ESTUDIOS ECONÓMICOS ESTADÍSTICOS

BANCO CENTRAL DE CHILE

Análisis de Información Faltante en Encuestas Microeconómicas

Rodrigo Alfaro
Marcelo Fuenzalida

N.° 67 - Septiembre 2008

STUDIES IN ECONOMIC STATISTICS

CENTRAL BANK OF CHILE

BANCO CENTRAL DE CHILE

CENTRAL BANK OF CHILE

Los *Estudios Económicos Estadísticos* - hasta el número 49, *Serie de Estudios Económicos* - divulgan trabajos de investigación en el ámbito económico estadístico realizados por profesionales del Banco Central de Chile, o encargados por éste a especialistas o consultores externos. Su contenido se publica bajo exclusiva responsabilidad de sus autores y no compromete la opinión del Instituto Emisor. Estos trabajos tienen normalmente un carácter definitivo, en el sentido que, por lo general, no se vuelven a publicar con posterioridad en otro medio final, como una revista o un libro.

As from issue number 50, the *Series of Economic Studies* of the Central Bank of Chile will be called *Studies in Economic Statistics.*

Studies in Economic Statistics disseminates works of investigation in economic statistics carried out by professionals of the Central Bank of Chile or by specialists or external consultants. Its content is published under exclusive responsibility of its authors and it does not reflect the opinion of the Central Bank. These documents normally are definitives and are not made available in any other media such as books or magazines.

Estudios Económicos Estadísticos del Banco Central de Chile
Studies in Economic Statistics of the Central Bank of Chile
ISSN 0716 - 2502

Agustinas 1180, primer piso.
Teléfono: (56-2) 6702475; Fax: (56-2) 6702231

Estudios Económicos Estadísticos
N° 67

Studies in Economic Statistics
N° 67

Análisis de Información Faltante en Encuestas Microeconómicas (*)

Rodrigo Alfaro
Gerencia de Estabilidad Financiera
Banco Central de Chile

Marcelo Fuenzalida
Gerencia de Estabilidad Financiera
Banco Central de Chile

Resumen

En este trabajo presentamos metodologías para el manejo de información faltante que transforman la base con datos incompletos en varias bases completas las cuales pueden ser analizadas separadamente y sus resultados combinados con reglas claras y simples. La presentación del material se organiza en torno a una aplicación empírica tradicional en economía: la estimación de la ecuación de Mincer, utilizando la Encuesta de Protección Social 2004. Los resultados muestran que métodos tradicionales de manejo de información faltante reducen artificialmente la varianza de la variable en tratamiento.

Abstract

In this article we provide an executive survey of methods for missing data. We note that standard methods reduce the sample variances meanwhile bayesian methods keep track of the uncertainty associated with missing information. We discuss these methods in an empirical application using a well-known household survey. We know that the exercise is simple but it helps in the understanding of this material.

(*) Agradecemos los valiosos comentarios de Alejandra Marinovic y un árbitro anónimo. Este artículo es de exclusiva responsabilidad de los autores y no compromete necesariamente la visión del Banco Central de Chile o de su Consejo. Contactos: ralfaro@bcentral.cl y mfuenzal@bcentral.cl.

I. Introducción

Las encuestas con datos microeconómicos han sido ampliamente utilizadas para generar y evaluar políticas públicas, como por ejemplo programas para la superación de la pobreza con la Encuesta de Caracterización Socioeconómica (CASEN), y la evaluación del sistema de ahorro previsional con la Encuesta de Protección Social (EPS). 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.

El uso de este tipo de encuestas involucra un desafío estadístico que corresponde al manejo de la información faltante (*missing data*). Esta puede producirse por múltiples razones y su complejidad depende de la cantidad de variables no reportadas por el encuestado y del proceso estocástico que genera la omisión de dicha información. Como una forma de aminorar el efecto de la información faltante, los centros especializados en la toma de encuestas definen como completa aquella encuesta que tiene un porcentaje mínimo de respuestas para la información requerida en el cuestionario. Las encuestas que no cumplen con este requisito son reemplazadas por otras que tengan igual representatividad.

Cuestionarios incompletos son producto de una serie de eventos. Elementos asociados a la comprensión de las preguntas suelen ser controlados con una exhaustiva capacitación de los encuestadores, mientras que la disposición a revelar la información por parte de los encuestados puede abordarse mediante un cuestionario coherente que posea consistencia interna y cuyas preguntas no afecten la sensibilidad del encuestado. Esto último se logra a través de la aplicación de cuestionarios pilotos, los cuales son realizados previamente al levantamiento oficial de la encuesta.

Pese a los resguardos mencionados anteriormente, los cuestionarios finales presentan algún grado de información faltante. En general este problema está asociado a variables que son de interés para el analista como por ejemplo ingresos, gastos, deudas y activos. Datos

preliminares de la Encuesta Financiera de Hogares 2007 (EFH 2007) muestran que la información faltante constituye un problema real en los montos de estas variables, pero no así en la declaración de tenencia. Esto implica que resultados robustos pueden ser calculados con la estadística tradicional para la tenencia de deuda y activos (véase por el Informe de Estabilidad Financiera del primer semestre del 2008).

