¿Cómo pueden afectar los valores faltantes la robustez de las estimaciones? Una aplicación práctica a partir de la Encuesta de Condiciones de Vida (EUROSTAT)

Mauro Mediavilla[*]

Resumo

Habitualmente, la literatura empírica en el campo de la economía aplicada basa sus estimaciones en bases de datos que sólo contienen las observaciones con información válida para todas las variables implicadas en el análisis. En la gran mayoría de los casos, esta práctica tiene como consecuencia una pérdida de información que, a su vez,puede originar dos problemas: una menor efectividad de la estimación y un incremento en la probabilidad de llegar a una mala especificación del modelo. El objetivo del trabajo es comprobar la importancia de la imputación de datos faltantes en la minimización de ambas problemáticas. Para ello se aplica la imputación múltiple como alternativa para obtener estimaciones más robustas que la generada por la simple no imputación. Los resultados muestran que la imputación mejora la efectividad de la estimación en términos de errores estándar más reducidos y; en relación con la especificidad del modelo, se constatan sensibles cambios en la significatividad de algunas variables según la muestra empleada para la estimación.

Palavras-chave: observaciones faltantes, efectividad, imputación múltiple

JEL: C13; C81

1 Introducción

Habitualmente, la literatura empírica en el campo de la economía aplicada basa sus estimaciones en bases de datos que sólo contienen las observaciones con información válida para todas las variables implicadas. Como consecuencia de esta pérdida de información, se pueden originar dos problemas: una menor efectividad de la estimación y un incremento en la probabilidad de llegar a una mala especificación del modelo.

En el primer caso, la menor efectividad se basa en un incremento de la varianza y de los desvíos estándar, además de aumentar la probabilidad de realizar estimaciones sobre una muestra escasamente representativa de la población analizada. En este sentido,

[*]Universitat de València – IEB. Mail: Mauro.Mediavilla@uv.es

si bien la potencial disminución de las observaciones no tiene por qué ser un problema en sí mismo, la representatividad se ve afectada cuando este fenómeno influye no aleatoriamente en las variables con valores no observados. Además, la falta de respuesta no aleatoria conllevaría la incorporación de sesgo en las estimaciones. En el segundo caso, una base de datos con problemas de información podría inducir al investigador a escoger una estructura modelística que no responda a la realidad de la muestra, incurriendo en dos posibles problemas: omisión de variables relevantes (infraespecificación) y/o inclusión de variables irrelevantes (sobreespecificación). Las consecuencias directas implican calcular estadísticos t y F distorsionados (Maddala, 1996).

En un intento por mejorar las estimaciones realizadas, algunos autores han sugerido que la manera más adecuada para disminuir estos potenciales problemas era substituyendo los valores perdidos mediante alguna técnica de imputación (Schafer, 1999; Acock, 2005; Graham, 2009; entre otros).

Asimismo, existe una creciente preocupación en las instituciones generadoras de fuentes de información primaria (bases de datos) en relación al tratamiento que realizan los usuarios de los valores perdidos y las consecuencias directas en las estimaciones obtenidas. En este sentido, existen algunos antecedentes de bases de datos imputadas por sus propios creadores. Por ejemplo: en EE.UU, el Survey of Consumer Finances(Kennickell, 1991); a nivel europeo, el Survey of Health, Ageing and Retirement in Europe (Christelis, 2011) y la Encuesta de Condiciones de Vida; por último, en España, la Encuesta Financiera de las Familias (Bover, 2004, 2008).

Con el objeto de comprobar las ventajas de la imputación, en el presente trabajo se aplicará la técnica de imputación múltiple que se comparará con la no imputación (*list wise deletion*). En línea con el trabajo de Allison (2000), Musil et al. (2002) y Barceló (2008)[1], se plantea un caso práctico –la posible relación entre las becas y el logro educativo de los individuos en el nivel secundario post-obligatorio en España–que permite al lector observar las principales diferencias obtenidas a partir de una misma base de datos original que se ve modificada según sea la opción metodológica escogida.

Los resultados obtenidos en el análisis descriptivo y en la regresión logística posterior indican que, en relación con la no imputación, la metodología de imputación de datos faltantes reduce los errores estándar obtenidos en la estimación y obtienen una mayor bondad del ajuste. Resulta trascendente, desde el punto de vista de la interpretación de los propios resultados, algunos cambios de significatividad ocurridos. Los mismos implican que la imputación múltiple, al trabajar con una mayor cantidad de información, permite al investigador obtener resultados más robustos. Asimismo, y dada su característica estocástica, la imputación múltiple permite preservar las características de la distribución original de los datos (media, varianza y dispersión).

[1] Barceló (2008) compara la efectividad de una estimación a partir de la imputación múltiple en relación con otras metodologías de substitución de valores perdidos como la eliminación directa (*list wise deletion*), la aplicación de la técnica *hot-decking* y la imputación no estocástica.

El trabajo se inicia, en el capítulo 2, con una breve revisión teórica de las diferentes metodologías de imputación (o no imputación) haciendo hincapié en la importancia de conocer el patrón por el cual se generan los valores perdidos. En el capítulo 3, se introduce el análisis muestral y la aplicación de las diferentes metodologías de tratamiento de datos. En el capítulo 4, se realiza la comparación de las estimaciones resultantes y, en el capítulo final, se introducen las principales conclusiones.

2 Análisis de los datos perdidos

La presencia de información faltante es un problema constante con el que deben lidiar los investigadores en las diferentes áreas de la economía aplicada. La misma se puede originar por un registro defectuoso de la información, por la falta de respuesta a las preguntas del encuestador (sea la misma total o parcial) o, directamente, por la ausencia natural de la información (Allison, 2001; Perez, 2004). Habitualmente, la falta total de información se resuelve mediante pesos muestrales que tiene en cuenta este fenómeno. En cambio, en el caso de la falta de respuesta parcial, su tratamiento estadístico obliga a la selección de una metodología de imputación (o no imputación) que debe ser el resultado de un análisis previo sobre el patrón por el cual se generan los valores perdidos.

2.1 Patrones de comportamiento de los valores perdidos

En primer lugar, se puede determinar que el patrón seguido por los valores perdidos es totalmente aleatorio (MCAR, *Missing Completely At Random*) o establecer un supuesto no tan restrictivo indicando que su generación ha sido de manera aleatoria (MAR, Missing At Random) (Rubin, 1976). En este caso, los valores perdidos pueden ser determinados a partir de otras variables observables siguiendo la siguiente forma funcional:

$$Pr(Y_{miss}|Y, X, \phi) = Pr(Y_{miss}|Y, X_{obs}, \phi),$$

donde ϕ hace referencia a los parámetros desconocidos. Desafortunadamente, no existe un test que categóricamente indique si el supuesto MAR se satisface, a causa - en la mayoría de los casos - de la imposibilidad de conocer la información faltante. Por tanto, se deben optar por vías indirectas de control como la prueba de las correlaciones dicotómicas, el test conjunto de aleatoriedad de Little o el análisis de sensibilidad de la estabilidad de los resultados, inferidos a partir de diferentes modelos de imputación (Perez, 2004; Carpenter et al., 2007). Por último, se puede suponer que los valores perdidos fueron generados de manera no aleatoria (MNAR, *Missings Not At Random*), por lo que seguirían un patrón sistémico específico (Rubin, 1976).

