

---

## estudios estadísticos y prospectivos

# Imputación de datos: teoría y práctica

Fernando Medina

Marco Galván



División de Estadística y Proyecciones Económicas

Santiago de Chile, julio de 2007

Este documento fue preparado por Fernando Medina y Marco Galván, Asesor Regional y Asistente de Investigación, respectivamente, de la Unidad de Estadísticas Sociales de la División de Estadística y Proyecciones Económicas de la Comisión Económica para América Latina y el Caribe (CEPAL), bajo la supervisión de Juan Carlos Feres, Jefe de dicha unidad.

Las opiniones expresadas en este documento, que no ha sido sometido a revisión editorial, son de exclusiva responsabilidad de los autores y pueden no coincidir con las de la Organización.

---

Publicación de las Naciones Unidas

ISSN versión impresa 1680-8770

ISSN versión electrónica 1680-8789

ISBN: 978-92-1-323101-2

Nº de venta: S.07.II.G.109

LC/L.2772-P

Copyright © Naciones Unidas, julio de 2007. Todos los derechos reservados

Impreso en Naciones Unidas, Santiago de Chile

---

La autorización para reproducir total o parcialmente esta obra debe solicitarse al Secretario de la Junta de Publicaciones, Sede de las Naciones Unidas, Nueva York, N. Y. 10017, Estados Unidos. Los Estados miembros y sus instituciones gubernamentales pueden reproducir esta obra sin autorización previa. Sólo se les solicita que mencionen la fuente e informen a las Naciones Unidas de tal reproducción.

## Índice

---

<b>Resumen</b> .....	7
<b>I. Introducción</b> .....	9
<b>II. ¿Qué son los datos faltantes (<i>missing values</i>)?</b> .....	11
<b>III. Evolución histórica de los métodos de imputación</b> .....	13
<b>IV. Los objetivos teóricos y prácticos de la imputación</b> .....	15
<b>V. Patrones de comportamiento de los datos omitidos</b> .....	17
<b>VI. La distribución de los datos faltantes</b> .....	19
<b>VII. Procedimientos tradicionales de imputación</b> .....	21
A. Análisis con datos completos ( <i>Listwise o case deletion</i> LD).....	21
B. Análisis con los datos disponibles ( <i>pairwise deletion</i> ) .....	22
C. Reponderación.....	23
<b>VIII. Imputación simple</b> .....	25
A. Imputación por el método de medias no condicionadas.....	25
B. Imputación por medias condicionadas para datos agrupados.....	26
C. Imputación con variables ficticias .....	27
D. Imputación mediante una distribución no condicionada .....	27
E. Imputación por regresión.....	28
F. ¿Cuándo es adecuada la imputación simple? .....	29
G. Estimación por máxima verosimilitud (MV).....	29
<b>IX. Imputación múltiple</b> .....	31
A. ¿Qué es imputación múltiple (IM)?.....	31
B. Procedimiento de imputación múltiple (IM) .....	33
C. Consideraciones acerca del procedimiento IM.....	34
<b>X. Imputación de datos en encuestas complejas</b> .....	37

<b>XI. Efectos de la imputación en la pobreza y la desigualdad</b> .....	41
A. Objetivos del estudio .....	41
B. Características de los datos.....	42
C. Metodología.....	42
D. Resultados.....	43
1. Tasa de no respuesta .....	43
2. Distribuciones de frecuencias .....	45
3. Modelos ajustados .....	45
4. Ingreso promedio y <i>per capita</i> del hogar .....	47
5. Incidencia de la pobreza.....	52
6. Distribución el ingreso.....	55
E. Discusión .....	56
<b>XII. Conclusiones</b> .....	59
<b>Bibliografía</b> .....	61
<b>Anexos</b> .....	63
Anexo 1 Estadísticas complementarias .....	64
Anexo 2 Paquetes estadísticos para efectuar imputaciones.....	69
Anexo 3 Ejemplos de uso de los comandos para imputación en STATA.....	71
<b>Serie Estudios estadísticos y prospectivos: números publicados</b> .....	83

### Índice de cuadros

Cuadro 1 Resultado de los modelos ajustados.....	44
Cuadro 2 Distribución de datos faltantes por sexo .....	44
Cuadro 3 Resultados de los modelos ajustados .....	47
Cuadro 4 Sueldos y salarios imputados por distintos procedimientos .....	48
Cuadro 5 Ganancias imputadas por distintos procedimientos .....	49
Cuadro 6 Jubilaciones y pensiones imputadas por distintos procedimientos .....	50
Cuadro 7 Ingreso <i>per capita</i> y su error estándar.....	51
Cuadro 8 Estimaciones de pobreza y desigualdad con distintos métodos de imputación .....	52
Cuadro 9 Ordenamiento de los métodos de imputación según el error estándar estimado en la variable imputada.....	52
Cuadro 10 Errores de muestreo de las tasas de indigencia y pobreza .....	53
Cuadro 11 Efecto de la imputación en la tasa de pobreza por fuente de ingreso .....	54
Cuadro 12 Errores de muestreo de los índices de desigualdad.....	55
Cuadro 13 Distribución del ingreso por decil/distintos métodos de imputación.....	56

### Índice de recuadros

Recuadro 1 Estimación de máxima verosimilitud con el algoritmo EM.....	30
Recuadro 2 Procedimiento de imputación múltiple .....	32

### Índice de gráficos

Gráfico 1 Sueldos y salarios originales con <i>missing</i> .....	46
Gráfico 2 Ganancias originales con <i>missing</i> .....	46
Gráfico 3 Jubilaciones y pensiones originales con <i>missing</i> .....	46
Gráfico 4 Incidencia de la pobreza por distintos métodos de imputación .....	53
Gráfico 5 Estructura del ingreso <i>per capita</i> por deciles.....	56

**Índice de figuras**

Figura	1	Patrones de omisión de datos.....	18
Figura	2	Patrón con parámetros no identificados.....	18
Figura	3	Utilización de registros con información completa.....	22
Figura	4	Análisis con los datos disponibles ( <i>pairwise deletion</i> ).....	23
Figura	5	Efecto de la imputación con promedios.....	26
Figura	6	Sustitución de valores con covariables.....	28
Figura	7	Imputación múltiple.....	33
Figura	8	Eficiencia del proceso IM.....	35
Figura	9	Tipología de los métodos para la imputación de datos.....	39
Figura	10	Rango de variación de los sueldos y salarios.....	48
Figura	11	Rango de variación de las jubilaciones y pensiones.....	50
Figura	12	Rango de variación de las jubilaciones y pensiones.....	51

**Índice de los anexos**

Cuadro	A.1	Medidas de error, asimetría y <i>kurtosis</i> .....	67
Cuadro	A.2	Estimadores y su error estándar por distintos métodos de imputación.....	67
Cuadro	A.3	Índices de desigualdad para distintos procedimientos de imputación.....	68
Cuadro	A.4	Paquetes disponibles para imputación de datos.....	69
Cuadro	A.5	Ajuste de modelos de regresión para imputar datos.....	71
Cuadro	A.6	Patrón de datos faltantes.....	72
Cuadro	A.7	El procedimiento <i>hot-deck</i> .....	73
Cuadro	A.8	El procedimiento <i>hot-deck</i> con regresión.....	75
Cuadro	A.9	Imputación por regresión.....	77
Cuadro	A.10	Imputación simple.....	77
Cuadro	A.11	Imputación múltiple.....	78
Cuadro	A.12	Imputación múltiple (ICE).....	79
Gráfico	A.1	Sueldos y salarios.....	64
Gráfico	A.2	Ganancias.....	65
Gráfico	A.3	Jubilaciones y pensiones.....	66



## Resumen

---

La presencia de datos faltantes, es la situación a la que permanentemente se enfrentan investigadores y tomadores de decisiones. Disponer de un archivo de datos completos es ideal, pero aplicar métodos de imputación inapropiados para lograrlo, puede generar más problemas de los que resuelve. Durante las últimas décadas se han desarrollado procedimientos que tienen mejores propiedades estadísticas que las opciones tradicionales como la eliminación de datos (*listwise*), el pareo de observaciones (*pairwise*), el método de medias y el *hot-deck*. Los algoritmos de imputación múltiple (IM) se pueden aplicar utilizando paquetes comerciales y de acceso gratuito, pero imputar información no debe entenderse como un fin en sí mismo. Sus implicaciones en el análisis secundario de datos deben evaluarse con cautela, y este trabajo concluye que no existe el método de imputación ideal. Cada situación es diferente, y la tasa de no respuesta y su distribución espacial cambia entre encuestas, por lo que no es conveniente adoptar —*a priori*— el mismo procedimiento de imputación para todas las variables, en todas las encuestas. En la primera parte se analiza la teoría en la que se sustentan los procedimientos de imputación utilizados, y en la segunda se aplican ocho métodos alternativos para imputar distintos conceptos de ingreso para datos provenientes de una encuesta de hogares, y se evalúa la sensibilidad de los índices de pobreza y desigualdad (Gini, Theil y Atkinson ( $\epsilon = 2$ )), a las técnicas de imputación utilizadas. Se demuestra que los índices de pobreza son sensibles a los métodos de imputación, en tanto el procedimiento de sustitución de información tiene menor impacto en los indicadores de desigualdad.





