Imputación múltiple de ingresos individuales y familiares en la encuesta CAF 2017

Presentación, Métodos y Ejemplos

Emiro Molina – diciembre 2018

ÍNDICE GENERAL

Imputación Múltiple de los ingresos individuales y Familiares en la encuesta CAF 2017

Presentación, Métodos y Ejemplos.

1 INTRODUCCIÓN

Es común en las encuestas que indagan temas socio económicos encontrar porcentajes importantes de variables con observaciones sin respuesta, particularmente en el caso de las declaraciones de ingresos de las personas o los hogares. La encuesta CAF no escapa de este dilema. Los informantes pueden alegar no recordar el valor de sus ingresos o los de sus familiares o simplemente negarse a declarar los mismos. La información puede haberse perdido por problemas de campo, errores de transcripción, etc. Nos referiremos a estas observaciones como "valores perdidos" y las distinguiremos de los casos en los que no es posible realizar la entrevista, en cuyo caso es usual sustituir el hogar seleccionado en la muestra.

Al realizar análisis con variables que presentan datos perdidos, el ignorar esta información nos podría llevar a incurrir en sesgos de observación. El problema se agrava toda vez que al invocar procedimientos analíticos que incluyan una variable con valores perdidos, la mayoría de los programas para análisis de datos excluirán del análisis los valores declarados por el informante en el resto de las variables invocadas en el procedimiento ("listwise deletion" en inglés). Por ejemplo, en el caso de Argentina deberíamos contar con 679 observaciones para las cuales aplica la pregunta p58a (personas que devengan salarios, un 67% de toda la muestra). Sin embargo, la base de datos presenta 231 valores perdidos para la pregunta entre estas observaciones (un 34% de observaciones perdidas). Si quisiéramos ajustar, e.g., una regresión que incluya esta pregunta, la mayoría de los programas estadísticos, incluido Stata, utilizarán a lo sumo 448 observaciones. La razón es que estos programas eliminan las filas con valores perdidos en las variables invocadas para correr el algoritmo. En la práctica esto equivale a tener una encuesta con menos observaciones en comparación con la encuesta originalmente planificada. En nuestro caso, es equivalente a perder 231 observaciones de la muestra para el caso de Argentina.

Muchas veces se asume que la información se ha perdido completamente al azar (Missing Completely at Random: **MCAR**; ver, por ejemplo, Schafer y Graham, 2002) de modo que la perdida de los valores no es condicional a las otras variables de interés. Esta suposición es equivalente a sustraer al azar observaciones de la base de datos. Sin embargo, suponer que los datos se han perdido completamente al azar suele ser poco realista, sobre todo en el caso de la declaración de ingresos. La alternativa más utilizada es suponer que la pérdida de valores, aunque al azar, depende de otras variables en la muestra (Missing at Random: **MAR**, Schaffer y Graham, 2002[1]). Para ilustrar estas ideas, tomemos la pregunta p58a de la edición 2017 de la encuesta: "¿Cuál es su ingreso mensual neto (de bolsillo) habitual del trabajo principal que usted realiza?". En el caso de Argentina tenemos un 34% de valores perdidos entre las respuestas para las que la pregunta aplica (informantes trabajando). Creando una variable binaria que identifique las observaciones perdidas y realizando una regresión logística sobre algunas variables relacionadas con el hábitat del

[1] Como Schaffer y Graham (2002) subrayan, la escogencia de la terminología MCAR, MAR, MNAR por parte de Rubin (1967) es infeliz ya que induce en el lector nociones de *causalidad*, posiblemente por el uso común del término *azar* que puede sugerir procesos impredecibles o ajenos a las variables presentes en el estudio. Por el contrario, estas definiciones describen *relaciones estadísticas* entre los datos y los mecanismos de pérdida de información que son modelados *probabilísticamente*, no son considerados determinísticos.

individuo (apéndice 1) encontramos que los chances de obtener un valor perdido en los ingresos del informante son 2.5 veces mayores entre los que viven en apartamentos que entre los que viven en casas. Estos chances también se incrementan con el número de habitaciones en el hogar. Los chances se duplican entre aquellos que están conectados a las redes públicas de desagüe. Excluir la información sobre las otras variables que aportan los respondientes con valores perdidos en el ingreso puede sesgar nuestros análisis; la suposición MAR es más adecuada que la MCAR en nuestro caso. Al excluir las 231 observaciones podríamos estar excluyendo subgrupos de interés para el análisis.

Para aliviar esta situación, la Encuesta CAF 2017 incluye Imputaciones Múltiples de las variables relativas a los ingresos múltiples y los ingresos del hogar. El método, propuesto por Rubin (1978, 1987), se ha convertido en el procedimiento estándar para el tratamiento de observaciones perdidas en el caso de encuestas cuyos datos se hacen públicos. En esos casos es conveniente que la agencia que produce los datos entregue las imputaciones (National Center for Health Statistics, 2018; StataCorp LP. 2015; Schafer, 1997). Ello permite incorporar información que muchas veces solo es conocida por dichas agencias, produciendo un tratamiento consistente del tratamiento de la información perdida y análisis consistentes entre diferentes usuarios.

La base de datos de la encuesta es una base en formato estándar de Stata a la cual se han agregado 21 variables correspondientes a 10 imputaciones múltiples de los ingresos individuales, 10 imputaciones múltiples de los ingresos del hogar y una variable con pesos corregidos para no-respuesta. Los nombres de las variables agregadas y las instrucciones para su uso se encuentran en la sección 5 de este documento: *Estructura de la base y ejemplos*.

Este documento se divide en seis secciones. Las secciones 2 describe la incidencia de valores perdidos en la edición 2017 de la ECAF; la sección 3 presenta una descripción panorámica de las imputaciones múltiples para el tratamiento de valores perdidos; las secciones 4 y 5 detallan el procedimiento particular adoptado para la imputación y el uso de la base en Stata, con algunos ejemplos.

2 INCIDENCIA DE VALORES PERDIDOS EN LA ECAF 2017

La edición 2017 de la encuesta CAF presenta globalmente un 19% de valores perdidos en la pregunta de ingresos del informante (p58a) y un 26.8% para el caso de los ingresos del hogar (preguntas p59a y p98a). Sin embargo esta situación varía considerablemente de país a país. La tabla 1 resume la situación.

Tabla1. Porcentaje de valores perdidos según país y tipo de ingresos

	Porcentaje de valores perdidos		**Porcentaje de valores perdidos para los que se declara un intervalo**	
País	Ingresos Individuales	Ingresos del hogar	Ingresos Individuales	Ingresos del hogar
Argentina	34.0	45.7	27.7	23.6
Bolivia	10.7	18.3	27.6	31.2
Brasil	8.5	13.5	25.6	26.7
Chile	13.6	18.8	35.4	34.0
Colombia	17.9	25.1	32.8	31.9
Ecuador	6.2	10.8	60.0	46.3
Mexico	51.0	52.7	39.4	38.1
Panamá	8.4	15.9	48.3	44.3

Perú	8.8	18.9	58.0	54.0
Uruguay	14.1	20.0	50.9	35.6
Venezuela	35.4	49.9	61.3	56.7
Todos	19.0	26.8	41.5	38.6

México, Venezuela y Argentina –en ese orden de importancia– presentan porcentajes de información perdida considerablemente mayores que el resto de los países. Ecuador presenta los valores más bajos. Adicionalmente la tabla presenta para cada variable el porcentaje de informantes que declararon un intervalo de ingresos en lugar de los ingresos mismos. Por ejemplo, en el caso de México el 39.4% de los que no declararon su ingreso laboral, declararon un intervalo en el cual se encuentra dicho ingreso. Globalmente, el 41.5% del 19% que no declaró ingresos laborales, declaró un rango de ingresos. Ello se traduce en una *reducción* de la información perdida. Esta información ha sido incorporada al realizar las imputaciones.