En el siguiente apartado se detallan las principales opciones de imputación existente en la literatura y su aplicabilidad según el patrón que se suponga sigue los valores perdi-

dos.

2.2 Diferentes metodologías para el tratamiento de datos con valores perdidos

2.2.1. No imputación o eliminación directa (listwisedeletion)

En primer lugar, la eliminación directa es una técnica comúnmente empleada en el análisis empírico en la cual se elimina la fila donde existe un vacío de información. Con el objeto de obtener una base completa sólo con valores originariamente válidos se provoca una reducción de la base de datos inicial (Perez, 2004). En caso de que los datos sigan un patrón MCAR, la eliminación directa de las observaciones generaría una muestra más pequeña pero aún representativa, lo que permitiría una estimación no sesgada de los estimadores. Aun así, este proceso conllevaría una pérdida de información y un incremento en los errores estándar. No obstante, si la base de datos no sigue un patrón MCAR, tal eliminación introduce un sesgo a la hora de la estimación de los parámetros, que afecta la efectividad de la propia estimación y podría inducir a una mala especificación del modelo utilizado (Howell, 2007).

2.2.2. Imputación determinística: imputación a la media

En segundo lugar, un tipo de imputación determinística se basa en la sustitución del dato perdido por la media de las observaciones válidas en el caso de una variable cuantitativa y por la moda en el caso de una variable cualitativa. El fundamento teórico para su empleo está basado en el hecho de que tanto la media como la varianza seguirían el valor esperado en el caso de una observación seleccionada al azar de una distribución normal. En el caso de valores perdidos con un patrón no estrictamente al azar (MAR o MNAR), esta metodología genera valores que reflejan escasamente los valores originales y aquí radica su principal debilidad como método de imputación.

Si bien su aplicación es muy sencilla, tiene como desventaja que modifica la distribución de la variable reduciendo su varianza, efecto que no debe sorprender dado que incrementa artificialmente la muestra sin agregar nueva información (Howell, 2007). Asimismo, esta aproximación altera las relaciones entre variables, reduciendo las covarianzas y las correlaciones (Pavia et al., 2013). Estas deficiencias implican su no validez como técnica de imputación de datos.

2.2.3. Imputación estocástica:imputación múltiple

La técnica de imputación múltiple, si bien es conocida desde la década de 1970 (Rubin, 1976), su desarrollo y aplicación se ha ido extendiendo en los últimos años como consecuencia, principalmente, de tres factores. En primer lugar, a causa de su introducción en los programas econométricos que han permitido su generalización entre la comu-

nidad académica (Little y Rubin, 1987; Rubin, 1996; Barnard y Meng, 1999; Van Buuren et al., 1999; Royston, 2004, 2005a; Reiter y Raghunathan, 2007; Sterne et al., 2009). En segundo lugar, a partir de la publicación de diferentes estudios que han demostrado las ventajas de la imputación múltiple frente a los procedimientos tradicionales de tratamiento de los valores perdidos (Gómez y Palarea, 2003; Acock, 2005; Ambler y Omar, 2007)[2]. Finalmente, y no menos importante, gracias a la existencia de ordenadores más potentes que hacen posible su implementación.

Las principales bondades de esta técnica, de característica estocástica, es que permite hacer un uso completo de los datos, obtener estimadores no sesgados, reflejar la incertidumbre que la no-respuesta parcial introduce en la estimación de los parámetros y preservar la dispersión de la distribución de la variable imputada (Rubin, 1996). Su aplicación se basa en sustituir los datos no observados por m>1 valores posibles simulados[3]. La aplicabilidad de este método se ha visto potenciada con la incorporación, en su esquema general, de los métodos de Monte Carlo basados en cadenas de Markov, conocidos como algoritmos MICE (*Multiple imputation by chained equations*)[4][5].Asimismo, a la imputación múltiple se la considera una metodología flexible que permite trabajar con datos multivariados y con distribuciones monótonas o arbitrarias de los valores perdidos. Finalmente, su aplicabilidad requiere que el patrón de distribución de los valores perdidos sea aleatorio (MCAR o MAR).

El proceso de estimación mediante la imputación múltiple consta de tres etapas (White et al., 2011). En la primera, cada valor perdido se reemplaza por un conjunto de m>1 valores generados por simulación, con los que se crean m matrices de datos completas. Para generar estos valores posibles se debe establecer un método de estimación particular para cada variable a imputar a partir de sus características propias (en el caso de aplicar ecuaciones encadenadas) y se deben escoger las variables que permitirán la generación de estos valores. Dicha selección debe asegurar incluir el máximo de variables predictivas de los valores faltantes posibles. A partir de los resultados de la estimación, se generan los resultados esperados para los casos de información ausente. En la segunda etapa, el investigador debe aplicar a cada matriz simulada el análisis deseado que se hubiese aplicado a la base original en caso de no haber contenido observaciones perdidas. Por último, se combinan los resultados obtenidos en cada matriz, aplicando las reglas de Rubin, para obtener una única estimación del parámetro estimado y del error estándar (Rubin,

[2]Otros métodos empleados para imputar valores perdidos y aquí no desarrollados son la imputación Hot Decking propuesta por Todeschini (1990) que propone estimar los valores perdidos mediante una estimación basada en sus "k-vecinos" más próximos, la imputación pair wise deletion y la substitución vía regresión. Asimismo, otras aproximaciones son las imputaciones por criterio experto, en bases longitudinales y basada en reglas lógicas (Pavia et al., 2013)

[3]Para una explicación teórica detallada, véase Schafer (1999).

[4]Otros métodos que han sido también empleados son el algoritmo EM (Expectation-Maximization) y la aproximación de Monte Carlo mediante máxima verosimilitud (Schafer, 1997).

[5]En los últimos años se han desarrollado extensiones, como la imputación múltiple mediante árboles de clasificación (Bacallao y Bacallao, 2010).

1987).Tales reglas, al tener en cuenta la variabilidad de los resultados entre las diferentes imputaciones, permiten reflejar la incerteza asociada con los valores faltantes.

En cuanto al número óptimo de bases de datos simuladas (m) no existe un total consenso en la literatura más allá de una relación directa entre el número de imputaciones y el porcentaje de información perdida. Se considera óptimo realizar entre 3 y 10 imputaciones en caso de tener una baja fracción de información perdida y hasta 50 imputaciones en caso de proporciones altas de datos no observados (Rubin, 1987; Horton y Lipsitz, 2001; Schafer, 1999; van Buurenet al., 1999; Kenward y Carpenter, 2007; Graham, 2009). Por su parte, Bodner (2008), von Hippel (2009) y White et al. (2011) recomiendan que el número de imputaciones sea igual (o mayor en algunos casos) al porcentaje de información perdida de la variable. Asimismo, STATA recomienda realizar un mínimo de 20 imputaciones con el objetivo de reducir los posibles errores muestrales generados a partir de las propias imputaciones (StataCorp, 2017). Finalmente, se debe destacar que además del porcentaje de información perdida, el propio tamaño de la base de datos es otro elemento que podría generar imprecisiones en el proceso de imputación (Bodner, 2008).