## I. Introducción

---

La falta de repuesta —total o parcial— es una situación recurrente en las encuestas de hogares, y no tenerlo en cuenta puede generar situaciones no deseadas en la fase de inferencia estadística.<sup>1</sup> Hay quienes afirman que la ausencia de información y la presencia de datos aberrantes son un mal endémico en las ciencias sociales y el análisis económico (Juster y Smith, 1998).

A pesar de que los usuarios están informados de la presencia de registros sin información (*missing values*), y reconocen la existencia de observaciones “aberrantes” (*outliers*), es frecuente que lo pasen por alto, debido a que no son conscientes de las implicaciones estadísticas que conlleva trabajar con datos faltantes o aplicar procedimientos de imputación deficientes.<sup>2</sup>

Roth (1994), analizó distintas investigaciones publicadas en el *Journal of Applied Psychology* (JAP) y en el *Personnel Psychology* (PP) para el período 1989-1991, y afirma que el 42% de los artículos publicados en JAP y el 77% de la revista PP, requerían algún comentario respecto a la falta de datos y no se incluyó.<sup>3</sup> Por su parte, King *et al.* (2001), analiza algunas revistas de ciencia política para el

---

<sup>1</sup> De acuerdo con Cochran (1977), la falta de respuesta en las encuestas de hogares se presenta por: a) Cobertura. No se pudo ubicar algunas de las unidades seleccionadas debido a problemas de acceso. b) Localización. Cuando no fue posible localizar a ningún miembro del hogar durante la visita a la vivienda seleccionada. c) Informante inadecuado. Cuando la persona entrevistada no está en posibilidad de proporcionar la información que se le demanda. d) Rechazo. Cuando los hogares seleccionados se niegan a participar en la encuesta.

<sup>2</sup> La presencia de datos faltantes en las encuestas ha sido analizado por los especialistas de las Oficinas Nacionales de Estadística de la región, quedando en evidencia la magnitud del problema y las soluciones que los países están aplicando (CEPAL, 2002).

<sup>3</sup> En esta investigación también se señala que entre el 23 y 39% de las investigaciones publicadas no se requería ninguna mención al tema de omisión de datos.

período 1993-1997, y concluye que sólo el 19% de los trabajos publicados hacían alguna referencia a la ausencia de información. Asimismo, indica que el 94% de los investigadores que sustituyeron datos optaron por eliminar información y el resto aplicó el método de promedios.<sup>4</sup>

Las rutinas de los paquetes estadísticos asumen que se trabaja con datos completos e incorporan opciones —no siempre las más adecuadas— para imputar observaciones sin que el usuario se de cuenta de ello. Está ampliamente documentado que la aplicación de procedimientos inapropiados de sustitución de información introduce sesgos y reduce el poder explicativo de los métodos estadísticos, le resta eficiencia a la fase de inferencia y puede incluso invalidar las conclusiones del estudio.

Los procedimientos de imputación que se utilizan con mayor frecuencia limitan o sobredimensionan el poder explicativo de los modelos (Acock, 2005), y generan estimadores sesgados que distorsionan las relaciones de causalidad entre las variables, generan subestimación en la varianza y alteran el valor de los coeficientes de correlación.

Durante las últimas décadas se han propuesto distintas metodologías para sustituir datos faltantes; sin embargo, es frecuente que estos procedimientos se apliquen sin tener en cuenta sus fundamentos teóricos y sus limitaciones prácticas (véase anexo 2).

Rubin (1987), sustenta que los procedimientos de imputación múltiple (IM) deben aplicarse en forma intensiva, pero no aclara que es altamente probable que en la práctica no se satisfagan los supuestos en que se fundamenta su metodología, ya que es común que el patrón de datos omitidos esté asociado a las características de la población de referencia, lo cual invalida el supuesto de aleatoriedad en el que se sustenta la técnica IM, y que asume que la falta de información tiene una distribución aleatoria en la población de referencia.

Entre los estadísticos de encuestas existe consenso de que la mejor forma de enfrentar la falta de respuesta es evitándola. No obstante, se reconoce que eliminarla es imposible, por lo que toda vez que esta se presenta existen procedimientos para sustituir información, pero bajo ninguna circunstancia es adecuado afirmar que una cifra imputada es mejor que el dato observado. Está ampliamente documentado que las unidades de observación que no responden, pueden diferir en forma importante de las que sí lo hacen, lo cual induce a sesgos de estimación.

En este trabajo, se analizan los fundamentos teóricos de un conjunto amplio de métodos de imputación. En la primera parte se describe la teoría en la que se sustentan y la forma en que se aplican, haciendo énfasis en sus bondades y limitaciones, así como en los sesgos que se generan cuando se utilizan de manera acrítica.

En la segunda parte se aplican ocho algoritmos de imputación con el objetivo de sustituir valores omitidos en los sueldos y salarios, las ganancias y las jubilaciones y pensiones, para información proveniente de encuestas de hogares. Posteriormente, se evalúa la sensibilidad de los indicadores de pobreza y desigualdad a los procedimientos utilizados, y a modo de conclusión se presentan algunas reflexiones que permiten afirmar que no existe el mejor método de imputación, y se sugiere no utilizar los métodos de sustitución de datos como un fin en sí mismo.

---

<sup>4</sup> No se dispone de una investigación similar en el caso de los estudios de pobreza y condiciones de vida. No obstante, se intuye que las cifras aquí reportadas deben ser muy similares en este ámbito de estudio.

## II. ¿Qué son los datos faltantes (*missing values*)?

---

En las encuestas de hogares la falta de respuesta se asocia a diversas causas.<sup>5</sup> A la fatiga del informante, al desconocimiento de la información solicitada, al rechazo de las personas a informar acerca de temas sensibles, a la negativa de los hogares a participar en la investigación, así como a problemas asociados a la calidad del marco de muestreo.<sup>6</sup>

A pesar de que un cuestionario sea considerado correcto, la realidad indica que frecuentemente los archivos contienen observaciones “aberrantes” o poco probables,<sup>7</sup> y existen situaciones en que debido a los objetivos de la investigación, deliberadamente se omite información de personas que no forman parte de la población de estudio. Por ejemplo, en una encuesta de demografía y salud, la fecundidad tradicionalmente se estudia en las mujeres de 14 años y más, y por definición, en la base de datos existirán registros sin información para las personas que no forman parte de la población objetivo.

---

<sup>5</sup> Dreesbeke y Lavallée (1996), clasifican la falta de respuesta de la manera siguiente: a) No respuesta total o parcial. La no respuesta total existe cuando no se ha conseguido la entrevista, en tanto que la falta de respuesta parcial ocurre cuando sólo fue posible recabar información para un grupo de preguntas. b) No respuesta ignorable y no ignorable. Se presenta el caso de no respuesta ignorable cuando la probabilidad de que un hogar entrevistado no responda no depende de las características del hogar, en tanto que en la falta de respuesta no ignorable existe correlación entre la omisión de datos y las características de las unidades que no quisieron colaborar en la investigación.

