

NOTA TÉCNICA

IMPUTACIÓN MÚLTIPLE
EN ENCUESTAS MICROECONÓMICAS*

RODRIGO ALFARO

Banco Central de Chile

MARCELO FUENZALIDA

Columbia University

In the survey analysis, the missing data problem can be managed by using Multiple Imputation (MI) methods. In this paper we show the empirical application of MI methods to the financial variables included in Chile's Social Protection Survey 2004. Based on a brief review of MI methods we conclude that Multivariate Normal one is more appropriate for our case. In addition, we consider two empirical adjustments: (1) use of the variables in their logistic versions, and (2) implementation of the method by groups of individuals. Our results show that both adjustments improves the performance of the MI method.

JEL: C11, C15

Keywords: Información Faltante (Missing Data), Imputación Múltiple, Algoritmo EM/DA

1. INTRODUCCIÓN

Las encuestas con datos microeconómicos han sido ampliamente utilizadas para generar y evaluar políticas públicas en Chile. Programas para la superación de la pobreza se han evaluado con la Encuesta de Caracterización Socioeconómica (CASEN) y el sistema de ahorro previsional con la Encuesta de Protección Social (EPS), por ejemplo. Sin embargo, en el último tiempo, esta fuente de información ha sido utilizada en el análisis macroeconómico como complemento a los datos agregados. Por ejemplo, Cox, Parrado y Ruiz-Tagle (2006) caracterizan el ciclo de vida y la estructura de endeudamiento de los hogares utilizando información de la EPS 2004, mientras que Fuenzalida y Ruiz-Tagle (2009) analizan la vulnerabilidad de los hogares utilizando la Encuesta Financiera de Hogares 2007.

* Agradecemos los comentarios de Patrick Royston, Jaime Ruiz-Tagle y un árbitro anónimo. Investigadores interesados en obtener acceso a los datos para fines exclusivamente académicos deben comunicarse directamente con los autores del estudio.

Email: ralfaro@bcentral.cl

El uso de este tipo de encuestas involucra un desafío estadístico importante que corresponde al manejo de la información faltante. La complejidad del problema depende de la cantidad de variables no reportadas por el encuestado y del proceso estocástico que genera la omisión de dicha información. Encuestas incompletas son el resultado de una serie de eventos como, por ejemplo, son la comprensión de las preguntas y la disposición a revelar información sensible por parte del entrevistado. Es posible cuantificar este efecto a través de la aplicación de encuestas pilotos, lo que conlleva a mejoras tanto en el cuestionario como en las capacitaciones de los encuestadores. Desafortunadamente, esta solución no elimina totalmente el problema de la información faltante, por lo que el investigador termina con encuestas incompletas. Dada las tasas de no respuesta en algunas preguntas aisladas, no es conveniente trabajar solo con los casos para los cuales se dispone de información. Esto último supondría que los casos incompletos son una sub-muestra aleatoria de la muestra original, lo que puede no ser válido y el análisis podría generar resultados sesgados. Del mismo modo, en análisis multivariados, trabajar solo con las observaciones para las cuales todas las variables de interés tengan información completa, puede disminuir excesivamente el tamaño muestral. Alternativamente, algunos investigadores optan por reemplazar la información faltante con valores arbitrarios, como son cero, el promedio muestral o el valor proyectado de una regresión lineal (Allison, 2001). El problema de estas técnicas es que ignoran el hecho de que dichos valores no son efectivos, generándose una reducción artificial de la varianza muestral de dicha variable.

Rubin (1987) propone que los datos omitidos sean reemplazados por múltiples realizaciones aleatorias. Este proceso se conoce como Imputación Múltiple (MI por *Multiple Imputation*) y su sustento teórico se encuentra en la estadística bayesiana, la cual utiliza la información de la muestra para realizar inferencia respecto de los parámetros. En términos simples, nuestro análisis bajo MI estará basado sobre un conjunto de bases con información completa (bases imputadas), en cada una de las cuales se ha sustituido la información faltante por realizaciones aleatorias que consideran la incertidumbre asociada al hecho que los valores imputados fueron simulados. En cada una de las bases imputadas, el investigador debe realizar sus estimaciones y combinar los resultados a través de las distintas bases de datos para obtener la inferencia deseada. Para lograr este objetivo, Rubin propone reglas simples que permiten combinar los resultados obtenidos en las distintas bases imputadas, de modo de ajustar los errores estándares de los estimadores para considerar la incertidumbre generada de las imputaciones. Por otra parte, Rubin muestra que un pequeño número de bases imputadas, por ejemplo tres o cinco, son suficientes para aproximar apropiadamente la incertidumbre asociada a la información faltante.