Para su uso empírico, esta metodología ha sido trasladada a los diferentes paquetes econométricos a partir de los trabajos de van Buurenet al. (1999) y la implementación directa, en el caso de STATA 10, a través del comando elaborado por Royston (Royston, 2004, 2005a, 2005b) llamado ice, el cual permite realizar las estimaciones a partir de una distribución arbitraria de los datos perdidos o mediante toda una familia de comandos mi que se incorporan a partir de la versión 11 (Royston y White, 2011). Asimismo, se han desarrollado otras aproximaciones en el caso del programa R, SPSS, SOLAS, SAS, Mplus y S-Plus (Horton y Lipsitz, 2001; Muñoz y Álvarez, 2009).

3 Caso aplicado: evaluación de la relación entre las becas y el logro educativo en España

3.1 Base de datos empleada: Encuesta de Condiciones de Vida (ECV - EUROSTAT)

La Encuesta de Condiciones de Vida (en adelante, ECV[6]) es una base de datos dirigida a hogares que viene a reemplazar el Panel de Hogares de la Unión Europea (PHOGUE), realizado durante el periodo 1994-2001. El objetivo fundamental que se persigue con la ECV es disponer de una fuente de referencia sobre estadísticas comparativas de la distribución de ingresos y la exclusión social en el ámbito europeo. Aunque los datos se refieren tanto a la dimensión transversal como a la longitudinal, se da prioridad a la producción de datos transversales de alta calidad en lo que respecta a la puntualidad y a la comparabilidad.

La componente longitudinal permite seguir en el tiempo a las mismas personas, estu-

[6]EU-SILC (European Union Statistics on Income and Living Conditions), en sus siglas en inglés.

diar los cambios que se producen en sus vidas cuando las condiciones y las políticas socioeconómicas se modifican, y cómo reaccionan a estos cambios. Formalmente, la ECV comienza en 2004 (si bien algunos países comenzaron más tarde y otros en 2003) y los ficheros de microdatos (tanto transversales como longitudinales) se generan con una periodicidad anual. A partir del año 2005 se van introduciendo módulos adicionales en la componente transversal sobre diferentes temas de especial interés.

Especificidades técnicas

La población de referencia son los hogares y todas las personas mayores de 16 años que se encuentren residiendo en un hogar dentro del territorio de los estados miembros en el momento de realizarse la encuesta. Quedan excluidas las personas que viven en hogares colectivos (residencias para la tercera edad, por ejemplo) o en algunos territorios que no son incorporados por sus propios países en la base de datos (territorios franceses fuera de sus fronteras europeas, por ejemplo). Los datos son recogidos por cada país mediante una institución que, en España, es el Instituto Nacional de Estadística (INE).

Estrictamente, la población objeto de investigación (población objetivo) son las personas miembros de hogares privados que residen en viviendas familiares principales.Aunque las personas de todas las edades forman parte de la población objetivo no todas las personas son encuestadas exhaustivamente, ya que sólo son seleccionables en este caso, los miembros del hogar con 16 o más años el 31 de diciembre del año anterior a la fecha de la entrevista.

La base de datos proporciona de microdatos transversales y longitudinales con información personalizada sobre ingresos, educación, salud, ocupación, entre otros, que permite conocer las condiciones en que viven los encuestados y las posibles situaciones de pobreza y exclusión social. En el caso de los ingresos, es de especial interés para este trabajo la información relacionada con las transferencias dinerarias recibidas por el individuo en concepto de becas y, en el caso de las variables educativas, aquellas que permiten seguir su evolución dentro del sistema educativo.

Finalmente, y en cuanto a las metodologías de imputación empleadas, la Encuesta de Condiciones de Vida imputa las variables referidas a ingresos y gastos de los individuos de cara a conformar los ingresos y gastos del hogar. La imputación sólo se realiza en caso de valores perdidos de forma parcial. Para el caso de la base longitudinal, si existe información (de un ingreso o gasto individual en concreto) del año anterior, se multiplica este valor por el IPC para calcular el valor no observado. En caso que no exista este valor, se aplica una regresión secuencial multivariante (procedimiento estocástico que considera elementos aleatorios).

Caso español

En el caso de España, la encuesta es de tipo "panel rotante", es decir, al ser un panel se investiga a las mismas unidades a lo largo de los años, pero a diferencia del PHOGUE en que las unidades panel eran fijas a lo largo de los ocho años de duración del estudio, en

la ECV las unidades panel se encuestan durante cuatro años y luego son reemplazadas.

La selección de la muestra se realiza a partir del Padrón Municipal de habitantes de 2003 (INE) y se compone de 4 submuestras panel, de forma que cada año una de ellas se sustituye por una nueva. Para la selección de cada submuestra se sigue un diseño bietápico con estratificación de las unidades de primera etapa. La primera etapa la forman las secciones censales y la segunda etapa las viviendas familiares principales. Dentro de ellas no se realiza submuestreo alguno, investigándose a todos los hogares que tienen su residencia habitual en las mismas.

3.2 Selección muestral

Para el análisis empírico se emplean los datos correspondientes a la Encuesta de Condiciones de Vida (ECV), elaborada por EUROSTAT con datos longitudinales para el período 2004-2006, publicada en 2009. Los datos disponibles hacen referencia a los países de la Unión Europea y en el caso español, la muestra comprende 58.740 individuos. Para el estudio de impacto de las becas y ayudas al estudio en el logro educativo de los estudiantes, la variable dependiente hace referencia al nivel educativo que posee la persona a los 19 años (véase Cuadro 1).

Si bien la edad teórica para finalizar el nivel secundario post-obligatorio son los 18 años, se ha optado por seleccionar un año más para evitar encontrar individuos con 18 años que aún no tengan este nivel educativo alcanzado sólo porque la encuesta se ha realizado antes de finalizar su curso lectivo. Como variables independientes se consideran diferentes variables relacionadas con el individuo, sus progenitores y su hogar. Finalmente, un tema siempre presente en este tipo de análisis es la posible existencia de endogeneidad entre alguna de las variables explicativas y la explicada. Específicamente en este caso, conviene destacar que la supuesta endogeneidad existente entre las becas recibidas y el éxito escolar queda limitada debido a dos aspectos relacionados directamente con el diseño de las mismas. En primer lugar, a que las ayudas al estudio sólo siguen criterios económicos y, por tanto, no tienen en cuenta el rendimiento académico. En segundo lugar,que si bien las becas para su renovación sí tienen en cuenta ambos criterios, la endogeneidad sólo existiría en la renovación y no en su primera otorgación donde sólo se tienen en cuenta criterios económicos. En este sentido, conviene recordar que el objetivo principal del trabajo no es estimar el impacto causal de las becas en la consecución del título de Bachillerato o del Ciclo Formativo Medio, sino analizar las diferencias encontradas en los efectos estimados según distintas metodologías de imputación.