<sup>6</sup> Es importante que los responsables de las encuestas asignen códigos apropiados para distinguir en los archivos de datos la falta de respuesta, ya que la mayoría de los paquetes estadísticos identifican los datos faltantes con “.”. No se recomienda utilizar los dígitos “0” y “999” para identificar registros con datos faltantes, ya que los algoritmos utilizan estos valores en sus cálculos.

<sup>7</sup> Si se analiza, por ejemplo, el ingreso por sueldos y salarios, es probable que los registros con valores menores o superiores a 10 millones de unidades deban ser considerados como valores extremos.

En estas y otras situaciones, se aconseja asignar códigos para diferenciar entre las preguntas que presentan omisiones, de aquellos registros que se excluyeron debido a que el flujo del cuestionario no las consideraba parte de la población elegible.

El código “.” que comúnmente se asocia con información faltante, se debe reservar para situaciones en que no fue posible recabar datos, mientras que el dígito “0” (cero) se preserva para variables (discretas o continuas) que puedan asumir ese valor, y por tanto no se aconseja asignarlo a registros sin información.<sup>8</sup>

Asimismo, a pesar de que es común utilizar el “código 9999” para indicar la falta de respuesta, esta práctica no se considera adecuada. Cuando se presenta esta situación, se sugiere cambiar ese valor por el carácter que el lenguaje de programación o paquete utilizado identifique a los *missing values*, ya que si esto no se hace del conocimiento de los usuarios el “número 9999” será utilizado, por ejemplo, en el cálculo de las medidas de tendencia central y dispersión.<sup>9</sup>

Acock (2005), concluye que de la revisión de las principales revistas que se especializan en estudios de familia, muy pocos autores aclaran la manera en que trabajaron los datos faltantes y no informan de las consecuencias que genera la eliminación de información en la reducción del tamaño de muestra, en la varianza de los estimadores y en el sesgo de estimación.<sup>10</sup>

Los *missing values* forman parte de un conjunto de observaciones con características especiales que incluye a los datos agrupados, agregados, redondeados, censurados o truncados; es decir, a datos con información especial (Heitjan y Rubin, 1991).

Existen también las variables latentes que están relacionadas con *missing data*. Este tipo de variables son cantidades que no se pueden observar, y en el trabajo empírico sólo se puede lograr una medición imperfecta de ellas; por ejemplo, la medición de inteligencia o la asertividad de las personas.<sup>11</sup>

Un aspecto crucial en el análisis de datos se vincula al porcentaje máximo de omisiones que deben aceptarse. No existen criterios objetivos para dilucidar este tema, por lo que cada investigador debe hacerse cargo de sus propias decisiones.

Los que promueven el uso de la imputación múltiple como el método más adecuado para reponer información omitida (Rubin, 1987), afirman que los procedimientos de IM generan buenos resultados, aún con porcentajes de omisión del 30, 40 o 50%. No obstante, es preciso señalar que cuando se trabaja con una encuesta probabilística, el tamaño de muestra garantiza cierta precisión para una tasa máxima de no respuesta, y en la medida de que la omisión supera el umbral establecido se pone en riesgo la confiabilidad estadística de las variables principales. Por tanto, no se recomienda imputar datos en situaciones en que la omisión en una o más variables alcance porcentajes superiores al 20%.

Si se trabaja, por ejemplo, con una base de datos en donde la tasa de omisión en las variables de interés se ubica en 25%, se debe tener presente que modelar la respuesta en una de cada cuatro observaciones puede resultar adecuado en el ámbito académico, pero se considera poco útil desde el punto de vista práctico, sobre todo cuando los resultados se utilizarán apoyar el diseño o evaluación de políticas públicas.

---

<sup>8</sup> Si la falta de información se identifica con 0, las rutinas de los paquetes estadísticos no están en posibilidad (al menos que así se indique en el código del programa) de identificar que ese valor “0” es un *missing value*.

<sup>9</sup> Si en una investigación por muestreo interesa analizar las distintas causas de falta de respuesta, la asignación de un único código (por ejemplo, -9) para identificar este tipo de observaciones es un despropósito. Es necesario asignar códigos diferentes a cada causa, y asumir las precauciones adecuadas en el tratamiento de los datos. De acuerdo a Little y Rubin (2002), en los estudios de percepción si el entrevistado responde “no sé”, puede considerarse que su respuesta está a la mitad entre está de acuerdo o desacuerdo.

<sup>10</sup> Eliminar información en un estudio probabilístico, altera la estructura del diseño muestral, las probabilidades de selección y genera sesgos que afectan la inferencia estadística, y si no se corrigen los factores de expansión las observaciones que permanecen en muestra no estimarán en forma correcta los totales y promedio de la población de referencia.

<sup>11</sup> En psicología los investigadores establecen diferencias entre datos faltantes en las variables independientes o predictores y la variable dependiente o de resultado. En general, esta situación no es relevante; no obstante, se debe tener seguridad acerca de qué tipo de variable se analiza.

### III. Evolución histórica de los métodos de imputación

---

Durante la década de los setenta, la imputación de datos significaba identificar y sustituir los registros sin información. En ese contexto, los procedimientos *hot-deck*, en sus distintas variantes, se aplicaban profusamente para suplir información en censos y encuestas, y los métodos de ajuste por promedios y *cold-deck* eran frecuentemente utilizados.

En Rubin (1976), se propuso un marco conceptual para el análisis de datos faltantes sustentado en métodos de inferencia estadística. Posteriormente, la aparición del algoritmo Expectation Maximization (EM) permitió generar estimadores robustos a partir de la aplicación del método de máxima verosimilitud (MV) (Dempster, Laird, y Rubin, 1977), en donde las observaciones faltantes se asumen como variables aleatorias y los datos imputados se generan sin necesidad de ajustar modelos.

Más tarde, Rubin (1987), introdujo el concepto de imputación múltiple (IM), sustentado en la premisa de que cada dato faltante debe ser reemplazado a partir de  $m > 1$  simulaciones. La aplicación de esta técnica se facilitó con los avances computacionales y el desarrollo de los métodos *bayesianos* de simulación que aparecieron hacia finales de los ochenta (Schafer, 1997). A partir de ese momento, se propició el uso intensivo de estos algoritmos, debido a que distintas rutinas de IM y MV se incluyeron en paquetes comerciales y de acceso gratuito.

Durante el decenio de los noventa se registraron avances notables. El método de reponderación, históricamente utilizado por los estadísticos de encuestas, se modificó con el propósito de utilizarlo en la imputación de datos en modelos de regresión (Ibrahim, 1990). También se desarrollaron técnicas en el campo de la bioestadística, con el propósito de resolver situaciones en que los datos faltantes dependían de las características de la población.

El objetivo era disponer de algoritmos apropiados para analizar datos clínicos, ya que en este tipo de investigaciones es común que el número de pacientes se reduzca, y los estudios concluyan con menos personas de las que iniciaron (Little, 1995).

En Little (1992) y Little y Rubin (1987), los métodos de imputación se clasifican como se muestra a continuación:<sup>12</sup>

- Análisis de datos completos (*listwise*)
- Análisis de datos disponibles (*pairwise*)
- Imputación por medias no condicionadas
- Imputación por medias condicionadas mediante métodos de regresión
- Máxima verosimilitud (MV)
- Imputación múltiple (IM)

La literatura considera que el desarrollo del método IM ha zanjado el debate respecto a la mejor forma de imputar datos omitidos. Sin embargo, los estadísticos de encuestas han demostrado que las soluciones propuestas no tienen en cuenta el diseño de la muestra, ni las probabilidades de selección de las unidades de análisis, cuando los datos provienen de muestras complejas, situación que introduce sesgos en las estimaciones.

---

<sup>12</sup> Esta clasificación no es exhaustiva, ya que no incorpora el método de *hot-deck* en sus distintas variantes, las cuales se aplican desde hace mucho tiempo para imputar datos omitidos en censos y encuestas.

## IV. Los objetivos teóricos y prácticos de la imputación

---

Se acepta, sin embargo, que con o sin datos omitidos el objetivo del análisis estadístico es generar inferencia válida. No se trata únicamente de obtener estimadores insesgados y de mínima varianza, ni tampoco ajustar modelos para sustituir de cualquier forma la información faltante.<sup>13</sup>

La imputación debe considerarse parte del proceso de investigación con el propósito de arribar a conclusiones sustentadas en evidencia empírica sólida, por lo que la atención de los usuarios y diseñadores de políticas no debiera concentrarse en generar estimadores que satisfagan propiedades estadísticas deseables.

