# Imputación Múltiple de la Encuesta Longitudinal de Protección Social

## I. Introducción

En este documento se describen sintéticamente los procedimientos realizados para proceder con la imputación de un conjunto de datos faltantes en la Primera Ola de la Encuesta Longitudinal de Protección Social. Adicionalmente, se incluye una breve guía para el uso de la base de datos generada utilizando el programa Stata

El manejo de la información faltante (*missing data*) propone un desafío estadístico importante para el uso de encuestas.

Esta nota tiene como objetivo documentar el uso del método de imputación múltiple propuesto por Rubin[4] para resolver el problema de datos faltantes en la Encuesta Longitudinal de Protección Social (ELPS) de Uruguay.

Utilizando el algoritmo Full Conditional Specification[5], se realizó la imputación de un conjunto de variables, entre ellas los salarios, los ingresos de la seguridad social y el ingreso total de los hogares.

## II. Análisis estadístico con datos faltantes

Los mecanismos ad-hoc utilizados habitualmente para resolver el problema de los datos faltantes pueden introducir diversos problemas en el análisis de la información.

Omitir las observaciones con datos faltantes (**complete case analysis**) implica descartar información que resulta costoso relevar. Además, desde el punto de vista formal, implica mayor incertidumbre en las estimaciones. Finalmente es posible que los casos completos no sean representativos de la población de interés, lo que resultará en estimaciones sesgadas.

Otra solución de uso habitual es completar las variables con datos faltantes con sus respectivas medias (**fill-in with means**). Si bien este método no altera la media de la variable de interés, si reduce su varianza, y puede distorsionar correlaciones entre las variables de la muestra.

La **imputación estocástica** implica especificar un modelo de imputación para la variable con datos faltantes, y luego realizar la predicción de este modelo incluyendo un componente aleatorio. Esto permite preservar los momentos condicionales de primer y segundo orden en la variable imputada. Sin embargo, implica suponer que los datos imputados se observan con el mismo grado de certidumbre que los efectivamente observados.

El método de imputación múltiple (**Multiple Imputation**) propuesto por Rubin surge como una solución a este problema, obteniendo múltiples realizaciones de la distribución condicional de la variable imputada.

La cantidad de imputaciones (M) ha sido diversa en los trabajos que utilizan imputación múltiple. Sin embargo, hay un cierto consenso[6] respecto a que M=5 es una alternativa válida.

## III. Imputación Múltiple

El objetivo del procedimiento de imputación no es recrear los datos faltantes individualmente sino realizar inferencia estadística válida en presencia de datos faltantes.

El método de imputación múltiple consiste en realizar varias imputaciones de los datos faltantes, de forma de tener en cuenta la incertidumbre sobre esos datos a la hora de realizar inferencia.

Para hacer estimación e inferencia con los datos imputadas, se estiman los parámetros en las *M* bases imputadas y luego se combinan los resultados de acuerdo a las reglas propuestas por Rubin[4]. En el caso de la estimación de un estadístico Q de un escalar, se calcula:

$$H = \frac{1}{M}\sum_{i=1}^{M} Q_i$$

$$W = \frac{1}{M}\sum_{i=1}^{M} V_i^2 + \Big(1+\frac{1}{M}\Big)\Big(\frac{1}{M-1}\sum_{i=1}^{M}(Q_i - H)^2\Big)$$

donde $Q_i$ y $V_i$ son el estadístico y su desvío estándar en la imputación *i* y H y W serán las estimaciones finales.

La imputación de la ELPS fue realizada con Stata. Este paquete incluye programas que simplifican considerablemente la implementación del proceso de imputación múltiple y análisis de muestras imputadas con este método, ajustando los errores estándar de las estimaciones apropiadamente. En la sección final, ilustramos el uso de estos comandos con un ejemplo simple.

Un supuesto básico de trabajo es que los datos faltantes son Missing at Random (**MAR**). Este supuesto implica que, si bien el mecanismo que genera los datos faltantes no es completamente aleatorio, al condicionar por las variables observadas si lo es.

Este es un supuesto de trabajo habitual en las aplicaciones prácticas e implica que no es necesario modelar el mecanismo que produce los datos faltantes[1].

## IV. Imputación múltiple mediante FCS

El procedimiento FCS (Full Conditional Specification), también conocido como ICE (Imputation Using Chained Equations) requiere especificar un modelo de imputación para cada una de las variables a imputar.