3 IMPUTACIONES

Como se menciona en la sección anterior, la ECAF 2017 incluye Imputaciones Múltiples para aliviar el problema de la información perdida/no-declarada sobre los ingresos personales y los ingresos del hogar en la encuesta CAF 2017. La Imputación Múltiple es una aplicación de los métodos de Monte Carlo descrita por Rubin (1987) en el contexto de la falta de respuestas en encuestas de dominio público. El método se ha convertido en un estándar para el manejo de la información perdida en encuestas de largo alcance con bases de datos públicas. Algunos ejemplos son: The Cancer Care Outcomes Research and Surveillance (CanCORS) Consortium (Ayanian *et al*, 2003), The Fatal Accident Reporting System (Heitjan and Little, 1991), The Census Industry and Occupation Codes (Schenker, Treiman and Weidman,1993), The National Health and Nutrition Examination Survey (Schafer, 1997, sec. 6.4), The National Health Interview Survey (National Center for Health Statistics, 2018), The Survey of Consumer Finance (Kennickell, 2017), entre otros. Varias razones favorecen el uso de tales imputaciones: la posibilidad de ajustar las diferencias entre respondientes y no-respondientes, la posibilidad de utilizar software estadístico estándar sin perder valores observados en otras variables del estudio; además, cuando las imputaciones son realizadas por la agencia que produce los datos, existe la posibilidad de incorporar información que solo se maneja al nivel de producción de las bases de datos y se garantiza que el problema de la información perdida sea tratada de manera consistente por los usuarios de las bases de datos.

La imputación múltiple además permite incorporar al análisis el hecho de que los valores sustituidos son un conjunto de valores *plausibles* para las variables imputadas, no son los valores realmente perdidos. Ello produce una incertidumbre que es preciso introducir en los análisis. En el pasado, al utilizar imputaciones simples, se producían errores estándares e intervalos de cobertura muy pequeños, lo que llevaba al aumento del rechazo de hipótesis nulas aun siendo ciertas con una frecuencia mayor que la anunciada por el valor nominal de las pruebas estadísticas (En un estudio realizado por Rubin y Schenker, 1986, se encontraron intervalos de cobertura reales entre el 85% y 90% para valores nominales del 95% cuando se utilizaba imputación simple con tasas de información perdida entre el 20% y el 30%). La introducción de imputaciones múltiples permite estimar el efecto de imputar las variables e incorporar a los análisis la incertidumbre generada por la pérdida de información y estimar el efecto de la información perdida.

Es importante destacar este último punto. El objetivo de imputar *no es obtener las mejores sustituciones de los datos perdidos*, sino incorporar la información que aportan otras variables y que se pierde al no incorporar la observación con el dato perdido y, evaluar el impacto de esta pérdida de información.

3.1 Imputaciones Múltiples

La Imputación Múltiple sustituye cada valor perdido por una lista de $m > 1$ valores simulados, produciendo en la práctica m bases de datos completas *plausibles* bajo la suposición MAR. Cada una de estas m bases es analizada separadamente con métodos estándares para bases completas, y los resultados obtenidos son combinados utilizando cierto conjunto de reglas dirigidas a obtener estimadores de la incertidumbre generada por el proceso de imputación. Por ejemplo, supongamos que deseamos utilizar una regresión probit para investigar alguna hipótesis sobre los datos. Cada una de estas bases es analizada ajustando la misma regresión probit, y luego los resultados de los m análisis se combinan para obtener estimaciones globales y errores estándares que reflejen la incertidumbre generada por la pérdida de datos además de la variación finita inherente a los datos mismos. En el apéndice 2 se presenta una panorámica del proceso.

3.2 Número de imputaciones

La imputación múltiple no requiere de un gran número de imputaciones para obtener estimaciones adecuadas. Rubin (1987) muestra que la eficiencia relativa de un estimado basado en m imputaciones comparada con la eficiencia obtenida con infinitas imputaciones viene dada por

$$e = (1 + \lambda/m)^{-1},$$

donde λ es la proporción de información perdida de la variable de interés y m es el número de imputaciones. Tomando el caso con mayor fracción de información perdida, los ingresos del hogar en el caso de México ($\lambda = 0.527$) obtenemos una eficiencia relativa del 95% con 10 imputaciones. En general, eficiencias relativas del orden del 90% son adecuadas en la práctica. De hecho, en el caso de bases de datos públicas el número de imputaciones no suele pasar de 5 (National Center for Health Statistics, 2018). Rubin (1996) argumenta que un valor pequeño de m es apropiado porque las simulaciones envueltas en la imputación múltiple sólo tienen como objeto controlar las fracciones de información perdida, mientras que la información observada es manejada por los métodos de análisis utilizados con la base de datos completa. En nuestro caso se decidió publicar 10 imputaciones por variable de ingresos considerada, tomando el país con mayor cantidad de información perdida (México) como referencia para calcular la eficiencia relativa, dada la diversidad observada entre los países estudiados.

3.3 Selección del modelo de Imputación

Para realizar la imputación se postula un modelo que vincule las variables a imputar con otras variables en la encuesta. El modelo de imputación debe incluir los predictores que sean potencialmente relevantes para el mecanismo de pérdida de información. A sí mismo, debe preservar aquellas características que podrían ser incluidas en análisis que envuelvan las variables imputadas, bien sea como variables de respuesta o como predictores. El diseño muestral también debe incluirse en el modelo. En resumen, el modelo debe incluir las variables que potencialmente estén asociadas a las variables a imputar y al mecanismo de pérdida de información de dicha variable. (StataCorp LP, 2015, pag. 8; Schafer, 1997, sec. 4.5.5). Por otra parte, un exceso de variables en el modelo, sobre todo en el caso de variables categóricas, puede crear problemas de estimación de los modelos y/o evitar la convergencia de los mecanismos de imputación, de modo que el modelo debe ser tan parsimonioso como sea posible sin sacrificar información. La lista de variables escogida para realizar las imputaciones en esta entrega se encuentra en el apéndice 3. La exploración de la

bondad de estas variables se realizó país por país, utilizando el proceso de selección LASSO (Least Absolute Shrinkage and Selection Operator, ver, e.g. Bradley *et al*, 2004) con el criterio de información de Akaike (AIC: Akaike, 1973).

3.4 ANÁLISIS DE LOS DATOS IMPUTADOS

El estudio de un estimado poblacional de interés, tal como una media, proporción o un coeficiente en un modelo estadístico que envuelva *m* imputaciones múltiples, puede resumirse en los siguientes pasos:

- Se postula un modelo estadístico para la estimación del coeficiente de interés.
- Se aplica el mismo procedimiento de estimación a cada uno de los *m+1* conjuntos de datos producidos por las imputaciones (*m* imputaciones y los datos originales)
- Los *m+1* coeficientes obtenidos y sus errores estándares se combinan para obtener un estimado, su error estándar, intervalos de cobertura y otras estadísticas pertinentes al análisis.
- La inferencia procede con estos agregados utilizando las distribuciones de referencia resultantes de realizar los agregados.

El estimado puntual del coeficiente se obtiene promediando los estimados puntuales resultantes de las *m* estimaciones individuales. El segundo componente estima la variación entre los *m* estimadores puntuales. Para la construcción de intervalos de confianza y medidas de significancia se invoca una distribución *t* con grados de libertad que dependen del número de imputaciones practicadas. Los detalles técnicos de la implementación en Stata se resumen en el Apéndice 2 de este reporte.