En este trabajo presentamos la aplicación empírica de MI a la EPS 2004, pero focalizamos nuestro interés en el ingreso laboral, los activos financieros y las deudas de los individuos entrevistados. Basados en las características de las variables financieras consideradas, utilizamos la transformación logística, mientras que para controlar por la heterogeneidad de los hogares, proponemos realizar una imputación por grupos. Nuestros resultados establecen que ambos ajustes mejoran la estimación de los parámetros de la distribución subyacente.

2. IMPUTACIÓN MÚLTIPLE

Rubin (1987) recoge el tema de la incertidumbre de los valores imputados proponiendo el método de MI, por el que, a través de un proceso estocástico se seleccionan posibles valores para la información faltante y la utilización de dichos valores recoge el componente aleatorio del dato imputado. Usualmente, estas realizaciones se obtienen a través de la caracterización de la distribución conjunta de los datos, que por lo general se asume normal. También es posible obtenerlas a través de las formas funcionales de las distribuciones condicionales. En todos los casos la validez del proceso de imputación se basa en el supuesto de que la información faltante ha sido omitida en forma completamente aleatoria, es decir que la probabilidad de no reportar la información no depende del valor de la variable.

Es importante notar que el método de MI no soluciona el problema de la información faltante, sino que lo acomoda desde una perspectiva estadística. Así, el investigador podrá contar con información completa, pero deberá manejar múltiples bases de datos donde cada una de ellas tiene un valor posible para la observación faltante. El investigador entonces deberá desarrollar su análisis en cada una de las m bases de datos completas y luego combinar los resultados a fin de obtener las conclusiones finales de su investigación. La combinación de los resultados sigue las reglas propuestas por Rubin (1987), que en caso escalar se resumen en el estimador promedio (H) y su error estándar (V), siendo:

$$(1) \quad H = \frac{1}{m} \sum_{t=1}^m Q_t$$

$$(2) \quad V = \frac{1}{m} \sum_{t=1}^m V_t^2 + \left(1 + \frac{1}{m}\right) \left[\frac{1}{m-1} \sum_{t=1}^m (Q_t - H)^2 \right]$$

donde Q_t y V_t corresponden, respectivamente, al estimador y error estándar en la base t . Notamos que la primera expresión en V es el promedio de los errores estándares al cuadrado, mientras que la expresión en corchetes es un estimador de la dispersión de los estimadores obtenidos.

El número óptimo de bases de datos (m) depende del porcentaje de información faltante. Schafer (1997) discute sobre la eficiencia que se obtiene al incrementar m , la que puede analizarse empíricamente a través del grado de información faltante. Por ejemplo, bajo un 20% de información faltante el uso de tres bases imputadas incrementa el error estándar en 3.3% relativo al caso en que se consideren infinitas bases imputadas. En la práctica, una elección comúnmente utilizada es $m=5$ (Allison, 2001), sin embargo Royston (2004) propone una regla en donde se minimice empíricamente el rango de incertidumbre de la estimación del intervalo de confianza. Sus resultados son coherentes con la sugerencia de StataCorp (2009) en donde se establece que $m=20$ es un número razonable de imputaciones.

Por otra parte, distintos métodos de imputación han sido propuestos en la literatura: Hot-Deck, Condicional Univariado, Condicional Encadenado y Normal Multivariado¹. Todos ellos son apropiados, en el sentido que establece Rubin (1987), pues consideran la incertidumbre en la estimación de los parámetros. A continuación describimos brevemente cada uno de ellos, enfatizando cuales son los problemas empíricos con que se enfrenta en el investigador al aplicarlos. Una discusión detallada de los métodos en términos teóricos puede encontrarse en Allison (2001).

2.1 Método Hot-Deck

Este método asigna valores a los datos faltantes con la información existente en la muestra de acuerdo a la celda en la que se encuentra la observación con información faltante. El procedimiento consiste en completar en cada celda las observaciones faltantes utilizando datos de la misma celda, los cuales son seleccionados de forma aleatoria. Luego de hacer el procedimiento para cada celda, se logra una base de datos completa. El proceso se repite para construir las m bases de datos completas. Debido a que los valores imputados son efectivos, las características estadísticas de la celda se preservan. Esto resulta útil cuando la variable a imputar tiene características particulares como es el caso de las variables discretas.

