Graham, J. 2012. Missing data: Analysis and design. New York: Springer.

Jelicic, H., Phelps, E. and Lerner, R. 2009. Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology.

Jimenez, J. and Mejía L. 2014. Multivariate Stochastic Analysis CO_2 emission factor for carbon sequestration and sustainable development for Colombia. Ugciencia 20, 64-71.

Little, R. and Rubin, D. 2002. Statistical analysis with missing data. Hoboken, NJ: Wiley.

Meng, X.1994. Multiple-imputation inferences with uncongenial sources of input. Statistical Science, 9, 538e558.

Peugh, J. and Enders, C. 2004. Missing data in educational research: A review of reporting practices and suggestions for improvement. Review of Educational Research, 74, 525e556.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J. and Solenberger, P. 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology, 27, 85e95.

Raykov, T. 2011. On Testability of missing data mechanisms in incomplete datasets. Structural Equation Modeling: A Multidisciplinary Journal, 18(3), 419e429. http://dx.doi.org/10.1080/10705511.2011.582396.

Rubin, D. 1976. Inference and missing data. Biometrika, 63, 581e592.

Rubin, D. 1987. Multiple imputation for nonresponse in surveys. Hoboken, New Jersey: Wiley.

Rubin, D. 1996. Multiple imputation after 18þ years. Journal of the American Statistical Association, 91, 473e489.

Schafer, J. 1997. Analysis of incomplete multivariate data. New York: Chapman &Hall.

Schafer, J. 2003. Multiple imputation in multivariate problems when the imputation and analysis models differ. Statistica Neerlandica, 57, 19e35.

Schafer, J. and Graham, J. 2002. Missing data: Our view of the state of the art. Psychological Methods, 7, 147e177.

Van Buuren, S. 2012. Flexible imputation of missing data. New York: Chapman &Hall.

Van Buuren, S., Brand, J. Groothuis-Oudshoorn, C. and Rubin, D. 2006. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation, 76, 1049e1064.

Widaman, K. 2006. Missing data: What to do with or without them. Monographs of the Society for Research in Child Development, 71, 42e64.

Wood, A., White, I. and Thompson, S. 2004. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals.Clinical Trials, 1, 368e376.