4 PROCEDIMIENTO SEGUIDO PARA LA IMPUTACIÓN DE LA EDICIÓN 2017 DE LA ENCUESTA

Los ingresos de los informantes y los ingresos del hogar presentan una alta correlación en la encuesta, sin embargo los valores perdidos asociados a estas variables no están anidados (que una observación tenga un valor perdido en una de estas variables no necesariamente implica que para esta observación el valor esté perdido en la otra variable; ello puede verificarse utilizando el comando "`misstable nested ingind inghog`"). Ello, aunado al hecho de que algunas de las variables utilizadas como predictores en los modelos de imputación pueden tener valores perdidos, implica que debe evitarse realizar las imputaciones secuencialmente por separado. Por ello se procedió a utilizar el método de imputaciones múltiples por ecuaciones encadenadas, utilizando el procedimiento **mi impute chained** de Stata. El método realiza la imputación de varias variables utilizando una secuencia de imputaciones univariadas iterativamente con especificaciones condicionales en el resto de las variables: las variables a imputar son utilizadas iterativamente como variables de respuesta y predictores en una cadena de modelos secuencialmente, la primera variable utiliza el resto como predictores y los valores son imputados; luego la variable imputada pasa a ser utilizada como predictor de la siguiente variable a imputar y así sucesivamente. El proceso se repite hasta obtener convergencia. Supongamos que las variables Y1,Y2, …, Y_k en la base de datos presentan valores perdidos y han sido ordenadas según el número de valores perdidos, de menos a más. Supongamos que el grupo de predictores **Z** no posee valores perdidos. Supongamos que utilizamos un modelo de imputación *g* (por ejemplo, una regresión simple). Los valores imputados se obtienen de la cadena de modelos

$$Y_1^{(t+1)} \sim g(Y_1|Y_2^{(t)}, \ldots, Y_k^{(t)}, \mathbf{Z}; \varphi_1),$$

$Y_2^{(t+1)} \sim g(Y_2|Y_1^{(t+1)}, Y_3^{(t)}, \ldots, Y_k^{(t)}, \mathbf{Z}; \varphi_2)$,

…

$Y_k^{(t+1)} \sim g(Y_k|Y_1^{(t+1)}, Y_2^{(t+1}, \ldots, Y_{k-1}^{(t+1)}, \mathbf{Z}; \varphi_k)$,

iterando $t = 0, 1, \ldots, T$ veces hasta obtener convergencia en $t = T$; aquí φ_j, $j=1, \ldots, k$ representa los parámetros del modelo g para la variable Y_j. Ver StataCorp (2015, pp. 144-147) para más detalles.

Con el fin de incorporar el diseño muestral, para cada país la imputación se realizó por estrato muestral. En nuestro caso se utilizaron como variables a imputar los *logaritmos naturales* de los ingresos y, dependiendo de los datos de cada país, las variables de naturaleza continua escogidas para el modelo que exhibiesen valores perdidos; aquellas variables que no presentaron valores perdidos entraron en los modelos como Z-variables –en la notación descrita–. En el caso de las variables de naturaleza categórica, para los países donde estas variables presentaron valores perdidos, se trataron dichos valores perdidos como una categoría extra y se utilizaron las variables como variables completas (Z-variables en la notación anterior). Una vez obtenidas las imputaciones para los logaritmos de los ingresos estos fueron exponenciados para obtener los ingresos imputados en su escala natural. En el caso de aquellas observaciones que no declararon el valor de sus ingresos pero declararon un rango de ingresos, el procedimiento primero verificaba que el valor imputado cayese dentro de este rango. En los pocos casos en que el valor imputado cayó fuera de dicho rango, el extremo del intervalo más cercano a la imputación se tomó como valor para la imputación, ya que en ese caso el vecino más cercano encontrado tenía un valor o ligeramente mayor o ligeramente menor que los valores en el intervalo.

Para modelar la asociación entre las variables a imputar y el resto de las variables se utilizaron regresiones (el *g*-modelo en la notación anterior) combinadas con el método de "coincidencias de medias predichas" (predictive mean matching) implementado en Stata con el procedimiento **impute pmm** (StataCorp, 2015, pp. 243-248).

El procedimiento es un método semi-paramétrico que reemplaza el valor faltante con el valor observado más cercano a la predicción lineal correspondiente a dicho valor. Introducido por Little (1988), el método combina la regresión lineal con la imputación por vecino más cercano. Primero se ajusta una regresión lineal para obtener un valor predicho de la observación faltante; luego utiliza una medida de distancia a dicha predicción lineal para escoger un conjunto de vecinos más cercanos dentro de los valores observados (posibles donantes). Finalmente, selecciona aleatoriamente un valor de este conjunto para realizar la imputación.

Al escoger los valores imputados a partir de los datos observados, el método conserva la distribución de estos valores, lo que hace más robusta la selección. Por otra parte el método asegura que el valor imputado es *observable*: el valor predicho por una regresión es un valor plausible dentro del rango observado de los datos, pero no necesariamente puede ser encontrado en la población objeto, por ejemplo, un ingreso intermedio entre dos salarios reales que no exista en la práctica.

La implementación en Stata permite escoger el tamaño del subconjunto de valores más cercanos a la predicción lineal. Para ser eficiente el método requiere que hayan suficientes valores para ser donados en la vecindad. Dada las altas correlaciones intra-conglomerados de los ingresos para cada país (ver apéndice 4) se decidió utilizar subconjuntos de tamaño 3, ya que cada conglomerado en la muestra tiene tamaño 5 en promedio (ver Morris, White, and Royston, 2014, para una discusión sobre el tema).

5 ESTRUCTURA DE LAS BASES ENTREGADAS Y EJEMPLOS

La base de datos de la ECAF 2017 sido aumentada con 21 variables correspondientes a 10 imputaciones múltiples de los ingresos individuales, 10 imputaciones múltiples de los ingresos del hogar y una variable correspondiente a pesos muestrales corregidos para la falta de respuestas en caso de que el analista no desee utilizar las imputaciones.

La base "ecaf2017.dta" es una base en formato estándar de Stata. Las variables aumentadas son:

ingind: Esta es una copia de la variable p58a, los ingresos declarados del entrevistado.

ingind_1, ingind_2, …, ingind_10: Diez (10) imputaciones múltiples de Ingind, mediante el método de ecuaciones encadenadas implementado en Stata ("mi impute chained": StataCorp LP, 2015).

Inghog : Los ingresos declarados del hogar. En la encuesta esta pregunta se encuentra dividida en dos variables correspondientes a las preguntas p59a, los ingresos del hogar de los informantes que tienen un trabajo remunerado y la p98a, los informantes que no devengan salarios.

inghog_1, inghog_2, …, inghog_10: Diez (10) imputaciones múltiples de Inghog, mediante el método de ecuaciones encadenadas implementado en Stata ("mi impute chained": StataCorp LP, 2015).

peso_p58a: ponderaciones muestrales corregidas para los valores perdidos, para el caso en el cual el analista no desee utilizar las imputaciones en análisis que envuelvan los ingresos.

Como se mencionó anteriormente, esta es una base estándar en formato *dta* de Stata. El usuario puede utilizar la base directamente con cualquiera de los comandos de Stata. En la base ya ha sido declarada la estuctura muestral para el uso de los comandos con prefijo "`svy:`"

```
svyset psu_cod [pw= pesoper], strata( strata_mues)
```

En particular el usuario puede:

1. *Importar* la base al formato "mi" de Sstata de su preferencia. Por ejemplo para importar la base al formato *mi wide*:

```
use "ecaf2017_imputaciones-multiples.dta", clear
mi import wide, imputed( ingind= ingind_1 ingind_2 ingind_3 ///
      ingind_4 ingind_5 ingind_6 ingind_7 ingind_8 ingind_9 /// ingind_10
      ///
      inghog= inghog_1 inghog_2 inghog_3 inghog_4 inghog_5 inghog_6 ///
                  inghog_7 inghog_8 inghog_9 inghog_10) drop
save "ecaf2017_mi.dta", replace
```

2. Utilizar una metodología *ad hoc* para tratar los valores perdidos. El usuario puede utilizar otros procedimientos en Stata, e.g. el comando `implogit` (http://www.stata.com/stb/stb45). Por otra parte, cualquiera de las variables **ingind_1, ingind_2, …, ingind_10; inghog_1, inghog_2, …, inghog_10**, pude utilizarse para realizar imputaciones simples, constituyendo cada una de ellas una imputación simple *condicional estocástica*, aunque este procedimiento es inferior al uso de imputaciones múltiples. O el usuario puede exportar la base para utilizar otro software, etc.