El manejo de la información faltante usualmente se reduce a la eliminación de toda la observación (*listwise deletion*). Por ejemplo, supongamos que tenemos una encuesta con información en las variables *x* e *y* pero no posee información en la variable *z*. En un estudio que involucre a estas tres variables la metodología usual ignora esta encuesta, reduciendo de este modo la cantidad de observaciones disponibles. Es evidente que este procedimiento podría inducir una muestra pequeña si el problema de información faltante se encuentra presente en múltiples variables de interés para el analista.

Dentro de la literatura especializada (Rubin, 1987; Schafer, 1997; Allison, 2002), el problema de información faltante se analiza como si este fuese generado a través de un proceso aleatorio. En particular, el supuesto es que la información faltante ha sido omitida de forma aleatoria sin estructurar el proceso estocástico que induce la falta de información (*missing at random*). Así, la información observada en los cuestionarios permite hacer inferencia sobre aquella omitida.

Siguiendo con nuestro ejemplo y suponiendo que la pérdida de información en *z* se debe a un proceso aleatorio, podemos utilizar la información que sí fue reportada para asignar un valor posible para *z* que sea consistente con los valores observados de *x* e *y*. Sin embargo, este valor es una variable aleatoria y debe ser considerado como tal al momento de utilizarlo. Metodologías comúnmente aplicadas, completan o imputan la información faltante a través de promedios y emplean los valores imputados como valores efectivos, ignorando así que corresponden a variables aleatorias.

En este trabajo presentamos de manera didáctica distintos métodos de imputación de información faltante. Nuestros objetivos son: (1) revisar ejecutivamente los métodos

disponibles en la literatura y los que usualmente se aplican y (2) mostrar con un ejemplo concreto el uso de estas metodologías, para lo cual utilizamos la EPS 2004 y el software Stata 9.0, para el que indicamos explícitamente los comandos utilizados en nuestros cálculos.

II. Descripción de los datos

La base de datos utilizada en este estudio corresponde a 9648 jefes de hogar que reportan estar trabajando en la EPS 2004. El procedimiento entonces, debe imputar solo a aquellos jefes de hogar que responden positivamente a la pregunta de estar ocupados y no a los que no trabajan (Acock, 2005).

Para nuestro análisis, consideramos las siguientes cuatro variables:

ling	corresponde al logaritmo del ingreso líquido mensual del individuo
género	Dummy que corresponde a 0=hombre;1=mujer
edad	edad reportada por el individuo
educ	años de educación reportados

El hecho de tomar ling en vez del ingreso directamente permite evitar el ajuste de Jensen en la predicción del ingreso y se acomoda perfectamente con la estimación de las ecuaciones de retorno educacional presentadas más adelante. Del mismo modo, las variables edad y educ permiten generan experiencia potencial (exp) y experiencia potencial al cuadrado (exp2), bajo el supuesto de que exp = edad – educ - 6.

La estadística descriptiva de los datos muestra que solo la variable ling tiene información faltante (Tabla 1). Además para este ejemplo, el problema de la información faltante resulta ser menor dado el alto porcentaje de información disponible (9230/9648 = 96%).

Por otra parte, el ejercicio es simple en circunstancias en que los patrones de información son reducidos. Existen 2 patrones: ling tiene observación y ling no tiene observación.

Adicionalmente los patrones son monótonos dado que es posible escalar las observaciones conforme su grado de información faltante.

Lo anterior puede ser poco probable de observar en la práctica cuando existen varias variables con información faltante. En dichos es probable que los patrones de información no sean monótonos. Dichos casos han sido revisados en la literatura y son complicaciones técnicas de los métodos aquí presentados.

Tabla 1: Estadística descriptiva con datos existentes

Variable	Observaciones	Promedio	desviación estándar
ling	9230	12.024	0.775
género	9648	0.369	0.483
edad	9648	41.866	12.538
educ	9648	10.416	3.971
exp	9648	25.450	14.413
exp2	9648	855.441	869.786

Fuente: Elaboración propia en base a EPS 2004.

III. Metodologías usuales

En esta sección utilizamos métodos de imputación no aleatoria para completar los 418 datos faltantes en la variable ling. Rubin (1987) define estos métodos como no apropiados en el sentido estadístico, pero en la práctica resultan ser ampliamente utilizados.

1. Imputación por celda

Este método consiste en completar la información faltante con el promedio por celdas de la variable ling. La forma más simple es dividir la muestra por género y utilizar los valores promedio para completar la base de datos. Basados en la estadística descriptiva de esta variable separada entre hombres y mujeres (Tabla 2)[1]/, imputaremos la información faltante con la siguiente regla: la celda vacía tomará el valor 11.89 si corresponde a una mujer y 12.11 si corresponde a un hombre.

[1]/ En Stata tipeamos: table genero, c(N ling mean ling sd ling min ling max ling) row.