A partir del total de observaciones válidas para la variable dependiente (Post_Oblig_con_19), se genera una sub-base de datos con 783 observaciones de individuos de 19 años durante el período analizado que contiene valores perdidos para algunas de las variables independientes, básicamente localizadas en aquellas que hacen referencia a los progenitores (véase cuadro 2), si bien en ningún caso superan el umbral del 20 %. La existencia de

Cuadro 1. Variables utilizadas en el análisis empírico

Tipo de variable	Variable utilizada	Descripción
Individuo	**Nivel educativo a los 19 años**	Variable que indica el nivel educativo (ISCED-97) a los 19 años en 2006 (b). Se estructura como una dummy = 1 si la persona tiene un nivel educativo igual o superior al de secundaria post-obligatoria (**Post-oblig_con_19**).
	Beca	Dummy Becario. Percepción de una o más becas/ayudas al estudio, a título personal, en el nivel secundario post-obligatorio durante el período 2004-2005 (Beca).
	Género	Dummy género. Toma el valor 1 si el individuo es mujer (**Mujer**).
	"Efecto calendario"	Dummy mes de nacimiento. Toma el valor 1 si el individuo nació en el último trimestre del año (Último_Tri).
	Estado de salud	Dummy enfermedad crónica. Toma el valor 1 si el individuo padece una enfermedad o incapacidad crónica (**Enf_Crónica**).
	Orden entre hermanos	Variable que hace referencia al orden que ocupa el individuo en relación con sus hermanos (Ejemplo: el hermano mayor tiene un número de orden igual a 1) (**Posición**).
Progenitores	**Nivel educativo padre**	Máxima educación lograda por el padre (ISCED-97) (**Educ_Padre**).
	Nivel educativo madre	Máxima educación lograda por la madre (ISCED-97) (**Educ_Madre**).
	Actividad padre	Dummy activo. Toma el valor 1 si el individuo se encuentra activo (**Activo_Padre**).
	Actividad madre	Dummy activo. Toma el valor 1 si el individuo se encuentra activo (**Activo_Madre**).
Hogar	**Número de hermanos**	Variable que indica la cantidad de hermanos existentes en el hogar (**Nro_Hermanos**).
	Nivel de ingresos (I)	Quintil de ingresos disponibles equivalentes (**Quintil**) (c).
	Nivel de ingresos (II)	Dummy dificultades económicas. Toma el valor 1 si el hogar declara tener problemas para asumir los gastos habituales del mes (**Dificultad_Econ**).
	Régimen de la vivienda	Dummy propietario de la vivienda. Toma el valor 1 si los habitantes del hogar son propietarios de la misma (**Vivienda_Prop**).
	Problemas estructurales	Dummy problemas estructurales en la vivienda. Toma el valor 1 si existen problemas estructurales en la vivienda (**Prob_Estructural**).
	Dimensiones del hogar	Dummy si el hogar posee más de cuatro ambientes (**Mas_4_Dep**)(d).
	Grado de urbanización	Dummy si el individuo vive en una zona de baja o media urbanización. (**Baja_Media_Urb**)(e).

(a) Para la construcción de las variables (excluidas Post-oblig_con_19 y Becas) se ha empleado la información disponible en los tres años (2004-2006).
(b) ISCED-97: International Standard Classification of Education.
(c) El ingreso equivalente se calcula teniendo en cuenta el ingreso disponible anual del hogar y el tamaño equivalente del hogar, el cual pondera de manera diferencial a los adultos y a los menores del hogar (escala: OCDE modificada). En el caso del ingreso disponible anual del hogar, el INE ha aplicado metodologías previas de imputación en la conformación de esta variable agregada para las situaciones de no respuesta parcial de algún miembro del hogar.
(d) La variable utilizada incorpora habitaciones, salas para comer, salas de estar y altillos o sótanos habitables. Quedan excluidos el baño, la despensa, los pasillos y la cocina en caso que sólo se emplee para cocinar.
(e) La variable original cuenta con tres valores relativos al grado de urbanización (alto, medio y bajo). Interesa comparar un grado de urbanización alto con el resto de categorías.
Fuente: Elaboración propia a partir de microdatos de EUSILC LONGITUDINAL UDB 2006 – versión 2 – de Marzo 2009.

casos de respuesta parcial obliga a la aplicación de alguna medida correctiva previo a la estimación.

Cuadro 2. Valores perdidos en la base de datos original

	Nº obs. Válidas	**Nº obs. missings**	**% missings**
Variable dependiente			
Post_Oblig_con_19	783	0	0
Variables independientes			
Individual			
Beca	783	0	0
Mujer	783	0	0
Último_Tri	779	4	0,51
Enf_Crónica	783	0	0
Posición	752	31	3,96
Progenitores			
Educ_Padre	662	121	15,45
Educ_Madre	714	69	8,81
Activo_Padre	672	111	14,18
Activo_Madre	748	35	4,47
Hogar			
Nro_Hermanos	775	8	1,02
Quintil	783	0	0
Dificultad_Econ	783	0	0
Vivienda_Prop	783	0	0
Prob_Estructural	783	0	0
Mas_4_Dep	783	0	0
Baja_Media_Urb	783	0	0

Fuente: Elaboración propia a partir de microdatos de EUSILC LONGITUDINAL UDB 2006 – versión 2 – de Marzo 2009.

3.3 Aplicación de diferentes metodologíaspara el tratamiento de datos faltantes

3.3.1. Análisis del patrón de comportamiento de los datos perdidos

Como paso previo al proceso de imputación, se debe comprobar la aleatoriedad (sea parcial o total) de los valores ausentes. En este caso, se aplica la prueba de las correlaciones dicotomizadas. Para realizar la prueba, a cada variable incluida en el análisis se la dicotomiza asignándole el valor cero a los valores ausentes y el valor uno a los valores válidos. Seguidamente, se halla la matriz de correlaciones con sus respectivos constrastes de significatividad para cada coeficiente. La correlación indica el grado de asociación entre los valores perdidos de cada par de variables. Bajas correlaciones indican aleatoriedad (Perez, 2004). La prueba realizada con las variables incluidas en el modelo indica bajos niveles de correlación, por tanto se puede considerar que los valores perdidos se estarían

generando aleatoriamente[7]. Esta conclusión permite aplicar algunas de las metodologías de imputación desarrolladas en la literatura empírica.