3. Eliminar las variables ingind_* e inghog_* si no desea utilizar imputaciones. En ese caso se recomienda enfáticamente utilizar como pesos muestrales la variable **peso_p58a** siempre que los cálculos envuelvan las variable **ingind** (p58a) o **inghog** (que combina las preguntas p59a y p98a). Cuando hay valores perdidos Stata elimina de la matriz de análisis todas las observaciones con valores perdidos en estas variables. Ello puede resultar en una reducción drástica de los datos y en una expansión inadecuada de las observaciones, lo cual inducirá un sesgo de selección en los análisis. Por lo tanto, *si no se realizan imputaciones*, se aconseja cambiar la declaración **svyset** a:

   ```
   svyset psu_cod [pw= peso_p58a], strata( strata_mues)
   ```

   Una vez terminado el análisis que incluye las variables ingind o inghog *debe volver a declararse la estructura SVY con las ponderaciones originales:*

   ```
   svyset psu_cod [pw= pesoper], strata( strata_mues)
   ```

En el caso de usar el primer procedimiento, se creará la base de datos "ecaf2017_mi.dta" preparada para el uso de los comandos "mi" (Multiple Imputation) de Stata (StataCorp LP, 2015), estructurada en el formato "*wide"* (una de las cuatro estructuras de imputaciones múltiples que posibilita Stata: flongsep, flong, mlong, y wide). En el estilo wide, Stata agrega las nuevas variables imputadas como columnas, de manera similar a la presentada en una base estándar de Stata, junto con una variable indicadora de sistema "_mi_miss", que no debe ser eliminada de la base. Los nombres dados por Stata a estas variables no deben ser alterados por el usuario. En nuestro caso las nuevas variables son: _1_ingind, _1_inghog, _2_ingind, _2_inghog,…, _10_ingind _10_inghog. Ello hace más simple la manipulación de estas variables, en contraste con los otros estilos.

Por ejemplo, si queremos un resumen de la variable ingind y su primera y última imputación para el caso de Bolivia podemos utilizar:

```
summarize ingind  _1_ingind  _10_ingind if pais==2

Variable    Obs   Mean  Std. Dev.   Min   Max

ingind      634   2645.066    2065.545    200   20000

_1_ingind   710   2597.972    2009.262    200   20000

_10_ingind  710   2613.313    2106.848    200   20000
```

710 informantes declaran recibir ingresos, pero solo 634 de ellos declararon el valor de estos ingresos. La Imputación Múltiple sustituye cada valor perdido por una lista de ***m*** > *1* variables con valores simulados, produciendo en la práctica ***m*** bases de datos ***plausibles***.

Un comando alternativo que funciona en cualquier estilo (no solo en el wide) es:

```
. mi xeq 0 1 10: summarize ingind if pais==2

m=0 data:
-> summarize ingind if pais==2

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      ingind |        634    2645.066    2065.545        200      20000
```

```
m=1 data:
-> summarize ingind if pais==2

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      ingind |        710    2597.972    2009.262        200      20000

m=10 data:
-> summarize ingind if pais==2

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      ingind |        710    2613.313    2106.848        200      20000
```

Aquí el índice 0 corresponde a la variable sin imputar, el índice 1 a la primera versión de la variable imputada, el índice 2 a la segunda, etc. El comando sin índices:

```
mi xeq: summarize ingind
```

genera un resumen de todas las versiones disponibles de la variable y sus imputaciones.

Supongamos que deseamos utilizar una regresión para investigar alguna hipótesis sobre los datos. Cada una de estas bases es analizada ajustando la misma regresión, y luego los resultados de los *m* análisis se combinan para obtener estimaciones globales y errores estándares que reflejen la incertidumbre generada por la pérdida de datos además de la variación finita inherente a dichos datos. Es importante recordar que el objetivo de imputar es incorporar la información que aportan otras variables y que se pierde al no incorporar la observación con el dato perdido, y, evaluar el impacto de esta pérdida.

Cuando se realiza una transformación de una variable que ha sido imputada, esta trasformación debe realizarse sobre todas sus versiones. La forma más eficiente es con el comando **mi passive**, e.g.

```
mi passive: gen logingre=log(ingind)
```

Este comando además registra la variable automáticamente como pasiva (una variable derivada de una imputada). Las variables en una base mi deben registrase como pasivas (passive), imputadas (imputed), regulares (regular) o no registrarse, en cuyo caso Stata asume que son regulares.

El comando **mi describe** nos revela la estructura de la base si esta se encuentra en formato *mi*, además de indicar las variables imputadas, derivadas, regulares, etc:

```
. use "D:\My_Dir\ecaf2017_mi-wide.dta", clear
.mi describe
  Style:  wide
  Obs.:   complete        7,660
          incomplete      3,027  (M = 10 imputations)
          --------------------
          total          10,687

  Vars.:  imputed:  2; ingind(1247+4110) inghog(2861)

          passive:  0

          regular:  0

          system:   1; _mi_miss
```

```
          (there are 338 unregistered variables)
```

Stata nos indica que tenemos 2 variables imputadas, que no hemos registrado transformaciones de estas variables (passive: 0) y no hemos registrado el resto de las variables.

Ejemplo 1

Este primer ejemplo solo pretende ilustrar algunas de las estadísticas disponibles para el examen del impacto de los datos perdidos en el módulo *mi* de Stata. Para el caso de Colombia examinemos la variable imputada *ingind* (ingresos del informante):

```
. use "D:\My_Dir\ecaf2017_mi.dta", clear
. preserve
. mi estimate: svy: mean ingind if pais ==5

Multiple-imputation estimates     Imputations     =         10
Survey: Mean estimation           Number of obs   =        665

Number of strata  =         6     Population size =  3,135,468
Number of PSUs    =       196
                                  Average RVI     =     0.0745
                                  Largest FMI     =     0.0711
                                  Complete DF     =        190
DF adjustment:   Small sample     DF:     min     =     160.05
                                          avg     =     160.05
Within VCE type:   Linearized             max     =     160.05

------------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
      ingind |    2660790   418670.2       1833960     3487621
------------------------------------------------------------------
```

Obtenemos información sobre el diseño muestral (número de estratos y unidades primarias muestrales), el número de imputaciones utilizadas, el número de observaciones válidas para el análisis (en la base hay 1000 informantes para el caso de Colombia; de ellos 665 reportan estar ocupados, la pregunta no aplica para otros entrevistados), y varias estadísticas relativas al efecto de imputar. `Average RVI` es el incremento promedio relativo de las varianzas por efecto de las no-respuestas, en este caso un 7.45%; `Largest FMI` es la mayor fracción de información perdida asociada a cada coeficiente estimado En este caso solo tenemos un coeficiente (la media poblacional, FMI=7.11%). Este estimado es el promedio de los coeficientes obtenidos realizando la estimación con cada una de las 10 imputaciones. Una explicación más técnica puede verse en el Apéndice 2.

La información de RVI's y FMI's para cada coeficiente la podemos obtener con el comando:

```
. mi estimate, vartable nocitable

Multiple-imputation estimates                   Imputations       =         10
Survey: Mean estimation

Variance information

------------------------------------------------------------------------------
             |        Imputation variance                                Relative
             |    Within   Between     Total       RVI       FMI    efficiency
-------------+----------------------------------------------------------------
      ingind |   1.6e+11   1.1e+10   1.8e+11   .074467   .071096       .992941
------------------------------------------------------------------------------
```

Esta tabla aporta la varianza intra y entre imputaciones; la tercera columna es la suma de las dos primeras ajustadas para un número finito de imputaciones; la cuarte columna presenta el incremento relativo de la varianza debido a la no-respuesta (RVI: el cociente entre las varianzas entre e intra corregido por el factor *(1 + 1/m)*); la quinta columna reporta el incremento en la varianza debido a la perdida de información. La ultima columna es la eficiencia relativa de utilizar ***m*** imputaciones en lugar de un hipotético número infinito de imputaciones. La opción dftable nos presenta una estimación del incremento porcentual en los errores estándares de los estimados debido a la falta de respuestas:

```
. mi estimate, dftable

Multiple-imputation estimates     Imputations     =         10
Survey: Mean estimation           Number of obs   =        665
Number of strata  =         6     Population size =  3,135,468
Number of PSUs    =       196
                                  Average RVI     =     0.0745
                                  Largest FMI     =     0.0711
                                  Complete DF     =        190
DF adjustment:   Small sample     DF:     min     =     160.05
                                          avg     =     160.05
Within VCE type:   Linearized             max     =     160.05

------------------------------------------------------------
             |                                      % Increase
             |       Mean   Std. Err.           DF   Std. Err.
-------------+----------------------------------------------
      ingind |    2660790   418670.2         160.1        3.66
------------------------------------------------------------
```

El error estándar de la estimación se incrementa en un 3.66% por la pérdida de respuestas. De no tomar en cuenta esta pérdida, utilizaríamos errores estándares más pequeños, induciendo p-valores mayores y su consecuente impacto en la inferencia.