Tabla 2: Estadística descriptiva del logaritmo del ingreso por género

Logaritmo del Ingreso (ling)					
Género	Observaciones	Promedio	Desviación estándar	Min.	Max.
Hombre	5833	12.105	0.764	7.601	15.761
Mujer	3397	11.886	0.773	7.824	14.509
Total	9230	12.024	0.775	7.601	15.761

Fuente: Elaboración propia en base a EPS 2004.

Con la información reportada podemos comparar las medias de ling por género. Bajo el supuesto de varianzas iguales este ejercicio corresponde a un *test* clásico en estadística que entrega un estadístico t de 13.2[2]/, el cual permite rechazar la hipótesis de que las medias son iguales. El resultado cualitativo no varía al levantar el supuesto de varianzas iguales.

Un procedimiento más elaborado implicaría abrir los datos por celdas de género, tramos de edad y tramos de educación (Tabla 3)[3]/, de este modo, se incorpora el hecho económico que a mayor educación y/o experiencia, mayor es el ingreso percibido.

Es importante notar que la aplicación de esta metodología requiere generar tramos de las variables edad y educación para aumentar el número posible de observaciones en cada celda. En nuestro ejercicio, dicha estructuración corresponde a enseñanza básica, media, universitaria y postgrado en el caso de la educación y a tramos de 10 años para la edad.

Con el mayor detalle de la información podemos refinar el cálculo anterior concentrándonos en un cruce específico. En particular, consideremos a los encuestados menores a 25 años con nivel de educación universitaria, es decir, aquellos que recién entran al mercado laboral. Para este ejercicio, el estadístico relevante tiene un valor de 3.5, el cual implica que las conclusiones cualitativas no han cambiado respecto del ejercicio que sólo distinguía entre hombres y mujeres.

[2]/ En Stata tipeamos: ttesti 5833 12.105 0.764 3397 11.886 0.773.
[3]/ En Stata tipeamos: bysort sexo: table teduc tedad, c(N ling N genero mean ling sd ling).

Tabla 3: Estadística descriptiva de la información utilizada para la imputación por celdas.

Logaritmo del Ingreso (ling)							
Hombres							
		Tramos de edad					
		menor de 25 años	25 - 34	35 - 44	45 - 54	55 - 64	Mayor de 65 años
Nivel de escolaridad	0 a 8 años	67 / 69	276 / 281	556 / 575	568 / 589	458 / 481	219 / 237
	media	11.737	11.744	11.768	11.718	11.738	11.419
	σ	(0.494)	(0.542)	(0.577)	(0.687)	(0.723)	(0.933)
	9 a 12 años	290 / 298	684 / 700	742 / 762	552 / 570	199 / 206	59 / 63
	media	11.949	12.114	12.182	12.227	12.252	12.005
	σ	(0.384)	(0.483)	(0.589)	(0.661)	(0.739)	(1.129)
	13 a 16 años	82 / 83	321 / 333	248 / 267	135 / 143	67 / 71	9 / 10
	media	12.092	12.438	12.676	12.733	12.903	13.002
	σ	(0.534)	(0.609)	(0.672)	(0.801)	(0.724)	(0.725)
	17 o más	7 / 8	91 / 108	80 / 93	69 / 78	35 / 40	19 / 20
	media	12.753	12.948	13.326	13.380	13.550	13.579
	σ	(0.505)	(0.697)	(0.663)	(0.748)	(0.743)	(0.882)

Logaritmo del Ingreso (ling)							
Mujeres							
		Tramos de edad					
		menor de 25 años	25 - 34	35 - 44	45 - 54	55 - 64	Mayor de 65 años
Nivel de escolaridad	0 a 8 años	13 / 13	110 / 115	210 / 214	268 / 279	182 / 197	63 / 70
	media	11.427	11.372	11.502	11.386	11.464	11.171
	σ	(0.473)	(0.745)	(0.553)	(0.628)	(0.706)	(1.045)
	9 a 12 años	186 / 190	394 / 404	470 / 489	326 / 342	112 / 116	24 / 26
	media	11.729	11.830	11.723	11.805	11.872	11.514
	σ	(0.441)	(0.552)	(0.621)	(0.665)	(0.920)	(1.243)
	13 a 16 años	73 / 77	249 / 269	252 / 265	131 / 141	32 / 38	5 / 6
	media	11.782	12.206	12.331	12.478	12.830	12.461
	σ	(0.575)	(0.587)	(0.650)	(0.633)	(0.637)	(0.4253)
	17 o más	12 / 13	97 / 103	77 / 82	68 / 70	40 / 41	3 / 3
	media	11.994	12.603	12.845	12.982	12.984	13.136
	σ	(0.613)	(0.666)	(0.625)	(0.704)	(0.546)	(0.605)

Nota: La celda contiene el número de observaciones en la muestra / número de observaciones totales, seguido por la media y la desviación estándar (σ) entre paréntesis.

Fuente: Elaboración propia en base a EPS 2004.

2. Imputación económica

Este método es un refinamiento del anterior dado que la información faltante es imputada a través de un modelo de regresión lineal con las variables originales en vez de agrupadas en tramos. Adicionalmente, la regresión utilizada tiene un fundamento económico y es conocida como la ecuación de retornos a la educación o ecuación de Mincer[4]/.