3.3.2. Metodologías de imputación

En primer lugar, se emplea la metodología tradicionalmente utilizada por la literatura empírica: la eliminación de todas las variables con observaciones faltantes (list wise deletion). Es importante reiterar que, esta metodología y la imputación múltiple planteada a continuación, han sido aplicadas sobre una base de datos restringida a los individuos de 19 años en 2006. Otra opción metodológica hubiera sido imputar con toda la base de datos y luego generar la subpoblación de interés. Se ha preferido generar los valores imputados a partir, directamente, de la población de interés debido a que el estudio empírico centra su atención en un problema concreto que afecta a la población de 19 años y es su posibilidad o no de finalizar el nivel secundario post-obligatorio en España. Se entiende que es una tipología de persona particular y, por ello, se selecciona la información más cercana y con mayor valor informativo para generar la muestra con la cual se realizará la imputación y la estimación posterior. Como elemento adicional para fundamentar la decisión metodológica escogida, se ha tenido en cuenta que dentro del resto de la base de datos también existían observaciones faltantes.

En segundo lugar, se aplica la imputación múltiple para substituir los datos faltantes. Siguiendo la literatura, se aplica una estimación múltiple a partir del algoritmo MICE, sistematizado para el programa STATA (versión 11) mediante el comando mi impute chained[8] que emplea ecuaciones encadenadas. La imputación aquí aplicada genera valores posibles a partir de una serie de modelos univariantes en los cuales una variable única es imputada en base a un grupo de variables. La inferencia se realiza a partir de una distribución t de Student en lugar de una normal (StataCorp, 2017 – pp.46) y va en la línea propuesta por Reiter y Raghunathan (2007) para evitar sesgos en la estimación.

En este caso, y siguiendo larecomendación de Rubin (1996) y Acock (2005), se han utilizado todas las variables disponibles en el modelo para estimar los datos no observados a partir de dos aproximaciones empíricas diferentes (logit y logit ordenado) según las características particulares de cada variable (véase cuadro 3). Para cada observación perdida se generan 20 observaciones imputadas (m=20) a partir de la estimación escogida, teniendo en cuenta que el porcentaje máximo de observaciones no observadas es de 15,45 % para el caso de la educación del padre.

[7]Previo al cálculo de los coeficientes de correlación se comprobó la ausencia de valores atípicos a partir de gráficos de caja. Todo este análisis previo se encuentra disponible para los lectores que así lo soliciten.

[8]El comando permite aplicar otras aproximaciones. Por ejemplo: mi impute mvn supone una distribución normal de todas las variables a imputar y se puede estimar adicionalmente como robustez de la estimación propuesta.

Cuadro3. Aproximaciones empíricas utilizadas para la imputación

Variable a imputar	Característica	Aproximación empírica
Último_Tri	Dummy (0-1)	*Logit*
Activo_Padre	Dummy (0-1)	
Activo_Madre	Dummy (0-1)	
Posición	Discreta (1-3)	*Logit* ordenado
Educ_Padre	Discreta (1-5)	
Educ_Madre	Discreta (1-5)	
Nro_Hermanos	Discreta (0-3)	

3.4 Base de datos obtenida mediante las diferentes metodologías de tratamiento

Para analizar descriptivamente las bases de datos obtenidas, se presentan los valores de la media y la desviación estándar para todas las variables utilizadas. En el caso de la no imputación, los valores surgen a partir de las observaciones finalmente empleadas en la estimación (en este caso, 616 observaciones) y, en el caso de la imputación múltiple, el valor publicado refleja el promedio de las 20 bases generadas.

En el cuadro 4 se observan los resultados para la variable dependiente y las variables independientes referidas al ámbito del individuo. Teniendo en cuenta las diferencias estadísticamente significativas, el único caso detectado corresponde a la variable dependiente (nivel educativo de los alumnos).En este caso, la no imputación incrementa la media en un 7,47 %. Asimismo, en el caso de la no imputación (list wise deletion) se observa que el número de observaciones empleadas (616) implica una pérdida de 167 observaciones, un 21,32 % de la muestra original. Además, conviene recalcar que, si bien la imputación múltiple se aplica en todas las variables con valores no observados, y por tanto hace variar el valor promedio y la desviación estándar, en ningún caso se obtiene una diferencia significativa en relación a la muestra original. Finalmente, no existen diferencias destacables entre las diferentes metodologías de imputación en relación con la desviación estándar obtenidas para las diferentes variables.

En el cuadro 5 finaliza el análisis descriptivo con las variables independientes referidas a los progenitores y al hogar donde habita el individuo analizado. En cuanto a la media resultante, sólo se observan diferencias consideradas significativas en la base no imputada para tres variables referidas al hogar (dificultad económica, vivienda en propiedad y más de cuatro dependencias). En cuanto a la desviación estándar se observa que, claramente, la no imputación confiere desvíos importantes (aunque no significativos) en cinco de las once variables analizadas y con diferencias que superan el 8 % en dos casos (tasa de actividad del padre y quintil de renta del hogar) mientras que la imputación múltiple sólo genera una diferencia destacada (actividad del padre) de entre todas las características tenidas en cuenta.

Cuadro 4. Análisis descriptivo (I)

	Original	No imputación	I. Múltiple (*)	Original	No imputación	I. Múltiple (*)
	Media			Desviación Estándar		
Var. dependiente						
Post_oblig_con_19 (No missing)	0,4546	**0,4886 (+7,47 %)**	0,4546	0,4982	**0,5002 (+0,40 %)**	0,4982
Var. independientes						
Individual						
Beca (No missing)	0,1915	**0,2110 (+10,18 %)**	0,1915	0,3938	**0,4084 (+3,71 %)**	0,3938
Mujer (No missing)	0,4802	**0,4756 (-0,95 %)**	0,4802	0,4999	**0,4998 (-0,02 %)**	0,4999
Último_Tri (0,51 % missing)	0,2593	**0,2630 (+1,42 %)**	**0,2597 (+0,15 %)**	0,4385	**0,4406 (+0,48 %)**	**0,4388 (+0,07 %)**
Enf_Crónica (No missing)	0,074	**0,0730 (-1,35 %)**	0,074	0,262	**0,2604 (-0,61 %)**	0,262
Posición (3,96 % missing)	1,5771	**1,5568 (-1,29 %)**	**1,5718 (-0,33 %)**	0,6995	**0,6867 (-1,83 %)**	**0,6986 (-0,13 %)**
N	783	**616**	783	783	**616**	783

En negrita los valores que presentan desviaciones respecto al valor de la base original. (*) El valor imputado surge como el promedio de los valores obtenidos en las 20 bases de datos completas generadas por el proceso de imputación. 2009.
Para conocer el número de observaciones para cada variable en la base original, véase cuadro 2.
Fuente: Elaboración propia a partir de microdatos de EUSILC LONGITUDINAL UDB 2006 – versión 2 – de Marzo 2009.

Como conclusión al análisis descriptivo precedente resulta relevante explicitar que la base no imputada presenta los mayores problemas de representatividad en términos de diferencias tanto en la media como en la desviación estándar. A su vez, la imputación múltiple permite incorporar el máximo de información sin perder representatividad de la base imputada respecto a la original, manteniendo el grado de variabilidad de todas las variables involucradas en la estimación. Esta última característica se debe a la capacidad para incorporar la incertidumbre mediante la estimación de varios valores para cada dato ausente en la muestra.