Investigadores del área dudan que el método sea capaz de recoger en la varianza combinada toda la incertidumbre asociada a la información faltante². En la práctica, el método no es válido si las celdas en las cuales se realiza el proceso de imputación contienen pocos datos. Por ejemplo, Alfaro y Fuenzalida (2008) muestran que para la EPS 2004 el cruce: género, nivel de escolaridad y tramos de edad puede generar celdas con menos de 20 observaciones, lo que es relativamente pequeño si se considera por ejemplo $m=10$.

2.2 Método Condicional Univariado

Este método está basado en un modelo de regresión. Las imputaciones se obtienen de ajustar los valores predichos a fin de que estos contengan la incertidumbre asociada a la estimación de los parámetros. Por ejemplo en el caso de la estimación de una regresión lineal con información faltante, los parámetros estimados a consi-

¹ Los tres métodos están implementados en Stata por diversos investigadores. Hot-Deck fue implementado por Adrian Mander y David Clayton (MRC Biostatistics Unit, Cambridge, UK). Patrick Royston (MRC Clinical Trials Unit, London, UK) implementó tanto el método Condicional Univariado como el método Condicional Encadenado, mientras que John Galati y John Carlin (Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, and University of Melbourne Department of Paediatrics) implementaron el método normal multivariado. La versión 11 de Stata incorpora varios mecanismos de imputación univariada que son válidos para una variable, así como también el método normal multivariado.

² Por ejemplo, en una discusión con el profesor Gary King (Harvard University) a mediados del 2008, él comparte su experiencia empírica en MI, destacando que los resultados obtenidos con Hot-Deck presentan fuertes reducciones de los errores estándares de las variables. Similar comentario se expone en Allison (2009).

derar son el vector de regresores y la varianza del error de la ecuación. En este caso, Rubin (1987) establece un procedimiento para generar los valores imputados, el cual puede resumirse en los siguientes pasos: (1) simular la varianza estimada del error a través del uso de la distribución asintótica de este estimador, que en este caso corresponde a una chi-cuadrado, (2) simular el vector de regresores considerando tanto la incertidumbre de dichos parámetros, a través de la distribución asintótica —que es este caso es la normal multivariada— como la asociada a la estimación de la varianza del error obtenida en el paso anterior y (3) generar el valor imputado considerando los valores simulados de los puntos anteriores.

Es importante notar que este método se basa en la existencia del valor condicional de la variable a imputar con respecto a las variables exógenas que se utilizan en el proceso de imputación. Por este motivo, es posible ajustar la forma funcional adecuadamente para reflejar la naturaleza de la variable a imputar. Por ejemplo, en el caso de una variable dicotómica se pueden utilizar los modelos de variable dependiente limitada.

2.3 Método Condicional Encadenado

Este método es una extensión del método anterior cuando hay múltiples variables con información faltante. En dicho caso Van Buuren (2006) propone que se realice el método anterior de forma secuencial encadenada, esto es, imputando primero una variable específica y luego utilizando dichos valores como verdaderos para imputar el resto de las variables. Realizando este algoritmo reiteradas veces se obtienen las distintas bases imputadas.

Este método ha sido exitoso en el área empírica (Royston, 2004) debido a que permite ajustar variadas formas funcionales para las relaciones condicionales, además la implementación computacional es relativamente económica debido a que se basa en ciclos. Dos elementos que hay que considerar de este método son: (1) un criterio que permita establecer la estabilidad del proceso secuencial y (2) que las distribuciones condicionales mantengan coherencia con el modelo de análisis. Para el primer caso es posible hacer un registro de la estabilidad de los parámetros de todas las distribuciones condicionales, mientras que el segundo punto recoge la discusión de Schafer (1997) quien establece que el modelo de imputación tiene que ser coherente con el que será ocupado por el investigador en sus análisis.