Ejemplo 2

Para ilustrar otras posibilidades, estudiaremos la relación entre las respuestas a la pregunta P12 (*En una escala del 1 al 10 donde 1 es "Nada satisfecho" y 10 es "Totalmente satisfecho", ¿qué tan satisfecho está usted con su vivienda?*) y los ingresos laborales del informante. Los ejemplos son ilustrativos de los comandos y no pretenden sugerir la bondad de los modelos utilizados.

La respuesta a la pregunta 12 es una escala ordinal discreta, de modo que utilizaremos una regresión logística ordinal. Supongamos que deseamos explorar el caso de Bolivia y nos restringimos al logaritmo de los ingresos como variable explicativa. La base ya ha sido preparada para los comandos svy por medio del comando: `mi svyset psu_cod [pw= pesoper], strata( strata_mues).` Notar que el comando es "*mi* svyset". Si intentamos utilizar solo el comando "svyset" Stata arrojará un error.

```
.use "D:\My_Dir\ecaf2017_mi.dta", clear
. svyset
       no; data are mi set
       Use mi svyset to set or query these data; mi svyset has the same syntax as
       svyset.
. mi svyset

      pweight: pesoper
          VCE: linearized
  Single unit: missing
     Strata 1: strata_mues
```

```
         SU 1: psu_cod
        FPC 1: <zero>
```

Continuamos con el ejemplo:

```
.preserve

** Generamos el logaritmo natural de los ingresos para los ingresos y sus 10
**imputaciones:

.mi passive: gen logingre=log(ingind)
```

(Omitimos la salida por brevedad. El comando **mi passive** registra la variable como pasiva: una transformación de una variable imputada. La transformación se aplica a la variable original y sus 10 versiones imputadas)

```
.mi update //actualiza la información sobre las variables en la base
.mi estimate: svy: ologit p12 logingre if pais==2 // ajustamos el modelo para Bolivia

Multiple-imputation estimates                   Imputations       =         10
Survey: Ordered logistic regression             Number of obs     =        710
Number of strata  =         6                   Population size   =    668,443
Number of PSUs    =       200
                                                Average RVI       =     0.0142
                                                Largest FMI       =     0.1002
                                                Complete DF       =        194
DF adjustment:   Small sample                   DF:     min       =     146.62
                                                        avg       =     152.95
                                                        max       =     168.28
Model F test:       Equal FMI                   F(   1,  146.6)   =       5.06
Within VCE type:   Linearized                   Prob > F          =     0.0259

------------------------------------------------------------------------------
         p12 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    logingre |   .2718657   .1208102     2.25   0.026     .0331113    .5106201
-------------+----------------------------------------------------------------
       /cut1 |  -1.747904   1.153182    -1.52   0.131    -4.024471    .5286626
       /cut2 |  -1.153061   1.048958    -1.10   0.273    -3.224544    .9184226
       /cut3 |   -.533334   1.002188    -0.53   0.595    -2.512865    1.446197
       /cut4 |   .2665337   .9619686     0.28   0.782    -1.634003     2.16707
       /cut5 |    1.15929   .9496498     1.22   0.224    -.7171217    3.035702
       /cut6 |    1.70425   .9440527     1.81   0.073    -.1612189    3.569719
       /cut7 |   2.356241    .945326     2.49   0.014     .4882183    4.224263
       /cut8 |   3.216855   .9460499     3.40   0.001     1.347362    5.086348
       /cut9 |   3.931607   .9530647     4.13   0.000     2.048317    5.814896
------------------------------------------------------------------------------
```

Nótese el uso del prefijo "**mi estimate:"**. La salida es el resultado de agregar once (11) regresiones ordinales, equivalentes a analizar 11 bases de datos, la original y las 10 bases formadas por cada imputación y las variables derivadas de ellas, y luego agregar estas regresiones (para los detalles, ver la documentación en el manual de referencia de la Imputación Múltiple de Stata, que es muy completo). Obtenemos además información sobre el diseño muestral y varias estadísticas dirigidas a explorar el efecto de imputar.

Con el objetivo de examinar el impacto de la pérdida de observaciones, producimos una tabulación complementaria:

```
. mi estimate, vartable nocitable
```

```
Multiple-imputation estimates                   Imputations       =         10
Survey: Ordered logistic regression

Variance information
------------------------------------------------------------------------------
             |        Imputation variance                               Relative
             |    Within   Between     Total       RVI       FMI    efficiency
-------------+----------------------------------------------------------------
    logingre |   .013175   .001291   .014595   .107778   .100224       .990077
-------------+----------------------------------------------------------------
       /cut1 |   1.24986   .072694   1.32983   .063978   .061578        .99388
       /cut2 |   1.02007   .072947   1.10031   .078663   .074834       .992572
       /cut3 |   .923777   .073276   1.00438   .087255   .082448       .991823
       /cut4 |   .843939   .074041   .925384   .096505   .090532       .991028
       /cut5 |   .819257   .075071   .901835   .100796   .094242       .990664
       /cut6 |   .807934   .075729   .891236   .103105   .096227       .990469
       /cut7 |   .809577   .076422   .893641   .103838   .096856       .990407
       /cut8 |   .810236   .077067    .89501   .104629   .097534       .990341
       /cut9 |   .823254   .077344   .908332   .103345   .096433       .990449
------------------------------------------------------------------------------
```

vartable produce una tabla con información sobre la varianza de los estimados MI. La tabla contiene estimaciones de las varianzas intra-imputaciones, entre-imputaciones, varianzas totales, el incremento relativo de la varianza debido a la no-respuesta (RVI), fracciones de la información perdida en la estimación debida a la no-respuesta (FMI), y la eficiencia relativa de utilizar ***m*** imputaciones en lugar de un hipotético número infinito de imputaciones. Como puede verse, el uso de 10 imputaciones produce eficiencias superiores al 99% para estas estimaciones. Stata produce un número considerable de estadísticas para evaluar el efecto de la información perdida. En particular se deben explorar los comandos de post-estimación del módulo MI de Stata.