Distintas especificaciones han sido utilizadas en la literatura empírica (Tabla 4). Observamos que la especificación 0 corresponde exactamente a la imputación por celdas presentada

[4]/ El marco teórico se debe a Mincer (1974). Para aplicaciones al caso chileno véase Sapelli (2003).

anteriormente, esto porque el modelo de regresión lineal coincide en ese caso con la imputación por celdas.

En la literatura empírica la especificación 2 es la más utilizada debido a que recoge los resultados de un modelo de acumulación de capital humano, los posibles efectos de discriminación por género y el efecto de rendimientos decrecientes en la variable experiencia potencial.

Tabla 4: Estimación de ecuaciones de Mincer con información incompleta

	Especificación 0	**Especificación 1**	**Especificación 2**
género	-0.219*	-0.334*	-0.335*
	(-13.23)	(-23.7)	(-23.91)
educ		0.120*	0.118*
		(57.35)	(56.42)
exp		0.008*	0.023*
		(14.15)	(14.12)
exp2			-0.0003*
			(-9.69)
Constante	12.105*	10.690*	10.562*
	(1204.76)	(322.29)	(297.10)
Observaciones	9230	9230	9230
R2	0.019	0.305	0.312

*Notas: ling es la variable dependiente. * Parámetro significativo al 1%. Test t entre paréntesis.*

Fuente: Elaboración propia en base a EPS 2004.

Los resultados de las imputaciones tanto utilizando el promedio por celda como, el método de valores predichos a partir de las ecuaciones de Mincer 1 y 2 nos muestran una pequeña variación en el promedio de los datos generados relativo a los originales (Tabla 5).

Tabla 5: Estadística descriptiva de datos imputados en forma no aleatoria

Variable	Observaciones	Promedio	Desviación estándar
ling	**9230**	**12.024**	**0.775**
ling_prom1_genero	9648	12.037	0.760
ling_prom2_tedad teduc	9648	12.038	0.770
ling_mincer 1	9648	12.028	0.765
ling_mincer 2	9648	12.028	0.765

Fuente: Elaboración propia en base a EPS 2004.

Sin embargo, las desviaciones estándar de estas nuevas variables que incluyen datos imputados son menores que la variable original (Acock, 2005).

El resultado anterior es esperable debido a que las observaciones imputadas —que originalmente eran variables aleatorias— pierden su componente estocástico al asumirse que son conocidas. Para el caso simple del promedio se muestra este hecho algebraicamente en el Apéndice A.

IV. Imputación múltiple

Rubin (1987) recoge el tema de incertidumbre de los valores imputados proponiendo el método de imputaciones múltiples (*multiple imputation*). Esto quiere decir que a través de un proceso aleatorio se muestran posibles valores para la información faltante y la utilización de dichos valores recoge el componente aleatorio del dato imputado.

El método de imputaciones múltiples no soluciona el tema 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 con información 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 de Rubin (1987). Para el análisis descriptivo, la media (*H*) y varianza (*V*) combinadas son:

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

donde Q_t y V_t corresponden al promedio y varianza estimados en la base *t*. La primera expresión en V es el promedio de las varianzas obtenidas (*within imputation variance*), mientras que la expresión en corchetes es un estimador de la dispersión de los promedios obtenidos (*between imputation variance*).

El número óptimo de bases de datos (*m*) depende del grado de información faltante. Schafer (1997) discute sobre el grado de eficiencia que se obtiene al incrementar *m*, el cual puede analizarse empíricamente a través del grado de información faltante.

Sin embargo, la práctica muestra que para un porcentaje de información faltante pequeño, *m*=5 es una elección razonable (Royston, 2005). Para simplicidad del análisis utilizaremos este número.

1. Imputación por *Hot-Deck*

Este método asigna valores a los datos faltantes con la información existente en la muestra de acuerdo a la celda en que se encuentra la observación con información faltante. El procedimiento consiste en que en cada celda se completan 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 *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 ser una variable discreta.

En nuestro ejemplo, las celdas utilizadas corresponden a las mismas que se construyeron para la imputación no aleatoria. El primer ejercicio considera dos grandes celdas, que son las definidas por el género. Realizando el remuestreo de forma aleatoria, se generan las cinco bases de datos completas. Igual procedimiento se aplica cuando se utilizan celdas más

detalladas, es decir cuando se realizan las separaciones por género, tramos de edad y tramos de educación.

Los resultados muestran que las bases imputadas solo por género presentan un promedio similar por ello, la desviación estándar combinada es cercana a la que se obtiene de los datos originales. Cuando las bases son imputadas por celdas más detalladas, tanto el promedio como la desviación estándar presentan leves variaciones (Tabla 6).