2.4 Método Normal Multivariado

Este método supone que todas las variables en el análisis tienen una distribución normal multivariada. A través de la maximización de la verosimilitud es posible recuperar los parámetros que caracterizan la distribución multivariada. Lo anterior se implementa con el algoritmo EM (*Expectation Maximization*) el cual realiza el proceso de maximización previo cálculo del valor esperado de la condición de primer orden. Esto permite que la estimación de los parámetros sea consistente bajo información faltante. Sobre dicha estimación se realiza una simulación suponiendo que las distribuciones asintóticas de cada uno de ellos

son válidas (esto es Normal para el caso del vector de medias y Wishart para la matriz de varianzas y covarianzas). Con estos parámetros simulados se generan imputaciones para las observaciones que presentan información faltante. Este algoritmo se conoce como DA (*Data Augmentation*), para el cual los valores imputados contienen explícitamente la incertidumbre asociada a la estimación del modelo de imputación.

Los fundamentos teóricos de DA están basados en el paso EM. Esto quiere decir que la apropiada convergencia de EM permite que DA sea consistente. Por este motivo EM/DA resulta ser una combinación atractiva. Sin embargo, la principal debilidad es el supuesto de normalidad conjunta, el cual es poco realista para las aplicaciones empíricas. Esto último ha sido abordado por los investigadores considerando transformaciones acordes con la naturaleza de las variables que permitan que las variables en análisis presenten un perfil más cercano a la distribución normal. De este modo, algunas variables son remplazadas por sus logaritmos o su transformación logística, mientras que en el caso de variables discretas lo habitual es redondear las cifras al entero más cercano (Allison, 2001). Para las variables dicotómicas estos procedimientos podrían no ser válidos (Allison, 2009).

3. IMPLEMENTACIÓN EN LA EPS 2004

En esta sección utilizamos la Encuesta de Protección Social 2004, para ejemplificar la implementación de MI. A fin de simplificar el problema consideraremos sólo a los jefes de hogar que reportan estar trabajando, lo que reduce la muestra a poco menos de diez mil individuos. Del mismo modo, consideraremos para nuestro análisis cuatro variables financieras: ingreso laboral, activos financieros, deuda con bancos o financieras y deuda con casas comerciales y tres variables que caracterizan al entrevistado: género, edad y años de educación. En lo que sigue presentamos la estadística descriptiva de estas variables y cuantificamos el problema de información faltante. Posteriormente discutimos sobre el método de imputación utilizado, el cual está basado en EM/DA, pero considera dos variaciones que resultan ser importantes en el trabajo empírico.

3.1. Descripción de los datos

La base de datos utilizada en este estudio corresponde a 9.648 jefes de hogar que reportan estar trabajando en la EPS 2004. Las variables a imputar son: (1) los ingresos laborales obtenidos de la ocupación principal, (2) los activos financieros, los que incluyen ahorro bancario para la compra de viviendas, ahorro en administradora de fondos para la vivienda, ahorro provisional voluntario, ahorro en cuenta 2, cuenta de ahorro bancaria, depósitos a plazo, inversiones en fondos mutuos, acciones o bonos de empresas, préstamos a terceros y otros ahorros, (3) la deuda en casas comerciales que corresponde a la deuda contraída mediante tarjetas de crédito propias y (4) la deuda bancaria, deuda de consumo contraída con el sistema bancario, excluyendo tarjetas de crédito y líneas de crédito (Ver el Cuadro 1).

CUADRO 1
ESTADÍSTICA DESCRIPTIVA

Variable	Observaciones	Valores positivos	Promedio	Desviación estándar	Mínimo	Máximo
Ingreso	9,230	9,230	241,393	297,447	2,000	7,000,000
Activos financieros	9,351	1,426	1,835,945	8,960,278	1,000	240,000,000
Deuda casas comerciales	9,475	4,123	290,990	977,009	10,000	60,000,000
Deuda bancaria	9,608	836	2,373,384	5,238,643	10,000	60,000,000
Edad (años)	9,648	9,648	41,962	12,669	16	90
Educación (años)	9,648	9,648	10,591	4,009	0	19
Experiencia (años)	9,648	9,648	25,373	14,542	0	83

Fuente: Elaboración propia en base a EPS 2004.

A partir de la estadística descriptiva, es posible observar que el ingreso promedio de la muestra supera los 240 mil pesos y el valor promedio de los activos financieros es de 1,84 millones de pesos. Por el lado de las deudas, el promedio de la deuda bancaria supera a la de casas comerciales y alcanzan a 2,4 millones y 290 mil pesos respectivamente.

En términos de información faltante, notamos que en este caso el problema es pequeño debido a que se reportan el 95,7% de los ingresos, el 97% de los activos financieros, el 98,2% de la deuda en casas comerciales y el 99,6% de la deuda bancaria. Sin embargo, las observaciones que presentan información completa en todas las variables representan el 91,7% de la muestra. Es decir que, de utilizarse en el análisis todas las variables perdemos un 8,3% de la muestra debido a la falta de información en alguna de ellas.