El resultado obtenido en la primera tabla respecto al coeficiente de la variable logingre sugiere una asociación creciente entre la satisfacción con la vivienda y los ingresos en el caso de Bolivia (la probabilidad de un incremento en la satisfacción aumenta con los ingresos) . Un resultado similar se obtiene para el caso de Ecuador:

```
. mi estimate: svy: ologit p12 logingre if pais==6

Multiple-imputation estimates                   Imputations       =         10
Survey: Ordered logistic regression             Number of obs     =        640

Number of strata  =         3                   Population size   =    652,384
Number of PSUs    =       200
                                                Average RVI       =     0.0023
                                                Largest FMI       =     0.0209
                                                Complete DF       =        197
DF adjustment:   Small sample                   DF:     min       =     189.35
                                                        avg       =     190.27
                                                        max       =     192.30
Model F test:       Equal FMI                   F(   1,  189.3)   =      13.86
Within VCE type:   Linearized                   Prob > F          =     0.0003

------------------------------------------------------------------------------
         p12 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    logingre |   .4253473   .1142632     3.72   0.000     .1999549    .6507397
-------------+----------------------------------------------------------------
       /cut1 |  -2.395968   .8894148    -2.69   0.008    -4.150229   -.6417059
       /cut2 |  -1.381701   .7553139    -1.83   0.069    -2.871531    .1081298
```

```
       /cut3 |  -.9247831   .7310642    -1.26   0.207      -2.3668    .5172334
       /cut4 |  -.1814069   .7237029    -0.25   0.802    -1.608912    1.246098
       /cut5 |   .6968206   .7070886     0.99   0.326    -.6979306    2.091572
       /cut6 |   1.150114   .7150563     1.61   0.109    -.2603498    2.560579
       /cut7 |     1.7928    .706681     2.54   0.012      .398845    3.186755
       /cut8 |   2.836677   .7023424     4.04   0.000     1.451271    4.222083
       /cut9 |   3.745694   .7013287     5.34   0.000      2.36228    5.129108
------------------------------------------------------------------------------
```

De nuevo el coeficiente correspondiente a logingre muestra un valor positivo significativo, cuyo estimado es mayor que en el caso de Bolivia.

¿Es significativa la diferencia estimada? Para la comparación es deseable que los ingresos se encuentren en escalas similares. Para el objeto de este ejercicio los llevaremos a su valor en USD al momento de la recolección de acuerdo a las tasas de cambio oficiales. Para realizar las comparaciones nos conviene restringir los datos a estos dos países; en el formato mi wide es muy sencillo substraer una base restringida a estos países:

```
.use "D:\My_Dir\ecaf2017_mi.dta", clear

.preserve
. keep if pais ==6 | pais==2
(8,687 observations deleted)
.mi update
.save "D:\My_Dir\Bolivia_Ecuador_mi", replace
.restore
use "D:\My_Dir\Bolivia_Ecuador_mi", clear
```

Ahora procedemos a comparar los coeficientes de interés; nótese el uso de las variables pasivas:

```
. quietly mi passive: gen ing_usd= ingind/6.91 if pais==2
. quietly mi passive: replace ing_usd= ingind if pais==6
. quietly mi passive: gen loging_usd=log(ing_usd)
**(Utilizamos quietly para suprimir las salidas)

. mi estimate, saving(m1, replace): svy: ologit p12 i.pais i.pais#c.loging_usd


Multiple-imputation estimates                   Imputations       =         10
Survey: Ordered logistic regression             Number of obs     =      1,350
Number of strata  =         9                   Population size   =  1,320,827
Number of PSUs    =       400
                                                Average RVI       =     0.0129
                                                Largest FMI       =     0.0972
                                                Complete DF       =        391
DF adjustment:   Small sample                   DF:     min       =     260.30
                                                        avg       =     290.51
                                                        max       =     374.82
Model F test:       Equal FMI                   F(   3,  376.1)   =      30.34
Within VCE type:   Linearized                   Prob > F          =     0.0000

------------------------------------------------------------------------------
             p12 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
            pais |
         Ecuador |  -.0033438   1.010766    -0.00   0.997    -1.991998     1.98531
                 |
pais#c.loging_usd |
         Bolivia |   .2749294   .1244225     2.21   0.028     .0299266    .5199322
```

```
         Ecuador  |   .4147046   .1109087     3.74   0.000     .1966234    .6327858
------------------+------------------------------------------------------------------
            /cut1 |  -2.310972   .9026638    -2.56   0.011    -4.086798   -.5351461
            /cut2 |  -1.593765   .8053596    -1.98   0.049    -3.178659   -.0088717
            /cut3 |  -1.026594   .7748014    -1.32   0.186    -2.551573    .4983856
            /cut4 |  -.2440128   .7459233    -0.33   0.744    -1.712417    1.224392
            /cut5 |   .6439392    .736864     0.87   0.383    -.8067583    2.094637
            /cut6 |   1.154362   .7331829     1.57   0.117    -.2891483    2.597872
            /cut7 |   1.800908   .7337834     2.45   0.015     .3561971    3.245618
            /cut8 |   2.757425   .7355418     3.75   0.000      1.30925      4.2056
            /cut9 |   3.590239   .7397003     4.85   0.000     2.133914    5.046564
------------------------------------------------------------------------------
```

Con el fin de comparar los países, hemos salvado los parámetros, errores estándares y datos de su estimación en el archivo m1.ster utilizando la opción `saving(m1, replace)`. Ahora estimamos la diferencia entre los parámetros invocando el modelo estimado con la opción using m1:

```
. mi estimate (diff: _b[6.pais#c.loging_usd] - _b[2.pais#c.loging_usd] ) using m1

Multiple-imputation estimates                   Imputations       =         10
Survey: Ordered logistic regression             Number of obs     =      1,350
Number of strata  =         9                   Population size   =  1,320,827
Number of PSUs    =       400
                                                Average RVI       =     0.0129
                                                Largest FMI       =     0.0972
                                                Complete DF       =        391
DF adjustment:   Small sample                   DF:     min       =     260.30
                                                        avg       =     290.51
                                                        max       =     374.82
Model F test:       Equal FMI                   F(   3,  376.1)   =      30.34
Within VCE type:   Linearized                   Prob > F          =     0.0000
------------------------------------------------------------------------------
             p12 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------------
            pais |
         Ecuador  |  -.0033438   1.010766    -0.00   0.997    -1.991998     1.98531
                  |
pais#c.loging_usd |
         Bolivia  |   .2749294   .1244225     2.21   0.028     .0299266    .5199322
         Ecuador  |   .4147046   .1109087     3.74   0.000     .1966234    .6327858
-----------------+------------------------------------------------------------------
            /cut1 |  -2.310972   .9026638    -2.56   0.011    -4.086798   -.5351461
            /cut2 |  -1.593765   .8053596    -1.98   0.049    -3.178659   -.0088717
            /cut3 |  -1.026594   .7748014    -1.32   0.186    -2.551573    .4983856
            /cut4 |  -.2440128   .7459233    -0.33   0.744    -1.712417    1.224392
            /cut5 |   .6439392    .736864     0.87   0.383    -.8067583    2.094637
            /cut6 |   1.154362   .7331829     1.57   0.117    -.2891483    2.597872
            /cut7 |   1.800908   .7337834     2.45   0.015     .3561971    3.245618
            /cut8 |   2.757425   .7355418     3.75   0.000      1.30925      4.2056
            /cut9 |   3.590239   .7397003     4.85   0.000     2.133914    5.046564
------------------------------------------------------------------------------
Transformations                                 Average RVI       =     0.0706
                                                Largest FMI       =     0.0672
                                                Complete DF       =        391
DF adjustment:   Small sample                   DF:     min       =     309.04
                                                        avg       =     309.04
Within VCE type:   Linearized                           max       =     309.04
```

```
         diff: _b[6.pais#c.loging_usd] - _b[2.pais#c.loging_usd]
------------------------------------------------------------------------------
         p12 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        diff |   .1397752   .1678659     0.83   0.406    -.1905295      .47008
------------------------------------------------------------------------------

. mi testtransform diff
note: assuming equal fractions of missing information
         diff: _b[6.pais#c.loging_usd] - _b[2.pais#c.loging_usd]
 ( 1)  diff = 0
       F(  1, 309.0) =    0.69
            Prob > F =    0.4057
```

La diferencia observada no es estadísticamente significativa. Hemos invocado al comando "mi testtransform" para ilustrar su uso. Dicho comando también puede utilizarse para estudiar hipótesis no lineales, por ejemplo, la razón entre parámetros.

Una nota importante se refiere a la codificación de valores perdidos, observaciones para las que no es válida una pregunta, etc. Stata requiere que los valores perdidos (missing values) se codifiquen con "."; otros valores que deben ser excluidos se deben codificar con "valores perdidos extendidos (extended missing values)": .a, .b, .c, etc. Por ejemplo, los casos para los cuales la información "no aplica" e.g. personas que no reciben ingresos, los codificamos ".a", etc. Esta codificación no debe ser alterada. En caso contrario los comandos tipo "mi" no trabajarán.