Tabla 6: Resultados de la imputación por Hot Deck

Logaritmo del Ingreso imputado por celda género			
Variable	Observaciones	Promedio	Desviación estándar
ling	**9230**	**12.024**	**0.775**
ling_1	9648	12.024	0.775
ling_2	9648	12.024	0.772
ling_3	9648	12.026	0.774
ling_4	9648	12.025	0.777
ling_5	9648	12.026	0.773
ling_HD_1		12.025	0.774

Logaritmo del Ingreso imputado por celda género, tramos edad y educación			
Variable	Observaciones	Promedio	Desviación estándar
ling	**9230**	**12.024**	**0.775**
ling_1	9648	12.029	0.779
ling_2	9648	12.029	0.781
ling_3	9648	12.028	0.778
ling_4	9648	12.030	0.778
ling_5	9648	12.029	0.780
ling_HD_2		12.029	0.779

Fuente: Elaboración propia en base a EPS 2004.

Los resultados anteriores son consistentes con la estadística descriptiva de los datos originales. En particular, celdas con alto nivel de educación presentan también alta dispersión, generando de esta forma una varianza combinada que es mayor que la de los datos originales.

Este método es sencillo de ocupar y su implementación sólo requiere un algoritmo de generación de números aleatorios uniformes. Sin embargo, investigadores del área dudan que

el método sea capaz de recoger en la varianza combinada toda la incertidumbre asociada a la información faltante.

2. Imputación condicional (uvis)

Este método está basado en un modelo de regresión que utiliza simulaciones de los parámetros estimados a fin de obtener una imputación aleatoria que contenga la incertidumbre asociada a la estimación de los parámetros.

Siguiendo a Rubin (1987), el procedimiento consiste en utilizar los parámetros estimados y la matriz de varianzas y covarianzas a partir de una regresión con los n_1 datos disponibles y los vectores (Y_i, X_i), con $i \in obs$:

$$Y_i = X_i \beta + \mu_i$$

De ella se obtienen los siguientes estimadores:

$$\hat{\sigma}_1^2 = \frac{\sum_{obs} (Y_i - X_i \hat{\beta}_1)^2}{n_1 - q}$$

$$\hat{\beta}_1 = V \left[\sum_{obs} X_i^t Y_i \right]$$

donde $V = \left[\sum_{obs} X_i^t X_i \right]^{-1}$

Luego, el proceso de imputación se puede describir con los siguientes pasos:

1. Construir la variable aleatoria g que se distribuye $\chi^2_{n_1-q}$ y con ella construir:

$$\sigma_*^2 = \frac{\hat{\sigma}_1^2 (n_1 - q)}{g}$$

2. Generar q valores a partir de una distribución $N(0,1)$ de manera independiente para crear un vector Z de dimensión q. Con este se construye:

$$\beta_* = \hat{\beta}_1 + \sigma_* [V]^{1/2} Z$$

donde $[V]^{1/2}$ corresponde a la descomposición de Cholesky.

3. Generar los n_0 valores de Y_{miss} como:

$$Y_{i*} = X_i \beta_* + z_i \sigma_*$$

donde los z_i se distribuyen normales y son independientes entre ellos.

Un nuevo valor para Y_{miss} se obtiene al generar un nuevo parámetro σ_*^2. En consecuencia, si se quieren *m* bases de datos, los pasos anteriores deben repetirse *m* veces de manera independiente.

Los resultados de la aplicación de esta metodología para el caso de la variable ling, utilizando las especificaciones 1 y 2 recogidas de las ecuaciones de Mincer, son consistentes con los obtenidos por el método *Hot-Deck* dando cuenta que en este ejemplo el problema de información faltante no es severo (Tabla 7)[5]/.

[5]/ Utilizamos el código uvis en Stata: uvis reg ling genero educ exp exp2, g(ling_uvis).

Tabla 7: Resultados de la imputación condicional

Logaritmo del Ingreso imputado condicional a género, educ y exp			
Variable	Observaciones	Promedio	Desviación estándar
ling	**9230**	**12.024**	**0.775**
ling_uvis1	9648	12.029	0.775
ling_uvis2	9648	12.029	0.778
ling_uvis3	9648	12.029	0.777
ling_uvis4	9648	12.029	0.782
ling_uvis5	9648	12.026	0.776
ling_uvis_I		12.029	0.778

Logaritmo del Ingreso imputado condicional a género, educ exp y exp2			
Variable	Observaciones	Promedio	Desviación estándar
ling	**9230**	**12.024**	**0.775**
ling_uvis1	9648	12.030	0.778
ling_uvis2	9648	12.028	0.775
ling_uvis3	9648	12.027	0.777
ling_uvis4	9648	12.027	0.779
ling_uvis5	9648	12.027	0.777
ling_uvis_II		12.028	0.777

Fuente: Elaboración propia en base a EPS 2004.

3. Imputación normal multivariada

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 puede ser fácilmente implementado para el caso en que exista información faltante a través del algoritmo EM (*expectation maximization*) el cual realiza el proceso de maximización previo cálculo del valor esperado de la condición de primer orden.

El algoritmo EM para el caso de la distribución normal multivariada es obtenido a través del uso del operador *sweep* (Schafer, 1997). En nuestro ejemplo, el problema es aún más sencillo en circunstancias que sólo existen dos patrones de datos y ellos son monótonos. De este modo,

la estimación de los parámetros del modelo pueden obtenerse por reiteradas aplicaciones del operador *sweep* y su inverso (*reverse-sweep*)[6]/.