La distribución de la información faltante a través de las variables permite la generación de patrones (Cuadro 2). Observamos que sólo un 0,01% de la muestra presenta falta de información en estas cuatro variables, un 0,17% en tres de ellas, mientras que un 1.01% de la muestra presenta falta de información en dos variables. Por último, un 7,1% posee falta de información en sólo una variable.

CUADRO 2
PATRÓN DE INFORMACIÓN FALTANTE

Ingreso	Activos financieros	Deuda		Observaciones	
		Casas comerciales	Bancos	Frecuencia	Porcentaje
F	F	F	F	1	0,01
F	F	F		6	0,06
F	F		F	2	0,02
F		F	F	6	0,06
	F	F	F	2	0,02
F	F			40	0,41
F		F		18	0,19
F			F	3	0,03
	F	F		17	0,18
		F	F	19	0,20
F				342	3,54
	F			229	2,37
		F		104	1,08
			F	7	0,07

Fuente: Elaboración propia en base a EPS 2004.

Nota: F se refiere que hay información faltante en la variable de dicha columna.

Notamos que el problema de información faltante en esta encuesta, es decir el 8,3% de la muestra, puede dividirse en 2 grandes casos: pérdida de información en una variable, que involucra el 7,1% de los casos y pérdida de información en más de una variable, cuya importancia es relativamente pequeña pues se refiere al 1.2% de la muestra³.

3.2. Método de Imputación

Notamos que todas las variables financieras consideradas en este ejercicio son montos en unidades monetarias, por lo que podremos asumir que ellas o sus formas funcionales tienen distribuciones continuas. Esto nos permite utilizar tanto el Método Condicional Encadenado como el Normal Multivariado. Utilizaremos este último debido a que posee un sustento teórico basado en la convergencia del algoritmo EM.

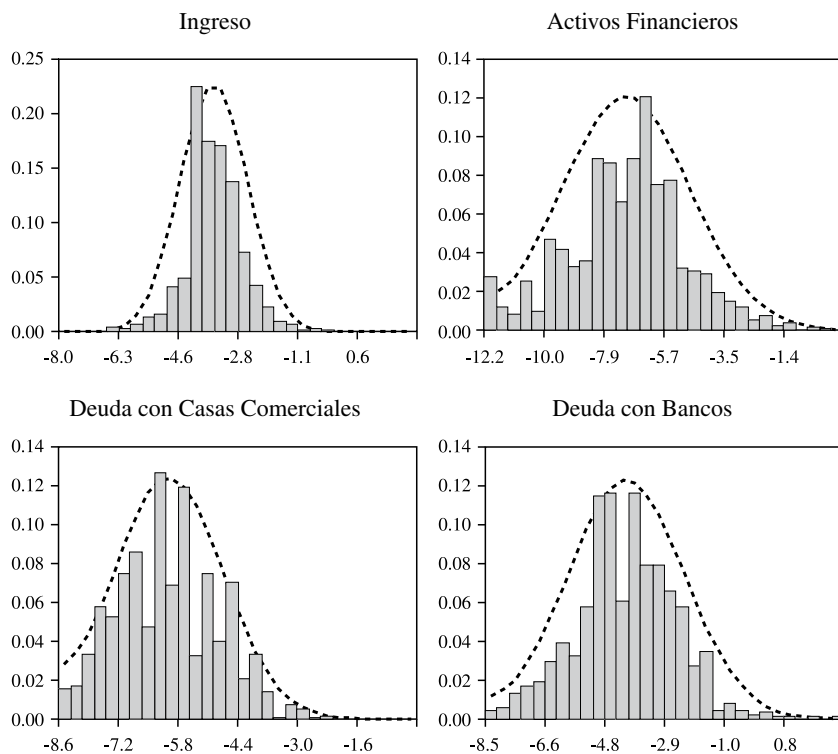
Dentro de los ajustes empíricos consideramos: (1) el uso de la transformación logística para corregir el tema de asimetrías en la distribución de las variables y generar perfiles más cercanos a la distribución normal y (2) la generación de grupos de individuos lo que permite una mejor identificación de la heterogeneidad de la muestra.