Los ejemplos dados solo pretender ilustrar las posibilidades del módulo *mi* de Stata. Existen una rica variedad de comandos de post-estimación, otras estadísticas para evaluar el efecto de perder información y una gran diversidad de modelos. El manual de referencia para imputaciones múltiples de Stata además de servir de guía para los comandos del programa constituye una excelente revisión sobre el tema.

6 APÉNDICES

APÉNDICE 1: EXPLORACIÓN DE VARIABLES ASOCIADAS A LOS VALORES PERDIDOS DE LOS INGRESOS DEL INFORMANTE – BASE DE DATOS PARA ARGENTINA 2017.

Creamos una variable que identifique los valores perdidos:

```
misstable sum p58a, gen(miss_)
replace miss_p58a=.a if p58a==.a
label define miss_p58a 0 "Not missing" 1 "missing" .a "No aplica", modify
label values miss_p58a miss_p58a
label variable miss_p58a "no-respuesta p58a(p58a=.)"
```

Simplificamos la codificación de algunas variables que utilizaremos como ejemplo:

```
gen redpub=(p20==1)
replace redpub=. if p20==.
replace redpub=.a if p20==.a
label var redpub "(p20==1)"
gen num_hab= p16
replace num_hab= 5 if p16 >4 & p16 <.
label define num_hab 5 "5+" .a "No aplica", modify
label val num_hab num_hab
label variable num_hab "No. de habitaciones"
```

Analizamos el impacto de los valores perdidos:

```
logistic miss_p58a i.p15 i.num_hab  redpub
```

```
Logistic regression                             Number of obs     =        668
                                                LR chi2(7)        =      46.24
                                                Prob > chi2       =     0.0000
Log likelihood = -405.67247                     Pseudo R2         =     0.0539
```

miss_p58a	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
p15						
apartamento	2.47056	.5143501	4.34	0.000	1.642791	3.715425
pieza en inquilinato,..	1.368889	.9952073	0.43	0.666	.3292524	5.691246
vivienda precaria o ..)	1	(empty)				
otro	1	(empty)				
num_hab						
2	2.197324	.7307924	2.37	0.018	1.144983	4.21686
3	2.276831	.775511	2.42	0.016	1.167894	4.438727
4	2.892838	1.07833	2.85	0.004	1.393258	6.006433
5+	2.109409	.8256655	1.91	0.057	.9794522	4.542954
redpub	2.160748	.5036967	3.31	0.001	1.368292	3.412162
_cons	.1049325	.0381519	-6.20	0.000	.0514548	.2139903

Los chances de obtener un valor perdido en los ingresos son 2.5 veces mayores entre los que viven en apartamentos que entre los que viven en casas. Estos chances también se incrementan con el número de habitaciones en el hogar. Los chances se duplican entre aquellos que están conectados a las redes públicas de desagüe. Claramente los valores perdidos de los ingresos del informante están asociados a los valores de otras variables, lo que descarta la posibilidad de considerar dichos valores como perdidos completamente al azar. La suposición MAR es más adecuada que la MCAR en nuestro caso.

APÉNDICE 2: RESUMEN TÉCNICO DE LOS PROCEDIMIENTOS DE ANÁLISIS DE IMPUTACIONES MÚLTIPLES.

A2.1 Estimadores

Sea *q* un parámetro poblacional sobre el cual se desea realizar ciertas inferencias (e.g. una media poblacional, un parámetro en una regresión lineal, etc.). Supongamos que se han obtenido *m* bases de datos completadas mediante imputación múltiple (*mi*). Denotemos por $\hat{q}_i$ el estimador de *q* obtenido de la base *i, i=1, …, m* y sea $\widehat{U}_i$ el estimador de su varianza. El *mi*-estimador de *q* viene dado por

$$\bar{q} = \frac{1}{m}\sum_{i=1}^{m} \hat{q}_i \; ;$$

el *mi*-estimador de la varianza *total* es

$T = \bar{U} + (1 + 1/m)B$,

donde $\bar{U} = \frac{1}{m}\sum_{i=1}^{m} \widehat{U}_i$ es el estimador de la varianza *intra*-imputación y $B = \sum_{i=1}^{m}\left(\hat{q}_i - \bar{q}\right)^2/(m-1)$ es el estimador de la varianza *entre*-imputaciones. Para realizar inferencias se asume que la estadística

$T^{-½}(q - \bar{q})$

se distribuye aproximadamente como una *t* con *v* grados de libertad. Para muestras grandes

$v = (m-1)\left(1 + \frac{1}{r}\right)^2$, $r = \frac{(1+m^{-1})B}{\bar{U}}$;

ver StataCorp (2015), pag.62 para más detalles.

En el caso multipramétrico, denotemos por $\boldsymbol{q}$ un vector de *p* parámetros poblacionales de interés (e.g. todos los parámetros en una regresión lineal). Supongamos que se han obtenido *m* bases de datos completadas mediante imputación múltiple (*mi*). Denotemos por $\widehat{\boldsymbol{q}}_{\boldsymbol{i}}$ el estimador de $\boldsymbol{q}$ obtenido de la base *i, i=1, …, m* y sea $\widehat{\boldsymbol{U}_{\boldsymbol{i}}}$ el estimador de su matriz de covarianzas. El *mi*-estimador de $\boldsymbol{q}$ viene dado por

$$\bar{\boldsymbol{q}} = \frac{1}{m}\sum_{i=1}^{m} \widehat{\boldsymbol{q}}_{\boldsymbol{i}} \; ;$$

el *mi*-estimador de la matriz de covarianzas *total* es

$\boldsymbol{T} = \bar{\boldsymbol{U}} + (1 + 1/m)\boldsymbol{B}$,

donde $\bar{\boldsymbol{U}} = \frac{1}{m}\sum_{i=1}^{m} \widehat{\boldsymbol{U}_i}$ es el estimador de la matriz de covarianzas *intra*-imputación y

$$\boldsymbol{B} = \sum_{i=1}^{m}\left(\widehat{\boldsymbol{q}}_i - \bar{\boldsymbol{q}}\right)\left(\widehat{\boldsymbol{q}}_i - \bar{\boldsymbol{q}}\right)'/(m-1)$$

es el estimador de la matriz de covarianzas *entre*-imputaciones. Para la inferencia se asume que

$$(\boldsymbol{q} - \bar{\boldsymbol{q}})\boldsymbol{T}^{-1}(\boldsymbol{q} - \bar{\boldsymbol{q}})'/k \sim F_{k,v} \, ,$$

Una distribución F con *k y v* grados de libertad, $k = rango(\boldsymbol{T})$ y *v* dado por las fórmulas anteriores con *r* sustituido por

$r_{ave} = \left(1 + \frac{1}{m}\right) tr(\boldsymbol{B}\bar{\boldsymbol{U}}^{-1})/k.$

Ver StataCorp (2015), pag.62 para una discusión más detallada.

A2.2 Otras estadísticas para examinar el impacto de la información perdida.

RVI: El *incremento relativo promedio de las varianzas* por efecto de las no-respuestas es

$r_{ave} = \left(1 + \frac{1}{m}\right) tr(\boldsymbol{B}\bar{\boldsymbol{U}}^{-1})/k.$

Para cada componente de $\boldsymbol{q}$ el RVI es calculado como

$r = \frac{(1+m^{-1})B}{\bar{U}}$.

El *incremento porcentual de los errores estándares* debido a la información perdida reportado por el comando "mi estimate, dftable" es calculado como

$[(T/\bar{U})^{-1/2} - 1]$ x*100%*

FMI: La fracción de información perdida como consecuencia de los valores perdidos que reporta el commando "`mi estimate, vartable`" para muestras grandes es:

$FMI = \frac{(r+2)/(v+3)}{(r+1)}$

La *eficiencia relativa* debida a utilizar *m* imputaciones en lugar de un número infinito de ellas es:

$e = (1 + \lambda/m)^{-1}$,

donde λ es la proporción de información perdida de la variable de interés y *m* es el número de imputaciones. El lector es referido al Manual de referencia para Imputaciones Múltiples de Stata para una discusión más completa y las definiciones de otras estadísticas reportadas por el programa (StataCorp LTD, 2015, Release 14).