FCS es un algoritmo iterativo mediante el cual es posible obtener realizaciones de la distribución de probabilidad $f(\mathbf{Y}_M, \theta|\mathbf{Y}_O)$, donde $\mathbf{Y}_M$ son los datos faltantes, $\mathbf{Y}_O$ los datos observados y $\theta$ los parámetros del modelo de imputación.

Al inicializar el algoritmo, para cada variable se completan los datos faltantes sorteando aleatoriamente uno de los datos observados de esa variable.

Luego, en cada iteración se estiman (en dos etapas) los parámetros de los modelos de imputación y las predicciones para $\mathbf{Y}_M$. Las sucesivas pseudo-realizaciones de estas variables conforman una cadena de Markov. Muestreando esta cadena de Markov obtenemos valores de la distribución de probabilidad $f(\mathbf{Y}_M|\mathbf{Y}_O)$

Es importante tener en cuenta que en las primeras iteraciones del algoritmo los valores de la cadena van a estar altamente correlacionados con el valor incial. Por lo tanto, se descartan las primeras realizaciones del algoritmo de forma que la cadena converja a su distribución límite.

Dado que aún no hay consenso sobre tests formales para determinar la convergencia de las cadenas a su distribución límite, es habitual que en la práctica se inspeccionen los gráficos de las variables imputadas para verificar que no exhiben tendencia.

## V. Imputación ELPS

Las variables imputadas fueron las siguientes: ingresos salariales (e75, e75a, e78, e78a, e81, e81a, e85), jubilaciones (h5_1, h5_2, h5_3), pensiones (h22_1, h22_2), ingresos por asignaciones familiares (d7), tarjeta alimentaria (d13) y pensiones alimenticias (a10a) e ingreso del hogar (y1).

El siguiente cuadro muestra la estadística descriptiva de los valores observados para estas variables:

TABLA I
ESTADÍSTICOS DESCRIPTIVOS

| Variable | Obs | Media | E.E | Mín | Máx |
|---|---|---|---|---|---|
| e75 | 5652 | 19001.8 | 17233.89 | 1000 | 350000 |
| e75a | 6060 | 14097.17 | 11367.08 | 1000 | 175000 |
| e78 | 193 | 30739.98 | 32196.71 | 0 | 220000 |
| e78a | 1629 | 14774.26 | 26121.33 | 1000 | 800000 |
| e81 | 500 | 10857.96 | 11229.16 | 400 | 100000 |
| e81a | 546 | 7981.806 | 7457.057 | 200 | 70000 |
| e85 | 56 | 6222.143 | 8156.983 | 0 | 50000 |
| h5_1 | 3788 | 11660.29 | 10169.69 | 323 | 120000 |
| h5_2 | 158 | 12937.74 | 12116.45 | 440 | 65000 |
| h5_3 | 8 | 8831.25 | 5745.74 | 2650 | 19000 |
| h22_1 | 2456 | 6302.093 | 6767.228 | 100 | 131700 |
| h22_2 | 181 | 5638.293 | 4925.197 | 400 | 27000 |
| h22_3 | 9 | 5834.333 | 6206.645 | 2095 | 22000 |
| d7 | 2241 | 1386.585 | 891.7469 | 200 | 19600 |
| d13 | 1032 | 1356.808 | 1318.2 | 70 | 30000 |
| a10a | 611 | 4889.049 | 5141.904 | 161 | 45000 |
| y1 | 14301 | 22072.59 | 30071.12 | 700 | 1400000 |

Adicionalmente se imputaron variables incompletas usadas en los modelos de imputación: edad del entrevistado, cantidad de personas en el hogar y cantidad de hijos del entrevistado.

Dado que los modelos de imputación presentan mejor ajuste con la transformación logarítmica, se procedió a realizar esta transformación a las variables monetarias. Una vez realizada la imputación de las variables en logaritmos, se deshace la transformación para recuperar el valor imputado de la variable en niveles.

Esto implica que las variables en niveles son variables pasivas (funciones de variables imputadas). Los nombres de las variables en logaritmos comienzan con el prefijo ”ln” (ln_e75, ln_e75a) y los nombres de las variables en nivel terminan con el sufijo ”_imputada” (e75_imputada, e75a_imputada).