En el caso de la transformación logística, primero se normalizaron los valores entre cero y el máximo valor de cada variable en la muestra para luego aplicar la función logística. De este modo, las variables transformadas fueron usadas tanto en el paso EM como en el DA. Posterior a la imputación, estas variables se vuelven a transformar para recuperar sus valores en nivel. Las distribuciones de las variables transformadas tienen perfiles similares a la normal (Gráfico 1), lo que permite que el algoritmo EM basado en esta distribución, sea más adecuado después de utilizar esta transformación que considerando los valores originales de las variables.

Por otra parte, los grupos se construyeron utilizando la información interna de la encuestas, es decir sobre el reporte de tenencia que declararon los individuos. Debido a que la muestra se compone de individuos que reportan estar trabajando y al hecho de que la tenencia de deuda con casas comerciales es similar para todos los individuos, la construcción de los grupos se realizó a partir de la información de posesión de activos financieros y deuda bancaria. Esto permite la generación de 4 grupos mutuamente excluyentes que se reportan en el Cuadro 3.

³ Dada esta condición de la base de datos es posible considerar las imputaciones sólo para los casos en que hay una variable con información faltante a través de métodos univariados de imputación, realizando de este modo un proceso parcial de imputación. En este caso no tomaremos esta ruta pero podría ser un camino válido en problemas más complejos.

GRÁFICO 1
DISTRIBUCIÓN DE LAS VARIABLES
CON TRANSFORMACIÓN LOGÍSTICA



Fuente: Elaboración propia en base a EPS 2004.

Nota: Variables bajo transformación logística. -- Distribución Normal.

CUADRO 3
GRUPOS DE INDIVIDUOS

Variable	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Total
Activos financieros		T		T	
Deuda bancaria			T	T	
Número de individuos con información	6.820	1.240	645	147	8.852
Número de observaciones a imputar	401	311	59	25	796
Total	7.221	1.551	704	172	9.648

Fuente: Elaboración propia en base a EPS 2004.

Nota: T indica tenencia de esta variable financiera.

De esta forma, el primer grupo está compuesto por todos aquellos individuos que no tienen activos financieros ni deuda bancaria, con un total de 6.820 observaciones donde hay información completa. Notamos que los grupos donde el individuo declara tener activos financieros son los que presentan mayores grados de información faltante (Grupos 2 y 4). En particular, el segundo grupo alcanza un 20% de información faltante.

La estadística descriptiva de los grupos muestra la heterogeneidad de ellos como puede verse en el Cuadro 4. En particular, observamos que los individuos que no tienen activos financieros ni deuda con bancos (Grupo 1), tienen un menor ingreso promedio en comparación al resto de los grupos. De hecho su valor es estadísticamente menor que el que se obtiene para el Grupo 2. Sin embargo, en términos de deuda con casas comerciales ambos grupos no presentan diferencia significativa en sus promedios.

3.3 Resultados de la Imputación

Los resultados del algoritmo EM convergen en menos de 12 iteraciones, tanto para la base agregada como para cada uno de los grupos. Esto confirma empíricamente que el uso de la función logística colabora con el supuesto de normalidad impuesto en el modelo. En el Cuadro 5 presentamos los resultados de las matrices de varianzas y covarianzas que se obtienen de este algoritmo para cada uno de los grupos y el agregado.

En primer lugar notamos que, los resultados obtenidos para la varianza se encuentran fuertemente influenciados según se incluyan o no los individuos que reportan no poseer cada variable financiera. Por ejemplo, la varianza de la deuda con bancos se multiplica por cuatro al incluir estos individuos en el cálculo. Esta diferencia desaparece si: (1) se consideran imputaciones por variables en cuyo caso los individuos que no poseen dicha variable financiera desaparecen de la estimación o (2) si se realiza una imputación condicionada. Un segundo punto relevante de los resultados corresponde a los resultados obtenidos para las covarianzas entre dichas variables. Notamos que la correlación entre ingreso y deuda con casas comerciales es positiva en el agregado pero negativa para el Grupo 4. Por construcción este grupo accede a deuda bancaria, en cuyo caso podría existir sustitución entre este tipo de deudas, como lo sugiere la correlación entre estas últimas variables.

Utilizando el algoritmo DA generamos 50 imputaciones para cada grupo y para la base agregada. En el Cuadro 6 se reportan los resultados para el promedio y su error estándar en los casos que se utilicen 3, 5, 10, 20 y 50 bases imputadas. Utilizamos el Coeficiente de Variación (CV) para comparar los resultados obtenidos por la imputación con grupos y aquella que se hace de forma agregada⁴.