4 Comparación de las estimaciones obtenidas a partir de diferentes bases de datos: imputadas y no imputadas

En este apartado se comparan los resultados obtenidos en una regresión logística binomial aplicada a las dos bases de datos resultantes: la base no imputaday la base imputada mediante un proceso de imputación múltiple. En el caso de esta segunda opción, el valor medio, tanto de los coeficientes como del error estándar, se calcula aplicando las reglas de Rubin (1987) para combinar estimaciones. Las mismas indican que, para cada análisis, uno primero debe calcular los estimadores y sus respectivos errores estándar para cada base imputada. Siguiendo la explicación de Schafer (1997), suponemos que $\hat{Q}_j$ es un estimador de un escalar de interés (un coeficiente de un regresión, por ejemplo) obtenido a partir de las diferentes bases de datos "j" ($j = 1, 2, ..., m$) y U_j es el error estándar asociado. El estimador total es la media de cada una de las estimaciones ($\bar{Q} = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}_j$).

Cuadro 5. Análisis descriptivo (II)

	Original	No imputación	I. Múltiple (*)	Original	No imputación	I. Múltiple (*)
	Media			Desviación Estándar		
Progenitores						
Educ_Padre (15,45 % missing)	2,4577	**2,4496 (-0,33 %)**	**2,4669 (+0,37 %)**	1,4954	**1,4763 (-1,27 %)**	**1,5012 (+0,39 %)**
Educ_Madre (8,81 % missing)	2,2913	**2,2938 (+0,11 %)**	**2,2807 (-0,46 %)**	1,3797	**1,3792 (-0,04 %)**	**1,3818 (+0,15 %)**
Activo_Padre (14,18 % missing)	0,9196	**0,9334 (+1,50 %)**	**0,8900 (-3,22 %)**	0,272	**0,2494 (-8,31 %)**	**0,3128 (+15,00 %)**
Activo_Madre (4,47 % missing)	0,512	**0,5081 (-0,76 %)**	**0,5162 (+0,82 %)**	0,5001	**0,5003 (+0,04 %)**	**0,5000 (-0,02 %)**
Hogar						
Nro_Hermanos (1,02 % missing)	1,3303	**1,2824 (-3,60 %)**	**1,3208 (-0,71 %)**	0,8771	**0,8173 (-6,82 %)**	**0,8794 (+0,26 %)**
Quintil (No missing)	2,6321	**2,7029 (+2,69 %)**	2,6321	1,3682	**1,5654 (+14,41 %)**	1,3682
Dificultad_Econ (No missing)	0,6398	**0,6055 (-5,36 %)**	0,6398	0,4803	**0,4891 (+1,83 %)**	0,4803
Vivienda_Prop (No missing)	0,8352	**0,8620 (+3,21 %)**	0,8352	0,3712	**0,3452 (-2,60 %)**	0,3712
Prob_Estructural (No missing)	0,1775	**0,1558 (-12,22 %)**	0,1775	0,3823	**0,3630 (-5,04 %)**	0,3823
Mas_4_Dep (No missing)	0,7139	**0,7458 (+4,47 %)**	0,7139	0,4522	**0,4305 (-4,80 %)**	0,4522
Baja_Media_Urb (No missing)	0,5441	**0,5698 (+4,72 %)**	0,5441	0,4984	**0,4955 (-0,58 %)**	0,4984
N	783	616	783	783	616	783

En negrita los valores que presentan desviaciones respecto al valor de la base original. (*) El valor imputado surge como el promedio de los valores obtenidos en las 20 bases de datos completas generadas por el proceso de imputación.
Para conocer el número de observaciones para cada variable en la base original, véase cuadro 2.
Fuente: Elaboración propia a partir de microdatos de EUSILC LONGITUDINAL UDB 2006 – versión 2 – de Marzo 2009.

Para calcular el error estándar total, primero se debe calcular la varianza intra-imputación ($\bar{U} = \frac{1}{m}\sum_{j=1}^{m}\hat{U}_j$) y la varianza entre imputaciones ($B = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{Q}_j - \bar{Q}^2)$). Por tanto, la varianza total se calcula como: $T = \bar{U} + (1 + \frac{1}{m}) * B$, donde el error estándar es su raíz cuadrada.

El objetivo del ejercicio es determinar la importancia de las diferentes variables introducidas en el logro educativo de los alumnos.Específicamente, se estima el siguiente modelo logístico binomial:

$$Pr(y = 1 \vee x) = F(x\beta),$$

donde Y indica el nivel educativo del individuo a los 19 años (y un valor igual a uno indicaría que ha finalizado con éxito el nivel secundario post-obligatorio o superior), F es una función de densidad logística y "x" incluye la totalidad de variables independientes incorporadas al modelo (Long y Freese, 2006).

El análisis comparado de los resultados obtenidos muestra dos aspectos relevantes a

destacar (véase cuadro 6). En primer lugar, y en cuanto a la significatividad de las variables, se observan comportamientos similares para las diferentes metodologías de imputación utilizadas en el caso de la variable referida a las becas, el género, las relacionadas con la educación de los progenitores y algunas variables del hogar (vivienda en propiedad, problemas estructurales y grado de urbanización). Por el contrario, existen tres variables donde la significatividad varía. En primer término, la variable referida a la posición del individuo en relación con sus hermanos sólo es significativa (si bien, al 80 %) en el caso de la base no imputada. En segundo término, la variable quintil de renta y la variable que informa sobre las dificultades económicas del hogar sólo resultan significativas para la base imputada.

En segundo lugar, el análisis se centra en los valores de los coeficientes. Para ello, los coeficientes estadísticamente significativos se han comparado a partir de un test de diferencias de coeficientes propuesto por McDowell (2005). En primer lugar, en el caso de las becas, la variable género y la educación del padre existe una diferencia de coeficientes entre los resultados obtenidos a partir de la base imputada y aquellos resultantes de la base no imputada.

En el caso de las becas y la educación del padre los resultados reflejan un mayor impacto de las mismas en la obtención del secundario post-obligatorio en el caso de la base imputada. En sentido opuesto, la base listwise otorga un mayor impacto de la variable mujer. En segundo lugar, los coeficientes referidos al quintil de renta y a las dificultades económicas sólo serían tenidos en cuenta en el caso de la imputación múltiple. Finalmente, para la educación de la madre y la variable referente a los problemas estructurales del hogar existe un mayor impacto otorgado por la regresión empleando la base no imputada en relación con la base imputada a través de la imputación múltiple.

Por último, y con un objetivo marcadamente descriptivo, se calculan diferentes indicadores de la bondad del ajuste de las regresiones anteriormente comentadas(véase cuadro 7). Antes de iniciar este análisis conviene aclarar que los indicadores empleados a continuación, al ser calculados en base a muestras distintas, no sirven como elementos definitivos de decisión en relación a la calidad de cada una de ellas[9].