Las bondades de los procedimientos de imputación no deben valorarse por el sólo hecho de que permiten completar información para ajustar modelos y probar hipótesis. Los criterios para evaluar la pertinencia de un método estadístico fueron establecidos por Neyman y Pearson (1933) y Neyman (1937), y guardan relación con el error cuadrático medio (ECM) y no sólo con el sesgo del estimador.

Si la variable analizada (**S**) contiene datos faltantes, esta situación debe tenerse en cuenta en el proceso de construcción del estimador.<sup>14</sup> Si la imputación que se hace es adecuada, el estimador  $\hat{S}$  será cercano al verdadero valor del parámetro **S** en muestras repetidas.

---

<sup>13</sup> Por ejemplo, la sustitución de información faltante a partir de promedios —generales o para subpoblaciones— puede ser adecuada para lograr predicciones más precisas, pero esta práctica tiene implicaciones negativas en la varianza del estimador e introduce distorsiones en el patrón de correlación de los datos (Schafer y Graham, 2002).

<sup>14</sup> La mayoría de los métodos de imputación no tiene en cuenta este hecho, por lo que generan estimadores sesgados.

De esta forma, se logra minimizar el sesgo (la diferencia promedio entre  $\hat{S}$  y  $S$ ), la varianza y desviación estándar de  $\hat{S}$ .<sup>15</sup>

Cuando la falta de datos ocurre por razones ajenas al investigador, es necesario establecer supuestos acerca de las causas que generaron las omisiones, contrastando la factibilidad de las hipótesis con el comportamiento observado en los datos.

A pesar de que se reconoce que los criterios estadísticos son fundamentales para la elección del método de imputación, es necesario tener claridad sobre el uso que se hará de la información. Si el propósito es, por ejemplo, conocer los determinantes del nivel de vida de las familias, es probable que distintos métodos sean equivalentes desde la óptica estadística. No obstante, se debe tener presente que pequeñas diferencias en el ingreso *per capita* pueden generar cambios significativos en el volumen de personas en situación de indigencia y pobreza.

Si los datos serán utilizados para diseñar políticas públicas, el número de familias en pobreza es relevante en el presupuesto del proyecto, a pesar de que se demuestre que estadísticamente no existan diferencias significativas en el estimador generado a partir de procedimientos de imputación alternativos.

---

<sup>15</sup> Sesgo y varianza se combinan en una medida denominada error cuadrático medio (ECM) el cual se computa como el promedio de la distancia entre  $(\hat{S}-S)^2$ , sobre muestras repetidas. Por tanto, el  $ECM = \text{sesgo}^2 + V(\hat{S})$ . El sesgo, la varianza y el error cuadrático medio describen el comportamiento de un estimador. El error estándar del estimador  $es(\hat{S})$  debiera ser parecida a la desviación estándar de  $\hat{S}$ , en tanto que un intervalo de confianza debe incluir al verdadero valor del parámetro  $S$  con probabilidad cercana a la tasa nominal, por lo que se tendrán intervalos más pequeños lo cual reduce la probabilidad de error tipo II.



## V. Patrones de comportamiento de los datos omitidos

---

Los estadísticos de encuestas advierten acerca de las diferencias que existen entre falta de respuesta total (no se encontró al informante, se rechazó la entrevista, problemas de marco de muestreo, etc.), y la no respuesta parcial, que se asocia con situaciones en que no se obtuvo respuesta en algunas preguntas del cuestionario. La no respuesta total, se corrige eliminando las observaciones y ajustando los factores de expansión,<sup>16</sup> de modo que las unidades que permanezcan en muestra estimen —sin sesgos— el total poblacional. Por su parte, para corregir la omisión parcial es común que se aplique el procedimiento *hot-deck* o el método de promedios que se describirá más adelante.

Los distintos procedimientos de imputación establecen supuestos acerca del patrón de comportamiento de los datos omitidos. De acuerdo a la literatura, los métodos se clasifican en tradicionales y modernos, y en general se considera que los algoritmos de máxima verosimilitud (MV) y de imputación múltiple (IM) son los más robustos.<sup>17</sup> En estudios longitudinales —encuestas de mercado de trabajo o investigaciones en educación—, la ausencia de datos es una situación habitual. Puede darse el caso, por ejemplo, que en una encuesta continua

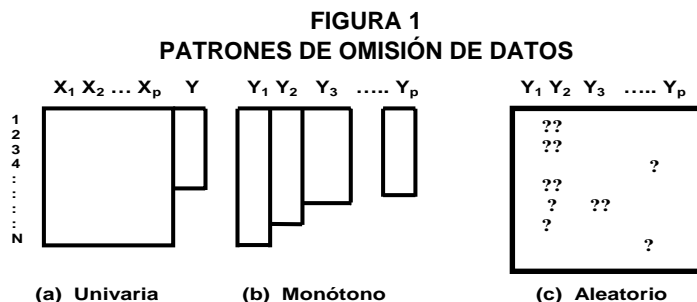
---

<sup>16</sup> Esta situación debe ser prevista en la determinación del tamaño de muestra, con el propósito de mantener la precisión de los estimadores ante la no respuesta total. Significa ajustar en forma apropiada las probabilidades de selección de las observaciones que permanecen en la muestra, de tal forma que sea posible de obtener estimadores insesgados de promedios y totales poblacionales. En ocasiones se utiliza el término reponderar la muestra; sin embargo, se debe distinguir del método de imputación de datos con el mismo nombre que será comentado en este trabajo.

<sup>17</sup> En Allison (2000), se indica que hay que asumir con cautela la afirmación de que siempre los métodos de imputación múltiple son mejores que los procedimientos univariados simples. El autor efectúa distintas simulaciones que demuestran lo ineficientes que pueden ser los algoritmos de imputación múltiple incorporados en algunos paquetes de cómputo comerciales.

se disponga de información para tres rondas, y en el último cuatrimestre del año se reporte que la familia se mudó. También es común que uno o más miembros del hogar se ausenten en una o más ocasiones, y posteriormente vuelvan a ser entrevistados. Debido a que las mediciones en estudios de panel están altamente correlacionadas, es común que los procedimientos de imputación hagan uso de información de períodos anteriores para sustituir datos omitidos.

Si la base de datos se interpreta como una matriz, en donde los renglones son las unidades de observación y las columnas representan a las variables de interés, la elección del método de imputación debiera tener en cuenta el comportamiento de los datos omitidos, ya que el análisis visual permite identificar patrones como los que se muestran en la figura 1.<sup>18</sup>

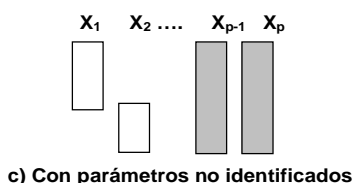


Fuente: Elaboración propia de los autores.

En 1a, la ausencia de respuesta se concentra en la variable **Y**, lo cual se conoce como “patrón univariado”, mientras que en 1b la omisión sigue un “patrón escalonado o monótono”, que es característico de estudios longitudinales, en donde  $Y_j$  representa el valor de la variable en la  $j$ -ésima ronda. Finalmente, 1c da cuenta de un “patrón aleatorio”, y en cualquier celda pueden existir datos faltantes; es decir, las omisiones no están dispuestas en una forma predeterminada.

Es posible que la falta de respuesta ocurra en dos o más variables en distintos períodos. Considere  $X_1, X_2, \dots, X_p$  y el patrón que se presenta en la figura 2, en donde  $X_1$  y  $X_2$  tienen datos para distintos registros. Esta situación corresponde a un “patrón con parámetros no identificados”.

**FIGURA 2**  
**PATRÓN CON PARÁMETROS NO IDENTIFICADOS**



Fuente: Elaboración propia de los autores.