⁴ En este caso el CV corresponde al ratio entre H y V obtenidos según las reglas de Rubin (1987) que fueron presentadas en la Sección 2. Por otra parte, los cambios no monotónicos en CV se deben al uso de factores de expansión para la obtención de los estadísticos.

CUADRO 4
ESTADÍSTICAS DESCRIPTIVAS POR GRUPOS

	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Total
Ingreso					
Promedio	211.708	292.464	364.722	493.225	241.393
Desviación estándar	(243.899)	(400.909)	(360.940)	(518.078)	(297.447)
Activos financieros					
Promedio	n.a	1.575.272	n.a	3.979.231	1.835.945
Desviación estándar	n.a	(5.862.768)	n.a	(21.300.000)	(8.960.278)
Deuda con casas comerciales					
Promedio	270.598	280.413	450.331	331.127	290.990
Desviación estándar	(1.083.145)	(605.732)	(767.789)	(412.289)	(977.009)
Deuda con bancos					
Promedio	n.a	n.a	2.361.635	2.420.850	2.373.384
Desviación estándar	n.a	n.a	(5.432.621)	(4.382.520)	(5.238.643)

Fuente: Elaboración propia en base a EPS 2004.

n.a: no aplica.

CUADRO 5
 VARIANZA-COVARIANZA DE LAS ESTIMACIONES EM

	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Total
Var(Ing)	0,657	0,662	0,620	0,775	0,691
Cov(Ing, CC)	0,457 (0,109)	0,238 (0,055)	0,457 (0,105)	-0,280 (-0,062)	0,515 (0,118)
Cov(Ing, Bcos)	n.a n.a	n.a n.a	0,452 (0,317)	0,506 (0,426)	0,611 (0,202)
Cov(Ing, AF)	n.a n.a	0,609 (0,350)	n.a n.a	0,869 (0,345)	0,372 (0,117)
Var(CC)	26,790	27,685	30,409	26,649	27,525
Cov(CC, Bcos)	n.a n.a	n.a n.a	0,336 (0,034)	-0,528 (-0,076)	2,023 (0,106)
Cov(CC, AF)	n.a n.a	-0,739 (-0,066)	n.a n.a	-0,805 (-0,054)	0,951 (0,048)
Var(Bcos)	n.a	n.a	3,2842	1,825	13,313
Cov(Bcos, AF)	n.a n.a	n.a n.a	n.a n.a	0,736 (0,190)	0,450 (0,032)
Var(AF)	n.a	4,578	n.a	8,206	14,473

Fuente: Elaboración propia en base a EPS 2004.

Nota: Var y Cov indican varianza y covarianza. Correlaciones entre paréntesis. Las variables son ingreso (Ing), deuda con casas comerciales (CC), deuda con bancos (Bcos) y activos financieros (AF).
 n.a: no aplica.

CUADRO 6
ESTADÍSTICAS DESCRIPTIVAS DE LAS IMPUTACIONES MÚLTIPLES

Variable	m	Grupos			Agregado		
		H	V	CV	H	V	CV
Ingreso	3	241.421	3.027	1,25	242.077	3.029	1,25
	5	241.484	3.029	1,25	241.860	3.028	1,25
	10	241.381	3.021	1,25	241.865	3.024	1,25
	20	241.115	3.033	1,26	241.916	3.039	1,26
	50	240.979	3.030	1,26	241.924	3.045	1,26
Activos Financieros	3	1.680.214	199.400	11,87	1.493.740	195.353	13,08
	5	1.682.030	199.299	11,85	1.495.441	195.398	13,07
	10	1.685.001	205.039	12,17	1.496.870	195.437	13,06
	20	1.690.231	206.329	12,21	1.497.906	195.531	13,05
	50	1.705.377	217.560	12,76	1.498.547	196.061	13,08
Deuda con casas comerciales	3	314.396	23.922	7,61	324.480	24.318	7,49
	5	312.165	26.528	8,50	322.369	24.067	7,47
	10	304.428	24.742	8,13	315.674	24.364	7,72
	20	303.482	23.773	7,83	315.179	25.251	8,01
	50	305.885	24.379	7,97	320.066	29.852	9,33
Deuda con bancos	3	2.397.496	177.794	7,42	2.250.393	173.234	7,70
	5	2.386.318	179.873	7,54	2.250.433	173.233	7,70
	10	2.396.975	180.944	7,55	2.250.406	173.234	7,70
	20	2.393.021	180.841	7,56	2.250.464	173.233	7,70
	50	2.414.774	188.435	7,80	2.250.444	173.233	7,70

Fuente: Elaboración propia en base a EPS 2004.