## VI. Flags

Para cada variable imputada se genera una variable bandera (”flag”) que indica las siguientes situaciones:

TABLA II
BANDERAS PARA VARIABLES IMPUTADAS

| | |
|---|---|
| 0 | True Missing |
| 1 | Respuesta válida |
| 2049 | Faltante por previa |
| 2050 | Valores cambiados por la crítica |
| 2051 | No Sabe |
| 2052 | No Contesta |
| 2054 | Ajuste post-imputación |

La bandera vale 0 para el caso en que no corresponde que el entrevistado conteste esa pregunta (ej: valor del salario para los que no trabajan). Además, hay valores que se cambian al momento de la crítica y valores faltantes debido a que el entrevistado no contestó una pregunta previa que implica llegar a la pregunta (ej: si no contestó si cobra jubilación, el ingreso por jubilación lleva esta bandera).

Los valores imputados son aquellos que tienen flag 2049, 2050, 2051 o 2052. La siguiente tabla muestra la cantidad de datos observados y faltantes para cada variable imputada. La imputación de otras variables de ingreso (descuentos sobre salarios, salarios en especie, etc.) no fue posible debido a la baja relación entre datos observados y faltantes.

TABLA III
VALORES OBSERVADOS E IMPUTADOS POR VARIABLE

| Variable | Observados | Imputados |
|---|---|---|
| e75 | 5.652 | 1.550 |
| e75a | 6.060 | 1.136 |
| e78 | 193 | 72 |
| e78a | 1.629 | 376 |
| e81 | 509 | 187 |
| e81a | 560 | 141 |
| e85 | 56 | 14 |
| h5_1 | 3.788 | 450 |
| h5_2 | 158 | 49 |
| h5_3 | 8 | 3 |
| h22_1 | 2.456 | 257 |
| h22_2 | 181 | 21 |
| d7 | 2.241 | 172 |
| d13 | 1.032 | 88 |
| a10a | 611 | 117 |

En el caso del ingreso del hogar, el cuestionario incluye dos preguntas, una del valor puntual del ingreso (y1) y otro del ingreso por tramos (y2). La siguiente tabla resume el valor de los flags para y1.

TABLA IV
BANDERAS PARA EL INGRESO DEL HOGAR

| Bandera | Obs. |
|---|---|
| 0 | 2,747 |
| 1 | 14,301 |
| 2049 | 124 |
| 2051 | 739 |
| 2052 | 523 |

Dado que no hay valores *true missing* para el ingreso del hogar, para evitar crear un valor de bandera específica para el

caso en que responden por tramos y no el valor puntual, se asignó el valor 0.

Como veremos más adelante, en el caso en que el entrevistado contesta y2, usamos esta información para imputar el valor de y1.

## VII. Modelos de Imputación

Como mencionamos anteriormente, el método de imputación mediante ecuaciones encadenadas requiere un modelo de imputación para cada variable con datos faltantes. Para todas las variables monetarias se usó la transformación logarítmica.

Para los modelos de los ingresos salariales, se usaron como regresores la educación del entrevistado, la antiguedad en su puesto de trabajo, su edad, sexo, raza, si tiene hijos, el tamaño del hogar y variables geográficas.

En el caso de las jubilaciones y pensiones, se utilizaron el tipo de jubilación, la categoría, la organización que paga la jubilación, la educación del entrevistado, la cantidad de integrantes del hogar así como variables sociodemográficas y geográficas.

Para las pensiones alimenticias se incluyeron la cantidad de hijos en el hogar que no son hijos de uno de los cónyuges

Para los ingresos por AFAM y tarjeta alimentaria se incluyeron variables sociodemográficas, el tamaño del hogar, si el hogar se encuentra en un asentamiento irregular

Finalmente, para imputar los ingresos del hogar, se usó el nivel de educación de todos los integrantes, y varios indicadores socio-demográficos (asentamiento, tenencia de vehículo, de computadora y raza).

Una de los principales ventajas del método FCS es que permite especificar formas funcionales específicas para cada variable. De esta forma, es posible especificar modelos distintos según la naturaleza de la variable.

Para utilizar la información de la variable y2 en la imputación del ingreso del hogar, especificamos una regresión por intervalo usando como límites los tramos de la pregunta y2.

Como notamos anteriormente, es importante verificar la convergencia de las cadenas construidas por el algoritmo FCS. Los gráficos de convergencia se encuentran en el Anexo II. Si bien la mayoría no muestra problemas, en el caso de la imputación de salarios se aprecian posibles problemas de autocorrelación. Para mitigar este problema y mejorar las imputaciones, es recomendable mejorar la especificación de los modelos de imputación agregando variables relativas a la historia laboral de los individuos.