A nivel general, si bien en ambos modelos se rechaza la hipótesis nula de que todos los términos incluidos en el modelo (excepto la constante) son cero para el resto de indicadores calculados (ratio de máxima verosimilitud -LR-; R2 de MacFadden; porcentaje de predicciones correctas y los criterios de información Akaike y Bayesiano) se observa un mejor comportamiento en el caso de la base imputada.

[9]En caso de que el objetivo del trabajo hubiese sido utilizar las medidas de bondad de ajuste como evidencia a favor/en contra de los métodos de imputación, se debería haber realizado un análisis específico de descomposición de la base según tenga valores perdidos o no; luego imputar los datos que faltan; estimar el modelo con las distintas muestras y, por último, predecir la variable dependiente y observar cuál tiene mayor capacidad predictiva. De esta manera, los resultados de la predicción extra-muestral darían una medida de las diferencias –si se encuentran– entre la calidad de los distintos modelos.

Cuadro 6. Estimación en base a las diferentes metodologías de imputación

		No imputación	Imputaciónmúltiple (*)
Beca	Coef	**1,105**	**1,194**
	Error	**0,227**	**0,208**
	T-ratio	**4,87**	**5,73**
Mujer	Coef	**0,305**	**0,242**
	Error	**0,179**	**0,161**
	T-ratio	**1,7**	**1,51**
Último_Tri	Coef	0,195	0,106
	Error	0,203	0,182
	T-ratio	0,96	0,58
Enf_Crónica	Coef	0,013	-0,239
	Error	0,343	0,31
	T-ratio	0,04	-0,77
Posición	Coef	**-0,207**	-0,116
	Error	**0,147**	0,13
	T-ratio	**-1,4**	-0,99
Educ_Padre	Coef	**0,246**	**0,33**
	Error	**0,076**	**0,065**
	T-ratio	**3,23**	**3,35**
Educ_Madre	Coef	**0,152**	**0,129**
	Error	**0,084**	**0,074**
	T-ratio	**1,81**	**1,74**
Activo_Padre	Coef	0,028	0,222
	Error	0,373	0,274
	T-ratio	0,08	0,75
Activo_Madre	Coef	0,146	-0,025
	Error	0,186	0,194
	T-ratio	0,79	-0,17
Nro_Hermanos	Coef	-0,086	-0,114
	Error	0,125	0,106
	T-ratio	-0,69	-1,08
Quintil	Coef	0,092	**0,121**
	Error	0,078	**0,067**
	T-ratio	1,18	**1,77**
Dificultad_Econ	Coef	-0,187	**-0,252**
	Error	0,198	**0,181**
	T-ratio	-0,94	**-1,4**
Vivienda_Prop	Coef	**0,797**	**0,742**
	Error	**0,284**	**0,239**
	T-ratio	**2,8**	**3,11**
Prob_Estructural	Coef	**-0,358**	**-0,303**
	Error	**0,259**	**0,222**
	T-ratio	**-1,38**	**-1,36**
Mas_4_Dep	Coef	-0,025	0,198
	Error	0,215	0,183
	T-ratio	-0,01	1,08
Baja_Media_Urb	Coef	**-0,497**	**-0,456**
	Error	**0,188**	**0,166**
	T-ratio	**-2,64**	**-2,75**
Constante	Coef	**-1,583**	**-1,854**
	Error	**0,596**	**0,494**
	T-ratio	**-2,65**	**-3,75**
N		616	783

En negrita: estimaciones significativas (>=80 %).(*) El valor imputado surge como el promedio de los valores obtenidos en las 20 bases de datos completas generadas por el proceso de imputación.

Cuadro 7. Estimación en base a las diferentes metodologías de imputación: análisis de la bondad del ajuste

	Log-Likelihood	Prob>LR	LR(16)	McFadden	Pred. Correctas (en %)	Akaike	Bayesiano
No imputación	-368,216	0	117,21	0,137	66,2	1,251	-3111,094
I. Múltiple (*)	-463,921	0	**151,18**	**0,14**	**68,59**	**1,229**	**-4176,117**

(∗) El valor imputado surge como el promedio de los valores obtenidos en las 20 bases de datos completas generadas por el proceso de imputación.

5 Conclusiones

El presente trabajo se planteó como objetivo comprobar las diferencias existentes entre la no imputación de las observaciones con datos faltantes y la imputación múltiple. Para ello se realizó un análisis comparativo entre las distintas bases de datos generadas a partir de un análisis descriptivoy de regresión logística.

Para analizar descriptivamente las bases de datos obtenidas, se presentan los valores de la media y la desviación estándar para todas las variables utilizadas. Los principales resultados obtenidos reflejan diferencias significativas de medias en el caso de la base no imputada – con respecto a la base original - para cuatro variables (nivel educativo de los alumnos; dificultad económica, vivienda en propiedad y más de cuatro dependencias).Asimismo, se comprueba que la imputación múltiple permite incorporar el máximo de información sin perder representatividad de la base imputada respecto a la original, manteniendo el grado de variabilidad de todas las variables involucradas en la estimación. Finalmente, en el caso de la no imputación (list wise deletion) se observa que el número de observaciones empleadas (616) implica una pérdida de 167 observaciones, un 21,32 % de la muestra original.

Los resultados obtenidos en la regresión logística muestran que la imputación múltiple es la más adecuada en términos de reducción del error estándar. Asimismo, los cambios de significatividad y de valor de los coeficientes denotan las diferencias que se generan a causa del empleo de diferentes bases de datos derivadas de una misma base original, reflexión que va en la línea de lo encontrado por Barceló (2008) y plantea un toque de atención para muchos investigadores que aún no consideran este potencial problema dentro de sus objetivos metodológicos de investigación y que podría derivar en una mala especificación del modelo.

La principal conclusión que se deriva del trabajo es que resulta recomendable (si es factible) imputar aquellas bases de datos con observaciones faltantes. Asimismo, y sobre todo para bases de datos no muy amplias (menos de 1000 observaciones, por ejemplo), la imputación no sólo permite una mayor precisión en los coeficientes, sino que también una menor probabilidad de realizar interpretaciones erróneas fruto de una base de datos

estimada que podría no ser representativa de la muestra original.

Abstract

The empirical literature in the field of applied economics frequently bases its estimates on databases that contain only valid observations for all the variables involved in the analysis. In the majority of the cases, this action results in a loss of information that can cause two problems: a less effective estimation and an increase in the probability of model misspecification. The objective of this work is to test the importance of the imputation of missing data in minimizing both problems. For this, I applied multiple imputationas a way for obtaining more robust estimates than those generated by simple listwise deletion. The results show that the imputation improves the effectiveness of the estimation in terms of smaller standard errors and; regarding the specificity of the model, there are important changes in the significance levels of some variables depending on the database used for performing the estimations.