El supuesto de que los datos faltantes siguen un patrón completamente aleatorio (*Missing Completely at Random, MCAR*)<sup>19</sup> fue introducido por Rubin (1977) y Little y Rubin (1987), y es el supuesto que se asume en la mayoría de los algoritmos de imputación. Sin embargo, es frecuente que en la práctica esta hipótesis no se satisfagan, ya que la falta de respuesta suele estar asociada a características de las familias y las personas.<sup>20</sup>

<sup>18</sup> Esta tipología se describe en detalle en Schafer y Graham (2002).

<sup>19</sup> Existen otros patrones de datos faltantes que serán comentados más adelante.

<sup>20</sup> En las encuestas de presupuestos familiares, por ejemplo, es habitual que las familias de mayores ingresos no estén dispuestas a informar de sus patrones de gasto, y tampoco les interesa entregar información de sus ingresos. Esto también se observa en encuestas con cuestionarios muy extensos que demandan mucho tiempo de entrevista a las familias.

## VI. La distribución de los datos faltantes

---

En algunas ocasiones a los registros sin información se le identifica con una variable binaria ( $Z$ ). De acuerdo a la figura 1a,  $Z$  asume el valor “1” si el dato existe y “0” en caso contrario. En 1b,  $Z$  tiene valores enteros (1, 2, ..., p), y el subíndice  $j$  identifica el registro para el cual la variable  $Y_j$  tiene información, en tanto que en 1c los valores de  $Z$  están dispuestos en un arreglo rectangular que tiene la misma dimensión que la matriz de datos.

Rubin (1976), advierte que la ausencia de datos debe analizarse como un fenómeno estocástico, por lo que  $Z$  debe considerarse como variable aleatoria con distribución de probabilidad conjunta, la cual da cuenta del porcentaje de omisión existente y de su relación con las observaciones completas. La tipología de datos faltantes propuesta por Rubin (1976), es ampliamente utilizada, pero a juicio de Schafer y Graham (2002), no ha sido bien entendida.

Sea  $Y_{com} = (Y_{obs}, Y_{mis})$  la variable de interés, en donde  $Y_{obs}$  y  $Y_{mis}$  corresponde al vector de datos observados y faltantes respectivamente. Se afirma que un proceso de datos omitidos se genera en forma aleatoria (*Missing at Random*, **MAR**),<sup>21</sup> si la distribución de los valores observados no depende del patrón de comportamiento de los registros sin información  $Y_{mis}$ :  $P(Z/Y_{com})=P(Z/Y_{obs})$ .

---

<sup>21</sup> Se dice que los datos faltantes para la variable ingreso siguen un patrón MAR, si la probabilidad de que existan omisiones dependen, por ejemplo, del nivel educativo de la persona, pero en cada categoría de escolaridad la falta de información no está relacionada con el ingreso. En la práctica esta hipótesis no puede ser comprobada, debido a que precisamente se desconoce el ingreso en algunas observaciones.

Esta situación se presenta en las encuestas de hogares, pero no en cualquier variable. En los estudios de nivel de vida, por ejemplo, se reconoce que la falta de respuesta en las familias “ricas”, está correlacionada con sus ingresos; y por tanto no se satisface el supuesto MAR. Sin embargo, cuando la omisión se presenta en la rama de actividad o en la categoría ocupacional, la falta de información no necesariamente guarda relación con el nivel de ingresos de las personas, y el supuesto MAR podría satisfacerse.

Un caso especial es el llamado proceso completamente aleatorio (MCAR),<sup>22</sup> el cual ocurre cuando la omisión no depende de los datos observados:  $P(\mathbf{R}/\mathbf{Y}_{\text{com}})=P(\mathbf{Z})$ . Por otra parte, cuando existe dependencia entre los datos completos y los faltantes ( $Y_{\text{mis}}$ ), se dice que  $Y_{\text{mis}}$  no sigue un proceso aleatorio (*Missing not at Random*, MNAR).<sup>23</sup> Es común referirse a los procesos MAR como mecanismos de no respuesta ignorable, en tanto que MNAR significa que la falta de respuesta no puede ser ignorada en el proceso de construcción del estimador ni al analizar las relaciones de causalidad entre variables.

En la práctica es común que se presenten situaciones en que los datos faltantes no siguen un patrón completamente aleatorio (MCAR) y tampoco aleatorio (MAR). Está demostrado que la falta de respuesta se asocia, por ejemplo, al estatus económico de la familia, al área geográfica de residencia del hogar, al nivel de estudio de las personas, y en estas situaciones el patrón de datos omitidos no puede ni debe ser ignorado (MNAR).

En las encuestas de educación y del mercado de trabajo la omisión está asociada a las unidades de observación, ya que es habitual que después de algún tiempo los informantes se nieguen a participar, se cambien de casa o los estudiantes emigren a otro establecimiento educativo.

Considere el patrón de datos faltantes de la figura 1a en el que los valores de la variable  $\mathbf{X}=(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p)$  existen en todas las observaciones, pero en algunos casos se desconoce el valor de  $Y$ . Cuando los datos provienen de una muestra aleatoria, los patrones de datos omitidos —MAR, MCAR y MNAR— pueden interpretarse a partir de  $\mathbf{X}$  y  $Y$ . MCAR significa que la probabilidad de que falte  $Y$  no depende de los valores de  $X$  e  $Y$  (ni de los propios ni de las otras observaciones), MAR asume que la probabilidad de que falte  $Y$  puede depender de  $X$  pero no de  $Y$ , en tanto que MNAR se interpreta en el sentido de que la probabilidad de que existan datos omitidos depende de  $Y$ .

Una manera de probar si los datos faltantes sigue un patrón MCAR se logra aplicando la prueba de hipótesis propuesta por Little (1988), que se incluye en la opción MVA del paquete SPSS, cuando se aplica el algoritmo EM (*Expectation-Maximization*).

Little propuso un estadístico de prueba que sigue una distribución  $X^2$  (ji-cuadrada) con  $f$  grados de libertad, en donde la hipótesis nula ( $H_0$ ) establece que los datos omitidos siguen un patrón MAR. Conforme a la regla de decisión, se debe rechazar  $H_0$  cuando el valor del estadístico de prueba ( $X^2$ ) para los datos observados, es mayor al valor en tablas conforme a un determinado nivel de significancia ( $\alpha$ ).

<sup>22</sup> En el caso del ingreso se dice que la omisión sigue un patrón MCAR, si se cumple que, en promedio, las personas que no respondieron tienen un ingreso similar a las que si lo informaron. En la práctica este supuesto no es posible validarlo.

<sup>23</sup> En las encuesta de empleo y de ingreso-gasto, la falta de respuesta en el ingreso está concentrada en las familias de mayor estatus económico. En este tipo de situaciones, se requiere valorar el método que se aplicará para suplir los datos faltantes.

## VII. Procedimientos tradicionales de imputación

---

Existen distintos procedimientos para sustituir la falta de observaciones. No obstante, antes de elegir alguno de ellos se recomienda analizar las distribuciones de frecuencias de las variables, así como las medidas de tendencia central, de dispersión, asimetría y *kurtosis*.<sup>24</sup>

### A. Análisis con datos completos (*Listwise o case deletion LD*)

Es habitual que los paquetes estadísticos trabajen —por defecto— sólo con información completa (*listwise*) como se menciona en la figura 3, a pesar de que se reconoce que esta práctica no es la más apropiada, ya que genera sesgos en los coeficientes de asociación y de correlación (Kalton y Kasprzyk, 1982).<sup>25</sup>

El *listwise* es el método que se utiliza con mayor frecuencia y asume que los datos faltantes siguen un patrón MC. Esta manera de proceder significa trabajar únicamente con las observaciones que disponen de información completa para todas las variables, tal como se observa en la figura 3.<sup>26</sup>

---

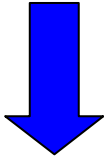
<sup>24</sup> El SAS (proc univariate) y el SPSS (explore) incorporan rutinas que generan diversos estadísticos y gráficos que permite conocer el recorrido de la variable, su dispersión, las medidas de tendencia central, e identificar valores extremos y datos faltantes.

<sup>25</sup> Cuando se opta por esta opción, se recomienda efectuar el análisis sin los datos faltantes y comparar los resultados que se obtienen con datos imputados a partir de la opción *listwise*.