El operador *sweep* se define para matrices simétricas. Así, G es una matriz simétrica de pxp con elementos g_{ij}. El operador SWP[k] reemplaza la matriz G por otra matriz H simétrica pxp:

$$H = SWP[k]G$$

donde los elementos de H se definen como:

$$\begin{aligned} h_{kk} &= -1/g_{kk} \\ h_{jk} &= h_{kj} = g_{jk}/g_{kk} \qquad \text{para} \quad j \neq k, \\ h_{jl} &= h_{lj} = g_{jl} - g_{jk}g_{kl}/g_{kk} \quad \text{para} \quad j \neq k \ \text{y} \ l \neq k \end{aligned}$$

El operador inverso de *sweep* (*RSW*) permite devolver una matriz que se le ha aplicado el operador *sweep* a su forma original. El operador *RSW* se define como:

$$H = RSW[k]G$$

y reemplaza los elementos de la matriz G con:

$$\begin{aligned} h_{kk} &= -1/g_{kk} \\ h_{jk} &= h_{kj} = -g_{jk}/g_{kk} \qquad \text{para} \quad j \neq k, \\ h_{jl} &= h_{lj} = g_{jl} - g_{jk}g_{kl}/g_{kk} \quad \text{para} \quad j \neq k \ \text{y} \ l \neq k \end{aligned}$$

[6]/ Para detalles técnicos véase Little y Rubin (2002).

En nuestro ejemplo, contamos con 2 patrones de información con los cuales generamos 2 grupos de variables. El primero de ellos está generado por el conjunto de las variables: género, educ, exp y exp2. Este grupo contiene 9648 observaciones completas.

El segundo grupo, asociado al segundo patrón de información, incluye la variable ling. Debido a que esta presenta información faltante, el número de observaciones de este grupo se reduce a 9230.

La matriz siguiente resume el vector de medias y la matriz de varianza y covarianzas del primer conjunto de variables. Notamos que esta matriz es simétrica y que por conveniencia matemática se denota un -1 en el extremo superior izquierdo[7]/.

$$A_1 = \begin{pmatrix} -1 & \mu_1^T \\ \mu_1 & \Sigma_1 \end{pmatrix} = \begin{pmatrix} -1 & & & & \\ 0.37 & 0.23 & & & \\ 10.42 & 0.26 & 15.77 & & \\ 25.45 & -0.60 & -33.17 & 207.72 & \\ 855.44 & -36.98 & -1995.55 & 11984.09 & 756448.9 \end{pmatrix}$$

Utilizando el segundo conjunto de variables obtenemos la siguiente matriz:

$$A_2 = \begin{pmatrix} -1 & & & & & \\ 0.37 & 0.23 & & & & \\ 10.38 & 0.26 & 15.50 & & & \\ 25.37 & -0.62 & -32.49 & 205.27 & & \\ 848.76 & -37.78 & -1945.93 & 11790.04 & 740803.48 & \\ 12.02 & -0.05 & 1.52 & -2.04 & -142.27 & 0.599 \end{pmatrix}$$

Ambas matrices corresponden a estimadores suficientes bajo el escenario de que las variables incorporadas se distribuyan multivariada normal (Schafer, 1997).

[7]/ Para código en Stata véase el Apéndice B.

Utilizando el operador *sweep* generamos $B_1 = SWP[2,3,4,5]A_1$ y $B_2 = SWP[2,3,4,5]A_2$, los cuales nos permiten computar el vector de medias y varianzas utilizando toda la información disponible (algoritmo EM).

$$B_1 = \begin{pmatrix} -28.22 & & & & \\ 0.93 & -4.37 & & & \\ 1.31 & 0.07 & -0.097 & & \\ 0.78 & 0.001 & -0.008 & -0.06 & \\ -0.01 & -0.000056 & -0.00012 & 0.001 & -0.00002 \end{pmatrix}$$

$$B_2 = \begin{pmatrix} -28.25 & & & & & \\ 0.90 & -4.38 & & & & \\ 1.31 & 0.069 & -0.099 & & & \\ 0.79 & 0.002 & -0.008 & -0.057 & & \\ -0.008 & -0.00008 & -0.00012 & 0.001 & -0.000016 & \\ 10.56 & -0.33 & 0.12 & 0.023 & -0.0003 & 0.413 \end{pmatrix}$$

La matriz final obtenida por EM consiste en completar la información de B_1 con las últimas filas columnas de B_2, la cual incorpora información de la variable ling:

$$B_3 = \begin{pmatrix} -28.22 & & & & & \\ 0.93 & -4.37 & & & & \\ 1.31 & 0.07 & -0.097 & & & \\ 0.78 & 0.001 & -0.008 & -0.06 & & \\ -0.01 & -0.000056 & -0.00012 & 0.001 & -0.00002 & \\ 10.56 & -0.33 & 0.12 & 0.023 & -0.0003 & 0.413 \end{pmatrix}$$