APÉNDICE 3: VARIABLES UTILIZADAS EN LOS MODELOS DE IMPUTACIÓN.

VARIABLE	DESCRIPCIÓN
P1	Género
P2	Edad
P3	Satisfacción con la vivienda
P7A	Incluyéndose usted, ¿cuántas personas conforman su hogar?
P12	¿Qué tan satisfecho se encuentra Ud. con su vivienda?
P13_1	¿Qué tan satisfecho se encuentra Ud. con: El tamaño de su vivienda?
P13_2	¿Qué tan satisfecho se encuentra Ud. con: Cercanía de su vivienda a medios de transporte?
P13_3	¿Qué tan satisfecho se encuentra Ud. con: Distancia de su vivienda a la actividad principal?
P14_1	Satisfacción: Calidad de los servicios provistos en el barrio /vecindario
P14_2	Satisfacción: Seguridad del barrio
P14_3	Satisfacción: Estado de limpieza y conservación del barrio
P15	Tipo de vivienda
P16	¿Cuántas habitaciones tiene la vivienda para su uso exclusivo?
P17	Material predominante de construcción del piso de la vivienda
P18	¿Esta vivienda accede al agua principalmente por…?
P20	Servicio de desagüe (eliminación de excretas) de la vivienda
P22	Combustible usado principalmente para cocinar o calentar el agua
P25	¿Ud. Se ha mudado alguna vez fuera de su núcleo familiar?
P31	¿Si tuviera la posibilidad de mudarse, lo haría?
P33	Propiedad de la vivienda/terreno
P38_3	Tolerancia al riesgo: opción 3
P38_4	Tolerancia al riesgo: opción 4
P41	Estado general de salud
P48_1	Últimos 3 meses viven de: subsidios, planes sociales o ayuda del gobierno
P48_2	Últimos 3 meses viven de: ayuda de familiares, vecinos, iglesias u otras organizaciones sociales
P48_3	Últimos 3 meses viven de: fuentes laborales, jubilación o pensión
P49	Situación laboral
P51	Satisfacción con el trabajo
P63	¿tiene usted una cuenta en alguna institución financiera?
P82	¿tiene usted una cuenta en alguna institución financiera?
P92	¿tiene usted una cuenta en alguna institución financiera?
P101_1	Máximo nivel educativo: usted
P101_3	Máximo nivel educativo: jefe del hogar
P149	En los 12 meses anteriores, ¿ud. o algún miembro de su hogar ha sufrido un accidente de tránsito que haya causado una lesión permanente?
P150_1	Tipo de acoso: Insinuación y/o acoso visual o verbal, toma de fotos sin consentimiento
P150_2	Tipo de acoso: Acoso físico, exhibicionismo o persecución hacia la persona
PESOPER	Ponderaciones muestrales
STRATA_MUES	Estratos Muestrales
PSU_COD	Unidades Muestrales Primarias

INGIND*	Ingreso mensual individual neto del trabajo principal
INGHOG *	Ingreso mensual neto del hogar

** En los modelos de imputación se utilizó el logaritmo natural de los ingresos individuales y del hogar; no se utilizaron los valores declarados en la encuesta directamente.*

APÉNDICE 4: CORRELACIONES INTRA-CONGLOMERADOS POR PAÍS.

PAÍS	CORRELACIÓN INTRA-CLUSTER
ARGENTINA	0.497
BOLIVIA	0.352
BRASIL	0.429
CHILE	0.117
COLOMBIA	0.311
ECUADOR	0.306
MEXICO	0.320
PANAMÁ	0.387
PERÚ	0.464
URUGUAY	0.225
VENEZUELA	0.457

7 REFERENCIAS

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle", in Petrov, B. N.; Csáki, F., *2nd International Symposium on Information Theory*, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akadémiai Kiadó, pp. 267–281

Ayanian JZ, Chrischilles EA, Fletcher RH, et al. (2003), "Understanding cancer treatment and outcomes: the Cancer Care Outcomes Research and Surveillance Consortium". *Journal of Clinical Oncology*, 22, 2292–2296.

Barnard, J., and Rubin, D.B. (1999), "Small-Sample Degrees of Freedom with Multiple Imputation," *Biometrika*, 86, 948-955.

Botman, S.L., and Jack, S.S. (1995), "Combining National Health Interview Survey Datasets: Issues and Approaches," *Statistics in Medicine*, 14, 669-677.

Box, G.E.P., and Cox, D.R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Series B, 26, 211-243.

Bradley, E., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), "Least Angle Regression". *The Annals of Statistics*, 32, 2. Institute of Mathematical Statistics: 407–51.

Graham, J. W. (2009), "Missing data analysis: Making it work in the real world". *Annual Review of Psychology* 60: 549–576.

Heitjan DF, Little RJA.(1991), "Multiple imputation for the Fatal Accident Reporting System". *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 40, 13–29

Kennickell, A.B. (2017), "Multiple imputation in the Survey of Consumer Finances". *Statistical Journal of the IAOS*, 33, 1, pp. 143-151.

Li, K.H., Meng, X.L., Raghunathan, T.E., and Rubin, D.B. (1991), "Significance Levels from Repeated p values with Multiply-Imputed Data," *Statistica Sinica*, 1, 65-92.

Li, K.-H., Raghunathan, T.E., and Rubin, D.B. (1991), "Large Sample Significance Levels from Multiply-Imputed Data Using Moment-Based Statistics and an F Reference Distribution," *Journal of the American Statistical Association*, 86, 1065-1073.

Little, R. J. A. (1988), "Missing-data adjustments in large surveys". *Journal of Business and Economic Statistics 6*, 287–296.

Little, R.J.A., and Rubin, D.B. (2002), *Statistical Analysis with Missing Data*, 2nd edition, Hoboken: Wiley.

Morris, T. P., I. R. White, and P. Royston. (2014), "Tuning multiple imputation by predictive mean matching and local residual draws". *BMC Medical Research Methodology*, 14, 75.

National Center for Health Statistics (2018), "Multiple Imputation of Family Income and Personal Earnings in the National Health Interview Survey: Methods and Examples" Division of Health Interview Statistics, National Center for Health Statistics. Available from the NHIS Web site (https://www.cdc.gov/nchs/data/nhis/tecdoc17.pdf).

Paulin, G.D., and Sweet, E.M. (1996), "Modeling Income in the U.S. Consumer Expenditure Survey," *Journal of Official Statistics*, 12, 403-419.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., and Solenberger, P. (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models, *Survey Methodology*, 27, 85-95.

Rubin, D.B. (1978), "Multiple Imputation in Sample Surveys – A Phenomenological Bayesian Approach to Nonresponse," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 20-34.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

Rubin, D.B. (1996), "Multiple Imputation After 18+ Years," *Journal of the American Statistical Association*, 91, 473-489.

Rubin, D.B., and Schenker, N. (1986), "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse," *Journal of the American Statistical Association*, 81, 366-374.

Schafer JL. (1997), *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

Schafer, J. L., and Graham, J. W. (2002), "Missing data: Our view of the state of the art". *Psychological Methods 7*, 147–177.

Schenker N, Treiman DJ, Weidman L.(1993), "Analyses of public use decennial census data with multiply imputed industry and occupation codes". *Journal of the Royal Statistical Society: Series C (Applied Statistics) 42*, 545–556.

Schenker, N., Raghunathan T.E., Chiu, P.-L., Makuc D.M., Zhang G., and Cohen A.J. (2006), "Multiple Imputation of Missing Income Data in the National Health Interview Survey," *Journal of the American Statistical Association*, 101, 924-933.

StataCorp LP. (2015), *Stata Multiple Imputation Reference Manual: Release 14*, College Station: Stata Press.