<sup>26</sup> El SAS y SPSS utilizan esta opción por “defecto”, aunque ambos incluyen la opción de “*pairwise deletion case*” y el método de promedios.

**FIGURA 3**  
**UTILIZACIÓN DE REGISTROS CON INFORMACIÓN COMPLETA**

Folio	Sexo	Edad	Escolaridad	Salario	Ocupación	Ponderación
1	Mujer	40	16	4 500	?	50
2	Hombre	35	15	?	1	75
3	Mujer	65	?	1 200	1	100
4	Hombre	23	12	2 200	2	80
5	Hombre	25	?	?	3	250
6	Mujer	38	15	1 800	4	140
....						



4	Hombre	23	12	2 200	2	80
5	Mujer	38	15	1 800	4	140

Fuente: Elaboración propia de los autores.

Al eliminar información se asume que la submuestra de datos excluidos tiene las mismas características que los datos completos, y que la falta de respuesta se generó de manera aleatoria lo cual en la mayoría de las situaciones prácticas no se cumple.

Cuando los datos analizados provienen de una muestra probabilística, eliminar observaciones no es correcto ya que se debe tener presente que las unidades fueron elegidas con un procedimiento aleatorio y con probabilidad de selección, conocida y distinta de cero, que no puede ser ignorada en el tratamiento de los datos ni en el cálculo de los estimadores y sus errores.

Si la eliminación de registros no se acompaña con el ajuste apropiado de los factores de expansión, los valores estimados por la muestra —razones, promedios y totales— pueden ser incompatibles con los parámetros observados en la población de referencia. Es decir, se obtendrán estimadores sesgados de los parámetros poblacionales lo que podría invalidar las conclusiones.

## **B. Análisis con los datos disponibles (*pairwise deletion*)**

Otra posibilidad es trabajar con información completa (*Available-case (AC)*). El procedimiento AC, en contraste con LD, utiliza distintos tamaños de muestra, por lo que también se le conoce como *pairwise deletion* o *pairwise inclusion*.

En la figura 4 se observa información completa para distintos registros de las variables salario y escolaridad, por lo que es posible calcular la correlación entre ambas utilizando los folios 1, 4 y 6, en tanto que la relación entre el salario y la ocupación se podría determinar con los datos de los registros 1, 3, 4 y 6. Sin embargo, debido a la diferencia en el tamaño de muestra, no es posible comparar los valores de los coeficientes obtenidos por ambos procedimientos.

Este método asume un patrón MCAR en los datos omitidos, y hace uso de toda la información disponible sin efectuar ningún tipo de corrección en los factores de expansión. Las observaciones que no tienen datos se eliminan, y los cálculos se realizan con diferentes tamaños de muestra lo que limita comparación de resultados.

**FIGURA 4**  
**ANÁLISIS CON LOS DATOS DISPONIBLES (PAIRWISE DELETION)**

Folio	Sexo	Edad	Escolaridad	Salario	Ocupación	Ponderación
1	Mujer	40	16	4 500	2	50
2	Hombre	35	15	?	1	75
3	Mujer	65	?	1 200	1	100
4	Hombre	23	12	2 200	2	80
5	Hombre	25	?	?	3	250
6	Mujer	38	15	1 800	4	140
....						

Fuente: Elaboración propia de los autores.

Este método asume un patrón MCAR en los datos omitidos, y hace uso de toda la información disponible sin efectuar ningún tipo de corrección en los factores de expansión. Las observaciones que no tienen datos se eliminan, y los cálculos se realizan con diferentes tamaños de muestra lo que limita comparación de resultados.

Si en un estudio de niveles de vida, por ejemplo, la correlación entre el ingreso de las familias y su condición de pobreza se determina con el 95% de la muestra, y el grado de asociación entre la escolaridad y las remuneraciones se calcula con el 85% de los registros, los coeficientes obtenidos no se pueden comparar. Esta es una de las razones por la cual no se recomienda la aplicación de este procedimiento de imputación.<sup>27</sup>

Cuando se le compara con el *listwise*, esta opción tiene la ventaja de que hace uso de toda la información disponible pero la mezcla de tamaños de muestra debilita su aplicación. Por ello, bajo ninguna circunstancia se recomienda el uso de esta metodología, y en la etapa de análisis se debe tener certeza acerca de la manera en que el paquete estadístico utilizado maneja las observaciones faltantes.

### C. Reponderación

El interés por reducir el sesgo de estimación cuando no se satisface el supuesto MCAR en los datos omitidos, tiene una larga historia en la literatura del muestreo probabilístico (Little y Rubin, 1987).<sup>28</sup>

Los procedimientos de imputación ponderada representan una manera de compensar la falta de respuesta, y es posible aplicar distintos métodos para reponderar las observaciones que se mantienen en la muestra.

Los ponderadores ( $w_i$ ) se interpretan como el número de unidades de la población que representa a cada elemento en la muestra, y es común que los algoritmos de reponderación se apliquen para compensar la falta de respuesta en subgrupos de interés.

Cuando en una subclase se detecta ausencia de información, los ponderadores de las unidades que sí respondieron se utilizan para ajustar los factores de expansión, de tal forma que la submuestra observada genere estimaciones compatibles con los valores poblacionales de la subclase de interés.

<sup>27</sup> En el análisis de varianza y regresión, la diferencia en el tamaño de muestra también genera problemas en el cómputo de los grados de libertad.

<sup>28</sup> Esta manera de proceder es distinta a aquella que ante la falta de respuesta total, elimina las observaciones que presentan esta característica y los factores de expansión se ajustan (Kalton y Kasprzyk, 1986).

Este procedimiento es similar al de *post-estratificación*, con la diferencia que para reponderar las observaciones se utilizan información de la muestra estudiada, en tanto que en la *post-estratificación* se recurre a datos exógenos provenientes de otras encuestas, censos o registros administrativos.

A pesar de que este tipo de ajustes es habitual en el trabajo empírico, se sugiere evitar la divulgación de bases de datos que contengan distintos factores de ajuste para cada subclase, además de los ponderadores originales inherentes al esquema de selección de la muestra.<sup>29</sup>

En bioestadística se han desarrollado aplicaciones que utilizan modelos de regresión con covariables incompletas para reponderar datos. El propósito de estos procedimientos es compensar la falta de respuesta, ajustando los ponderadores de las observaciones que permanecen en muestra (Kalton y Kasprzyk, 1986).

Robins *et al.* (1994), ajustó regresiones ponderadas para sustituir datos faltantes relajando algunos supuestos acerca del comportamiento esperado en los parámetros del modelo. El procedimiento es una extensión del método generalizado de estimación de ecuaciones (GEE), que se utiliza para modelar promedios poblacionales a partir de variables de respuesta y predictores (Zeger *et al.*, 1988). Este tipo de procedimientos se conocen como modelos semiparamétricos, debido a que requieren que la regresión tenga una forma específica.

Las observaciones se reponderan utilizando modelos de probabilidad con información completa y datos exógenos, con el propósito de mejorar la robustez de los estimadores de modo que el comportamiento de la muestra se asemeje al de la población de referencia.

Esta metodología es menos eficiente que los procedimientos de máxima verosimilitud o de estimación bayesiana, y en Rubin (1976), se demuestra que los métodos bayesianos y de máxima verosimilitud generan estimadores más eficientes cuando el patrón de datos faltantes es MAR, y se señala que los métodos semiparamétricos no tienen en cuenta el patrón que generó las observaciones faltantes.

---

<sup>29</sup> Reponderar o post-estatificar cada variable no es una buena práctica sobre todo cuando las bases de datos son de acceso libre y no se les entrega a los usuarios toda la información necesaria. Asimismo, la introducción de muchos factores de ajuste puede ser interpretado como una mala gestión del trabajo de campo y generar desconfianza acerca de la calidad de los datos recabados.



## VIII. Imputación simple

---

### A. Imputación por el método de medias no condicionadas

La sustitución de datos utilizando promedios es una vieja práctica entre investigadores de diversas disciplinas, a pesar de que por sus limitaciones teóricas no se considera un procedimiento apropiado. En su aplicación se asume que los datos faltantes siguen un patrón **MCAR**, y ha sido ampliamente documentado que su aplicación afecta la distribución de probabilidad de la variable imputada, atenúa la correlación con el resto de las variables y subestima la varianza, entre otras cosas.