Nota: H y V son el promedio y su error estándar obtenido usando fórmula de Rubin. CV es coeficiente de variación (en porcentaje).

Observamos que para el caso del ingreso no hay diferencias significativas entre ambos procedimientos, obteniéndose en ambos casos un ingreso promedio cercano a los 240 mil pesos. En el caso de los activos financieros los resultados indican que la estimación por grupos es levemente más precisa que la estimación agregada, incrementándose el promedio a aproximadamente 200 mil pesos. Al revisar las cifras de deudas notamos que en el caso de las casas comerciales la estimación por grupos ofrece una importante mejora cuando se considera $m=20$, lo cual es coherente con los resultados de Royston (2004). Es importante notar que en el caso de la imputación agregada el CV se incrementa conforme se aumenta el número de bases imputadas a considerar. En el caso de la deuda bancaria los resultados muestran una pequeña mejora del método de imputaciones por grupos.

Siguiendo lo sugerido por Royston (2004) en el caso analizado correspondería utilizar al menos 20 bases imputadas para considerar la incertidumbre de la información faltante y mantener intervalos de confianza que sean confiables.

4. CONCLUSIONES

En este trabajo, discutimos distintas técnicas que permiten enfrentar el problema de la información faltante en encuestas microeconómicas. En particular, revisamos de manera descriptiva los distintos métodos de imputación múltiple (MI), los cuales consideran la incertidumbre asociada a la estimación del modelo que genera los datos.

Utilizando la Encuesta de Protección Social 2004, mostramos la aplicación del método normal multivariado a través del algoritmo EM/DA. Se consideraron 2 modificaciones empíricas al algoritmo que son la transformación de las variables utilizando la función logística y la implementación de la imputación por grupos de individuos. Mostramos que estas dos modificaciones permiten al usuario mejorar los resultados de MI debido a que: (1) las variables transformadas presentan un perfil más cercano a la normalidad, que es el supuesto del modelo y (2) la estimación por grupos permite un mejor control sobre la heterogeneidad de la muestra. En particular, observamos que hay una mejora en la precisión de la estimación (reducción del CV) cuando la estimación se realiza por grupos de individuos que cuando se consideran todos de forma homogénea.

REFERENCIAS

- Alfaro, R. y M. Fuenzalida (2008), "Análisis de Información Faltante en Encuestas Microeconómicas" Estudios Económicos Estadísticos N° 67, Banco Central de Chile.
- Allison, P. (2001), *Missing Data*, Quantitative Applications in the Social Sciences, A Sage University Papers Series.
- Allison, P. (2009), Apuntes del Curso: Missing Data; Los Angeles, California, Mayo 8 y 9.
- Barceló, C. (2006), "Imputation of the 2002 wave of the Spanish survey of household finances (EFF)" Documentos Ocasionales N° 0603, Banco de España.

- Cox, P., E. Parrado y J. Ruiz-Tagle (2006), "The Distribution of Assets, Debt and Income among Chilean Households" Documento de Trabajo N° 388, Banco Central de Chile.
- Fuenzalida M. y J. Ruiz-Tagle (2009), "Household's Financial Vulnerability", *Journal Economía Chilena (The Chilean Economy)*, 12(2):35-53.
- Little, R. y D. Rubin (2002), *Statistical Analysis with Missing Data*, Second Edition J. Wiley & Sons, New York.
- Royston, P. (2004), "Multiple imputation of Missing Values" *Stata Journal* 4(3):227-241.
- Rubin, D. (1987), *Multiple Imputation for Nonresponse in Surveys*, J. Wiley & Sons, New York.
- Schafer, J. (1997), *Analysis of Incomplete Multivariate Data*, Chapman & Hall/CRC.
- StataCorp (2009) *Multiple-Imputation Reference Manual*, Stata Press.
- Van Buuren, S., J. Brand, C. Groothuis-Oudshoorn y D. Rubin (2006) "Fully Conditional Specification in Multivariate Imputation", *Journal of Statistical Computation and Simulations* 76(12):1049-1064.