A través del operador inverso de *sweep* se puede obtener el vector de medias y la matriz de varianzas y covarianzas del método EM:

$$A_3 = \begin{pmatrix} -1 & & & & & \\ 0.37 & 0.23 & & & & \\ 10.42 & 0.26 & 15.77 & & & \\ 25.45 & -0.60 & -33.17 & 207.72 & & \\ 855.44 & -36.98 & -1995.55 & 11984.09 & 756448.9 & \\ 12.03 & -0.05 & 1.55 & -2.12 & -148.06 & 0.60 \end{pmatrix}$$

Sobre la estimación de estos parámetros se realiza una simulación suponiendo que las distribuciones asintóticas de ellos son: 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*) y se interpreta como un Método de Monte Carlo con una cadena de Markov (*Markov Chain Monte Carlo, MCMC*).

Los resultados generados con este algoritmo (DA) son consistentes tanto con la metodología de imputación condicional y con el método de *Hot-Deck* (Tabla 8)[8]/.

Tabla 8: Resultados de la imputación multivariada normal

Logaritmo del Ingreso imputado por norm			
Variable	Observaciones	Promedio	Desviación estándar
ling	**9230**	**12.024**	**0.775**
ling_norm1	9648	12.027	0.776
ling_norm2	9648	12.028	0.778
ling_norm3	9648	12.029	0.777
ling_norm4	9648	12.028	0.776
ling_norm5	9648	12.028	0.776
ling_NORM		12.028	0.777

Fuente: Elaboración propia en base a EPS 2004.

[8]/ Los resultados están basados en el paquete estadístico NORM escrito por Joseph Schafer el cual está disponible en http://www.stat.psu.edu/~jls/misoftwa.html.

El primero de estos hechos era esperable en consideración que el modelo condicional utilizado es lineal, lo que lo hace replicable analíticamente por medio del modelo multivariado. Sin embargo, la correspondencia entre el método condicional y el método multivariado no siempre está disponible.

Por otra parte, el algoritmo EM en el cual DA descansa sus principios teóricos, ha sido demostrado su convergencia asintótica mientras que el método condicional y con ello el método condicional encadenado —que es la extensión presentada por van Buuren et al. (2005)— no posee propiedades teóricas que lo sustenten. Sin embargo, su aplicación resulta ser sencilla y extremadamente flexible lo que lo ha hecho muy popular entre los analistas que trabajan con imputaciones.

V. Conclusiones

En este trabajo, presentamos distintas técnicas que permiten enfrentar el problema de información faltante en encuestas microeconómicas. En particular, dimos una breve revisión de los métodos de imputación no aleatoria, los cuales reducen la varianza estimada; y los métodos de imputación aleatoria, los que permiten mantener la incertidumbre asociada al manejo de la información faltante pero requieren el manejo de múltiples bases de datos.

La discusión fue ilustrada con un ejercicio empírico sencillo donde el problema de información faltante era reducido y los patrones de información eran monótonos. Sin embargo, ante una mayor tasa de información faltante el problema puede tornarse extremadamente complejo y los distintos métodos de imputación pueden entregar resultados disímiles. Por otro lado, el levantamiento del supuesto de patrones de información monótonos genera complicaciones técnicas adicionales que se derivan en un uso intensivo del computador.

Referencias

Acock, A. (2005) “Working With Missing Values” *Journal of Marriage and Family* 67: 1012-1028.

Allison, P. (2002) *Missing Data*, Quantitative Applications in the Social Sciences, A Sage University Papers Series.

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.

Little, R. y D. Rubin (2002) *Statistical Analysis with Missing Data*, Second Edition J. Wiley & Sons, New York.

Mincer, J. (1974) *Schooling, Experience, and Earnings*, Gregg Revivals.

Royston, P. (2005) “Multiple imputation of missing values: update” *Stata Journal* 5(2): 188-201.

Rubin, D. (1987) Multiple Imputation for Nonresponse in Surveys, J. Wiley & Sons, New York.

Sapelli, C. (2003) “Ecuaciones de Mincer y las Tasas de Retorno a la Educación en Chile: 1990-1998” Documento de Trabajo N° 254, Instituto de Economía, Pontificia Universidad Católica de Chile.

Schafer, J. (1997) *Analysis of Incomplete Multivariate Data*, Chapman & Hall/CRC.

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.

Apéndice

1. Derivación de la varianza estimada con datos imputados por el promedio

Sean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ la media y $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ la varianza estimada utilizando solo los *n* datos observados.

Luego, el promedio y la varianza estimada de la variable cuando se reemplaza la información faltante con el promedio son como sigue:

$$\bar{x}_e = \frac{\sum_{i=1}^{n} x_i + (N-n)\bar{x}}{N} = \frac{n}{N}\bar{x} + \left(\frac{N-n}{N}\right)\bar{x} = \bar{x}$$

Es decir el promedio de la muestra extendida que incluye las *N* - *n* observaciones imputadas utilizando el promedio tiene el mismo promedio que los datos originales.