Por la manera en que se realiza la sustitución de los datos omitidos, la suma de cuadrados de las desviaciones de las observaciones respecto de la media permanece inalterada pero se incrementa el tamaño de muestra, lo cual origina que la varianza de la variable disminuya y se generen, en forma artificial, intervalos de confianza más estrechos.

A pesar de los inconvenientes señalados, es común el uso de esta metodología ya que existe la falsa creencia de que en una distribución de probabilidad normal el promedio de los datos es un buen estimador de las observaciones omitidas.

En caso de que las variables imputadas se utilicen en análisis secundario de datos, se demuestra, por ejemplo, que en los modelos de regresión se alteran los valores de los parámetros estimados, así como su significancia estadística.

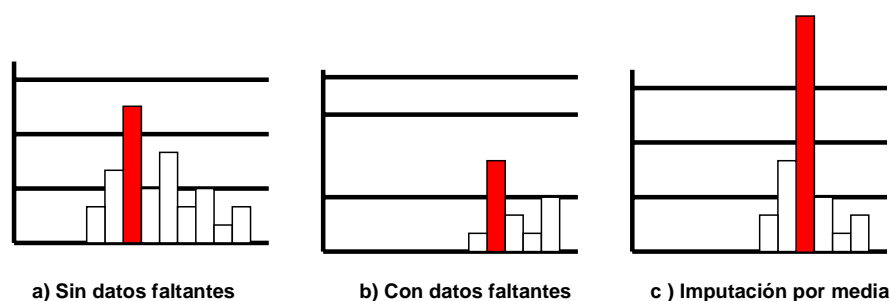
Asimismo, los problemas se multiplican en la medida de que la falta de respuesta se incrementa. Por ejemplo, si en una encuesta de hogares el 30% de los entrevistados no respondió su ingreso, y se decide imputar el valor promedio de las observaciones que disponen de información (por ejemplo, 650 pesos), el resultado será que el 30% de los registros tendrá como ingreso ese valor y varianza cero, lo cual subestimaré la dispersión de la muestra total.

Este hecho, además de que incide en el coeficiente de correlación del ingreso con otras variables, tendrá implicancias en la forma de la distribución del ingreso y los estimadores de pobreza y desigualdad.

Todas las decisiones que se asuman en el tratamiento de los datos se verán reflejadas en los valores de los indicadores que se construyan, e inciden en las conclusiones de la investigación.

En la figura 5 se ilustran las consecuencias de utilizar el método de promedios. En 5a se observa la distribución de frecuencia para las observaciones completas, en donde se asume que la desviación estándar del ingreso es 310 pesos. En 5b se considera que el 30% de los datos no tiene información, por lo que la desviación estándar se incrementa a 350 pesos, y se observa que cuando las omisiones son sustituidas por el promedio, la desviación estándar se reduce a 249 pesos (5c).

**FIGURA 5**  
**EFFECTO DE LA IMPUTACIÓN CON PROMEDIOS**



Fuente: Elaboración propia de los autores.

Bajo este procedimiento de imputación el valor medio de la variable se preserva, pero otros estadísticos que definen la forma de la distribución —varianza, covarianza, quantiles, sesgo, *kurtosis*, etc.— pueden ser afectados.<sup>30</sup> Para Acock (2005), este es el peor de los procedimientos de imputación, y por tanto no recomienda su uso.

## B. Imputación por medias condicionadas para datos agrupados

Una variante del procedimiento anterior consiste en formar categorías a partir de covariables correlacionadas con la variable de interés, e imputar los datos omitidos con observaciones provenientes de la submuestra que comparte características comunes (Acock y Demo, 1994).

<sup>30</sup> En Schafer y Graham (2002), se señala que para una muestra grande el intervalo de confianza está determinado por  $\bar{e} \pm 1,96\sqrt{S^2/n}$ , donde  $\bar{e}$  y  $S^2$  representan la media y varianza y  $n$  el tamaño de muestra. La sustitución de los datos faltantes por la media reduce la amplitud del intervalo de confianza debido a la disminución de la varianza del estimador. Suponiendo un patrón MCAR, la probabilidad de que el intervalo contenga al verdadero valor del parámetro es aproximadamente  $2\phi(1,96r)-1$ , donde  $\phi$  es la función de distribución normal y  $r$  es la tasa de respuesta. Con un 25% de datos faltantes ( $r=0,75$ ) la cobertura es 86% y la tasa de error es casi tres veces mayor.

Al igual que el procedimiento de medias, en este caso se asume que los datos faltantes siguen un patrón MCAR y existirán tantos promedios como categorías se formen, lo cual contribuye a atenuar los sesgos en cada celda pero de ninguna manera los elimina.

En la medida que la falta de información por categoría sea baja, los sesgos disminuyen pero no desaparecen. No obstante, no se sugiere utilizar este procedimiento en la medida de que se disponga de una mejor alternativa para sustituir la información omitida.

### C. Imputación con variables ficticias

Cohen y Cohen (1983) y Cohen *et al.* (2003), popularizaron esta metodología, la cual consiste en crear una variable indicador para identificar las observaciones con datos faltantes.

Suponga que la variable predictora ( $y$ ) es la condición de ocupación (“1” si no se conoce la condición de ocupación del entrevistado y “0” si la persona manifiesta estar ocupada),<sup>31</sup> entonces a las personas con datos faltantes se les asigna la media de la variable condición de ocupación.

Si se estima el modelo de regresión  $y = \alpha + \beta Z$ , el valor de  $\beta$  sería el mismo que se obtendría si se utilizará la opción *listwise*, y  $\beta$  daría cuenta de la diferencia entre la variable indicador  $Z$  y la media de  $y$ . Al igual que el procedimiento LD, la imputación por variables ficticias genera inconsistencias en la capacidad explicativa de los estimadores.

Schafer y Graham, (2002), sugieren evitar la aplicación de “trucos” que aparentemente resuelven problemas, pero que introducen sesgos en la interpretación de resultados.

Suponga que se desea estimar la relación entre  $Y$  y  $X$ , donde algunos valores de  $X$  presenta datos omitidos “.”. Se reemplazan los datos faltantes por “0”, y se crea una variable ficticia  $Z$  que asume el valor “1” si  $X=0$  y “0” en otro caso.

Los autores afirman que esta forma de proceder altera la interpretación de los coeficientes de regresión. En el modelo original,  $E(Y) = \beta_0 + \beta_1 X$ ,  $\beta_0$  y  $\beta_1$  representan la ordenada al origen y la pendiente de la recta de regresión respectivamente, en tanto que en el modelo extendido  $E(Y) = \beta_0 + \beta_1 X + \beta_2 Z$ ,  $\beta_0$  y  $\beta_1$  equivalen a la ordenada al origen y la pendiente de los individuos que respondieron, en tanto que  $\beta_0 + \beta_2$  es la media de la variable  $Y$ , entre las observaciones que no tuvieron respuesta.

Debido a las confusiones que introduce este procedimiento, no se recomienda el uso de esta técnica de imputación.

### D. Imputación mediante una distribución no condicionada

Con el propósito de preservar la distribución de probabilidad de las variables con datos incompletos, los estadísticos de encuestas desarrollaron el procedimiento de imputación no paramétrico denominado *hot-deck* (Madow, Nisselson y Olkin, 1983).

El método<sup>32</sup> tiene como objetivo llenar los registros vacíos (receptores) con información de campos con información completa (donantes), y los datos faltantes se reemplazan a partir de una selección aleatoria de valores observados, lo cual no introduce sesgos en la varianza del estimador.

El algoritmo consiste en ubicar registros completos e incompletos, identificar características comunes de donantes y receptores, y decidir los valores que se utilizarán para imputar los datos omitidos. Para la aplicación del procedimiento es fundamental generar agrupaciones que garanticen

<sup>31</sup> El promedio de la variable  $Z$  estima el porcentaje de datos omitidos.