## VIII. Comandos de Stata para datos imputados

La mayor parte de los comandos básicos de Stata tienen una versión MI, que permite tomar en cuenta que la base de datos contiene datos imputados. En esta sección describimos los principales comandos de Stata para trabajar con datos imputados.

Como ilustración, el Anexo III contiene una sesión de trabajo con la ELPS imputada. Al transcribir código de Stata, seguimos la convención de los manuales e incluimos los comandos a utilizar luego del punto (.).

Stata tiene tres formatos de archivo para trabajar datos con múltiples imputaciones. La ELPS se presenta con dos mlong y flongsep. El formato mlong incluye toda la información en un solo archivo. El formato flongsep guarda una imputación en cada archivo. Como se detalla en el anexo, para trabajar los archivos en formato flongsep, es necesario que los archivos con todas las imputaciones esten en el directorio de trabajo de Stata. Este directorio se puede especificar con el comando *cd*.

Stata agrega tres variables de sistema a la base de datos imputados: _mi_id , _mi_miss, _mi_m. Estas variables corresponden al identificador de cada observación, un indicador de si la observación está completa y el número de imputación al que corresponde cada observación (0 para la base original) respectivamente. Estas variables no deben ser modificadas directamente por el usuario.

El comando *mi describe* reporta la estructura de los datos, y la cantidad de variables imputadas. Las variables pasivas son transformaciones de las imputadas, por lo que varían entre imputaciones, pero no son imputadas.

Para abrir los datos en formato mlong:

```
.use "/ELPS/datos/mlong/ELPS_microdatos.dta"
```

Para abrir los datos en formato flongsep:

```
.cd "/ELPS/datos/flong"
.use "ELPS_microdatos
```

Para consultar la estructura de los datos:

```
.mi describe
```

Para obtener estadísticas descriptivas por imputación, usamos *mi xeq*.

```
.mi xeq 0 1 3: summarize log_y1
```

Para eliminar las imputaciones y recuperar los datos originales:

```
.mi set M=0
```

Para hacer estimaciones con los errores estándar corregidos por las reglas de Rubin usamos mi estimate. El siguiente comando estima la regresión del ingreso del hogar contra varios regresores disponibles en la base de datos:

```
mi estimate, ni(5): regress log_y1 edad
    edad_cuad aniosed afro hombre asentam
```

La opción *ni* permite especificar el número de imputaciones a incluir en el modelo.

El comando *mi test* permite hacer pruebas de hipótesis sobre los coeficientes de un modelo. El siguiente obtiene el test de significación conjunta que aparece en la salida de la regresión:

```
.mi test edad edad_cuad aniosed afro
        hombre asentam
```

Para hacer tests sobre un sub conjunto de parámetros:

```
.mi test edad edad_cuad
```

## IX. Comentarios Finales

En este documento se explica cómo fue utilizado el método FCS (Full Conditional Specification) para imputar un conjunto de variables monetarias de la FCS. También se informan los grupos de variables que fueron utilizados en los distintos modelos de imputación. Adicionalmente, se brindan ejemplos simples de cómo realizar estimaciones utilizando la base de datos imputada y el programa Stata 13.

El procedimiento utilizado es un procedimiento iterativo, que busca obtener soluciones que impliquen que las imputaciones realizadas no dependan de los valores a los que se inicializan las variables. Para el conjunto de las variables imputadas se obtuvo la convergencia de este proceso iterativo. No obstante, las pruebas de desempeño indican que sería deseable mejorar los modelos utilizados para imputar las variables vinculadas a los ingresos salariales.

## Referencias

[1] James R.Carpenter, Michael G. Kenward (2013) *Multiple Imputation and its Application.*, p.21, Wiley.
[2] StataCorp (2009), *Multiple Imputation Reference Manual*, StataPress.
[3] Cristina Barceló (2006), *Imputation of the 2002 wave of the Spanish survey of household finances (EFF)*, Documentos Ocasionales N° 0603, Banco de España.
[4] Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.
[5] van Buuren, S. (2007), *Multiple imputation of discrete and continuous data by fully conditional specification*, Statistical Methods in Medical Research 16: 219–242.
[6] Allison, P. (2001), *Missing Data, Quantitative Applications in the Social Sciences*, A Sage University Papers Series