Para el caso de la varianza estimada notamos que los datos imputados utilizando el promedio no aportan en varianza:

$$s_e^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i - \bar{x}_e)^2 + (N-n)(\bar{x} - \bar{x}_e)^2\right]$$

Con ello, tenemos que la varianza estimada de la muestra extendida es:

$$s_e^2 = \frac{1}{N-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\frac{n-1}{N-1}\right)s^2$$

Dado que por definición *N* es menor que *n*, la varianza estimada de la muestra extendida es menor que la varianza de la muestra con información incompleta.

2. Código en Stata

Programa

Los argumentos del programa son la posición (k), la matriz (G) y la dirección. Si esta última es positiva realiza *sweep*, sino realiza *reverse-sweep*.

```
program swp
    version 9
    args k G d
    local r =rowsof(‘G')
    tempname H
    mat ‘H' = ‘G'
    local d=sign(‘d')
    forv j=1/‘r' {
        forv l=1/‘r' {
            mat ‘H'[‘j',‘l'] = ‘G'[‘j',‘l'] ///
                - ‘G'[‘j',‘k']*‘G'[‘k',‘l']/‘G'[‘k',‘k']
        }
    }

    forv j=1/‘r' {
        mat ‘H’[‘j',‘k'] = ‘d'*‘G'[‘j',‘k']/‘G'[‘k',‘k']
        mat ‘H’[‘k',‘j'] = ‘d'*‘G'[‘j',‘k']/‘G'[‘k',‘k']
    }
    mat ‘H'[‘k',‘k'] = -1/‘G'[‘k',‘k']
    mat ‘G'=‘H'
End
```

Aplicación

Cargado el programa anterior las matrices A y B pueden ser generadas con el siguiente código. Notamos que la función mat acc genera los productos cruzados de las variables.

```
gen double e0=1
mat acc T = e0 genero educ exp exp2, noc
mat T=T/r(N)
mat A1=T
swp 1 A1 1
mat acc T=e0 genero educ exp exp2 ling, noc
mat T=T/r(N)
mat A2=T
swp 1 A2 1
mat B1=A1
mat B2=A2

forv i=2/5 {
    swp ‘i' B1 1
    swp ‘i' B2 1
}
mat C1=B2[1..5,6]
mat C2=B2[6,1..6]
mat B3=B1,C1
mat B3=B3\C2
mat A3=B3
forv i=2/5 {
    swp ‘i' A3 -1
}
```

Estudios Económicos Estadísticos
Banco Central de Chile

Studies in Economic Statistics
Central Bank of Chile

NÚMEROS ANTERIORES

PAST ISSUES

Los Estudios Económicos Estadísticos en versión PDF pueden consultarse en la página en Internet del Banco Central www.bcentral.cl. El precio de la copia impresa es de $500 dentro de Chile y US$12 al extranjero. Las solicitudes se pueden hacer por fax al: (56-2) 6702231 o por correo electrónico a: bcch@bcentral.cl

Studies in Economic Statistics in PDF format can be downloaded free of charge from the website www.bcentral.cl. Separate printed versions can be ordered at a price of Ch$500, or US$12 from overseas. Orders can be placed by fax: (56-2) 6702231 or email: bcch@bcentral.cl

SEE-66 Septiembre 2008
Consistencia Transversal en Cuentas Nacionales:
Métodos de Reconciliación a través de Técnicas de Optimización
Gerardo Aceituno Puga

SEE-65 Junio 2008
Inversión por Actividad Económica en Chile.
Período 2004-2005
Claudia Henriquez G.

SEE-64 Junio 2008
Índice de Avisos de Empleo
Marcus Cobb C. y Andrea Sánchez Y.

SEE-63 Abril 2008
Stock de Capital en Chile (1985-2005):
Metodología y Resultados
Claudia Henriquez G.

SEE-62 Diciembre 2007
Flujos de inversión de cartera hacia economías emergentes:
Caracterización de eventos de turbulencia
Karol Fernández Delgado

SEE-61 Diciembre 2007
Efecto de la Sustitución de Combustibles en el Valor Agregado de la Generación Eléctrica
Carmen Gloria Escobar y Marcelo Méndez

SEE-60 Julio 2007
Efectos de Valoración en la Posición de Inversión Internacional de España
Arturo Macías y Álvaro Nash

SEE-59 Julio 2007
Metodología de Cálculo de Índices de Valor Unitario de Exportaciones e Importaciones de Bienes
María Isabel Méndez

SEE-58 Julio 2007
Contenido de Importaciones en las Exportaciones Chilenas 1986-2005; Análisis de Insumo Producto
Claudia Henríquez G. y José Venegas M.

SEE-57 Abril 2007
Metodología de la Encuesta sobre Condiciones Generales y Estándares en el Mercado de Crédito Bancario
Alejandro Jara y Carmen Gloria Silva

SEE-56 Abril 2007
Mercados de Derivados: Swap de Tasas Promedio Cámara y Seguro Inflación
Felipe R. Varela Gana

SEE-55 Marzo 2007
Empalme del PIB y de los Componentes del Gasto: Series Anuales y Trimestrales 1986-2002, Base 2003
Michael Stanger V.