Keywords: missing values, efficiency, multiple imputation

JEL: C13; C81

Referencias

ALLISON, P. D. Multiple imputation for missing data: A cautionary tale. *Sociological methods & research*, Sage Publications, Inc., v. 28, n. 3, p. 301–309, 2000.

ALLISON, P. D. *Missing data*. [S.l.]: Sage publications, 2001. v. 136.

AMBLER, G.; OMAR, R. Z.; ROYSTON, P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Statistical methods in medical research*, Sage Publications Sage UK: London, England, v. 16, n. 3, p. 277–298, 2007.

BARCELÓ, C. The impact of alternative imputation methods on the measurement of income and wealth: Evidence from the spanish survey of household finances. 2008.

BARNARD, J.; MENG, X.-L. Applications of multiple imputation in medical studies: from aids to nhanes. *Statistical methods in medical research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 8, n. 1, p. 17–36, 1999.

BODNER, T. E. What improves with increased missing data imputations? *Structural Equation Modeling*, Taylor & Francis, v. 15, n. 4, p. 651–675, 2008.

BOVER, O.; CORONADO, E.; VELILLA, P. The spanish survey of household finances (eff): description and methods of the 2002 wave. *Occasional Paper*, n. 0409, 2004.

BOVER, O.; CORONADO, E.; VELILLA, P. The spanish survey of household finances (eff): description and methods of the 2005 wave. *Occasional Paper*, n. 0803, 2008.

BUUREN, S. V.; BOSHUIZEN, H. C.; KNOOK, D. L. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, Wiley Online Library, v. 18, n. 6, p. 681–694, 1999.

CARPENTER, J. R.; KENWARD, M. G.; WHITE, I. R. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical methods in medical research*, Sage Publications Sage UK: London, England, v. 16, n. 3, p. 259–275, 2007.

CHRISTELIS, D. Imputation of missing data in waves 1 and 2 of share. 2011.

GÓMEZ, J.; PALAREA, J. Inferencia basada en imputación múltiple en problemas con información incompleta. In: *Comunicación presentada en la IX Conferencia Española de Biometría*. [S.l.: s.n.], 2003.

GRAHAM, J. W. Missing data analysis: Making it work in the real world. *Annual review of psychology*, Annual Reviews, v. 60, p. 549–576, 2009.

GUERRA, J. B.; GALLESTEY, J. B. Imputacion multiple en variables categoricas usando data augmentation y arboles de clasificacion. *Investigación Operacional*, v. 31, n. 2, p. 133–139, 2014.

HIPPEL, P. T. V. 8. how to impute interactions, squares, and other transformed variables. *Sociological methodology*, SAGE Publications Sage CA: Los Angeles, CA, v. 39, n. 1, p. 265–291, 2009.

HORTON, N. J.; LIPSITZ, S. R. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician*, Taylor & Francis, v. 55, n. 3, p. 244–254, 2001.

KENNICKELL, A. B. Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation. In: *Proceedings of the Survey Research Methods Section of the American Statistical Association*. [S.l.: s.n.], 1991. v. 1, n. 10.

KENWARD, M. G.; CARPENTER, J. Multiple imputation: current perspectives. *Statistical methods in medical research*, Sage Publications Sage UK: London, England, v. 16, n. 3, p. 199–218, 2007.

LITTLE, R. J.; RUBIN, D. B. *Statistical analysis with missing data*. [S.l.]: John Wiley & Sons, 2014. v. 333.

LONG, J. S.; FREESE, J. *qRegression Models for Categorical Dependent Variables Using Stata, r Third Edition*. [S.l.]: Stata press, 2014.

MADDALA, G. *Introducción a la Econometría. Trad. por J. Jolly*. [S.l.]: Prentice-Hall Hispanoamericana, SA Segunda Edición. México DF, México, 1996.

MCDOWELL, A. How do you test equality of regression coefficients that are generated from two different regressions, estimated on two different samples? 2015. Consulta: 10/01/2013. Disponível em: <https://www.stata.com/support/faqs/statistics/test-equality-of-coefficients/>.

MUSIL, C. M. et al. A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 24, n. 7, p. 815–829, 2002.

OUTHWAITE, W.; TURNER, S. *The SAGE handbook of social science methodology*. [S.l.]: Sage, 2007.

PÉREZ, C. Técnicas de análisis multivariante de datos. aplicaciones con spss. *Capítulo*, v. 5, p. 155–191, 2004.

REITER, J. P.; RAGHUNATHAN, T. E. The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, Taylor & Francis, v. 102, n. 480, p. 1462–1471, 2007.

ROSAS, J. F. M.; VERDEJO, E. A. et al. Métodos de imputación para el tratamiento de datos faltantes: aplicación mediante r/splus. Universidad Pablo de Olavide, 2009.

ROYSTON, P. et al. Multiple imputation of missing values. *Stata journal*, Stata Press, v. 4, n. 3, p. 227–41, 2004.

ROYSTON, P. et al. Multiple imputation of missing values: update. *Stata Journal*, STATA PRESS 4905 LAKEWAY PARKWAY, COLLEGE STATION, TX 77845 USA, v. 5, n. 2, p. 188, 2005.

ROYSTON, P. et al. Multiple imputation of missing values: update of ice. *Stata Journal*, STATA PRESS, v. 5, n. 4, p. 527, 2005.

ROYSTON, P.; WHITE, I. R. et al. Multiple imputation by chained equations (mice): implementation in stata. *J Stat Softw*, v. 45, n. 4, p. 1–20, 2011.

RUBIN, D. B. Inference and missing data. *Biometrika*, Oxford University Press, v. 63, n. 3, p. 581–592, 1976.

RUBIN, D. B. Multiple imputation after 18+ years. *Journal of the American statistical Association*, Taylor & Francis Group, v. 91, n. 434, p. 473–489, 1996.

RUBIN, D. B. *Multiple imputation for nonresponse in surveys*. [S.l.]: John Wiley & Sons, 2004. v. 81.

SCHAFER, J. L. *Analysis of incomplete multivariate data*. [S.l.]: Chapman and Hall/CRC, 1997.

SCHAFER, J. L. Multiple imputation: a primer. *Statistical methods in medical research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 8, n. 1, p. 3–15, 1999.

STATACORP. Stata: Release 15. statistical software. College Station, TX: StataCorp LLC, 2017. Disponível em: <https://www.stata.com/manuals/mi.pdf>.

STERNE, J. A. et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, British Medical Journal Publishing Group, v. 338, p. b2393, 2009.

THRUSFIELD, M.; CHRISTLEY, R. *Veterinary epidemiology*. [S.l.]: Wiley Online Library, 2005. v. 9600.

TODESCHINI, R. Weighted k-nearest neighbour method for the calculation of missing values. *Chemometrics and Intelligent Laboratory Systems*, Elsevier, v. 9, n. 2, p. 201–205, 1990.

WHITE, I. R.; ROYSTON, P.; WOOD, A. M. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, Wiley Online Library, v. 30, n. 4, p. 377–399, 2011.