<sup>32</sup> Este método ha sido utilizado desde hace muchos años por la Oficina del Censo de los Estados Unidos de América.

que la imputación se llevará a cabo entre observaciones con características comunes, y la selección de los donantes se realiza en forma aleatoria evitando que se introduzcan sesgos en el estimador de la varianza.<sup>33</sup>

De acuerdo a la figura 6, el algoritmo identifica las omisiones y sustituye el valor “.”, por el ingreso de algún registro “similar”, utilizando para ello un conjunto de covariables correlacionadas con la variable de interés. La aplicación del método, requiere adoptar algún criterio que permita identificar cuál de los valores observados será utilizado en la imputación.

**FIGURA 6**  
**SUSTITUCIÓN DE VALORES CON COVARIABLES**

Folio	Ingreso	Rama	Categoría	Escolaridad	Ponderador
1	400	1	2	3	50
2	500	4	3	4	70
3	.	4	3	3	125
4	.	1	3	3	200

Fuente: Elaboración propia de los autores.

Existen variantes del procedimiento *hot-deck*. El “algoritmo secuencial”, parte de un proceso de ordenación de los datos en cada subgrupo y selecciona donantes en la medida que recorre el archivo de datos. Su aplicación supone que la falta de respuesta se distribuye en forma aleatoria en cada una de las categorías, pero en caso de que la falta de respuesta se concentre en un estrato con pocas observaciones, es posible que se generen estimadores sesgados en la medida que el procedimiento seleccione varias veces el mismo donante.<sup>34</sup>

Por su parte, el “método aleatorio” identifica registros sin datos y elige en forma estocástica al donante. También existe la posibilidad de que el donante sea el “vecino más cercano” al registro sin datos, y la selección se efectúa a partir de la definición de criterios de distancia.<sup>35</sup>

El *hot-deck* y las variantes que se han comentado se consideran mejores opciones que los procedimientos *listwise deletion*, *pairwise deletion*, y es superior los métodos de medias condicionadas y no condicionadas, ya que no introduce sesgos en el estimador y su error estándar. Si se desea preservar la distribución de probabilidad de las variables imputadas, conforme a la opinión de algunos autores se considera que el procedimiento *hot-deck* es más eficiente que el algoritmo la imputación múltiple y la regresión paramétrica (Durrent, 2005).

## E. Imputación por regresión

Ante la presencia de un patrón de datos faltantes MCRA es posible utilizar modelos de regresión para imputar información en la variable  $Y$ , a partir de un grupo de covariables  $(X_1, X_2, \dots, X_p)$  correlacionadas.

El procedimiento consiste en eliminar las observaciones con datos incompletos, y ajustar una ecuación de regresión para predecir los valores de  $\hat{y}$  que serán utilizados para sustituir los valores que faltan, de modo que el valor de  $\hat{y}$  se construye como una media condicionada de las covariables  $X$ 's.

<sup>33</sup> La aplicación del algoritmo conlleva a formar estratos con distintas variables y efectuar la imputación aleatoria en cada estrato.

<sup>34</sup> Para mayores detalles consúltese Lohr (1999) y Lettonen y Pahkinen (1996).

<sup>35</sup> Un procedimiento de imputación similar es el llamado *cold-deck*, el cual utiliza datos exógenos –censos o registros administrativos– para sustituir información.

Este método no se sugiere aplicar cuando el análisis secundario de datos involucra técnicas de análisis de covarianza o de correlación, ya que sobreestima la asociación entre variables, y en modelos de regresión múltiple puede sobredimensionar el valor del coeficiente de determinación  $R^2$ .

Si el método se aplica por estrato (subgrupos), es necesario garantizar suficientes grados de libertad (observaciones completas por subgrupo).<sup>36</sup> En este caso, a pesar de que el sesgo del estimador disminuye el modelo asignará el mismo valor a un grupo de observaciones, lo cual afecta el estimador de la varianza, el coeficiente de correlación de las covariables y la variable imputada, y en los modelos de regresión multivariada el  $R^2$  es sesgado.

Una variante a este procedimiento es la imputación por “regresión estocástica”, en donde los datos faltantes se obtienen con un modelo de regresión más un valor aleatorio asociado al término de error. Este procedimiento garantiza variabilidad en los valores imputados, y contribuye a reducir el sesgo en la varianza y en el coeficiente de determinación del modelo.

## F. ¿Cuándo es adecuada la imputación simple?

A pesar de que se reconoce que los métodos de imputación múltiple son más adecuados que los simples, la literatura da cuenta de situaciones en que éstos entregan resultados satisfactorios. A pesar de ello, no es posible definir reglas para decidir cuándo es factible favorecer la aplicación de un método simple, por lo que se recomienda actuar con prudencia y asumir en cada caso las mejores decisiones.

Si se trabaja con datos de una encuesta con diseño complejo, en donde el porcentaje de datos faltantes es bajo y está concentrado en una submuestra con características especiales, es probable que un método de imputación simple reproduzca mejor las características de la subpoblación de interés que un algoritmo que considera a toda la muestra.

## G. Estimación por máxima verosimilitud (MV)

Los métodos de máxima verosimilitud se pueden aplicar en cualquier problema de estimación. En el análisis de datos omitidos, y asumiendo que los datos faltantes siguen un patrón MAR, se demuestra que la distribución marginal de los registros observados está asociada a una función de verosimilitud para un parámetro  $\theta$  desconocido, bajo el supuesto de que el modelo es adecuado para el conjunto de datos completo.

De acuerdo a Little y Rubin (1987), a esta función se le conoce como la función de verosimilitud, la cual ignora el mecanismo que generó los datos faltantes (verosimilitud de los datos observados conforme a Shafer y Graham (2002).

El procedimiento para estimar los parámetros de un modelo utilizando una muestra con datos faltantes se resume a continuación:

- (i) Estimar los parámetros del modelo con los datos completos con la función de máxima verosimilitud.
- (ii) Utilizar los parámetros estimados para predecir los valores omitidos.
- (iii) Sustituir los datos por las predicciones, y obtener nuevos valores de los parámetros maximizando la verosimilitud de la muestra completa.

<sup>36</sup> La teoría estadística establece como mínimo 30 observaciones por celda para que de acuerdo al teorema del límite central se pueda asumir que, en el límite, la variable observada se asemeja a una distribución normal. No obstante, en el análisis de encuestas complejas esta “regla” debe interpretarse con cautela, considerando el esquema de selección de la muestra y las probabilidades de selección de las unidades de observación (los factores de expansión).

- (iv) El algoritmo se aplica hasta lograr la convergencia, la cual se obtiene cuando el valor de los parámetros no cambia entre dos iteraciones sucesivas.

Un procedimiento eficiente para maximizar la verosimilitud cuando existen datos faltantes es el algoritmo *Expectation-Maximization* (EM) (Dempster, Laird y Rubin (1977)), (véase recuadro 1).

#### RECUADRO 1 ESTIMACIÓN DE MÁXIMA VEROSIMILITUD CON EL ALGORITMO EM

Suponga una muestra de tamaño  $n$  de una variable aleatoria, en donde algunas de las variables no tienen información. Asumimos que los datos ausentes se generan al azar; es decir, que no existe relación entre los valores observados y los omitidos. Las dos situaciones más importantes de valores faltantes son: i) Algunos elementos de la muestra están completos ( $x_1, \dots, x_n$ ) y otros no tienen datos ( $x_{n+1}, \dots, x_m$ ). ii) Algunas variables no tienen datos.

Suponga que se trabaja con una matriz de datos  $Y=(y_1, \dots, y_n)$ , donde  $y_i$  es un vector de dimensión  $p_1 \times 1$ , y un conjunto de datos ausentes  $Z=(z_1, \dots, z_m)$ , con  $z_i$  un vector de dimensión  $p_2 \times 1$ , y el problema consiste en estimar el vector de parámetros  $\theta$  con la información disponible.

La función de densidad de probabilidad conjunta de las variables  $(Y, Z)$  se escribe como:  $f(Y, Z | \theta) = f(Z | Y, \theta) f(Y | \theta)$ , por lo que se tiene que  $\log f(Y | \theta) = \log f(Y, Z | \theta) - \log f(Z | Y, \theta)$ . En el procedimiento de máxima verosimilitud, el primer miembro de la expresión  $\log f(Y | \theta)$  es la función de los datos observados, cuya maximización en  $\theta$  genera el estimador de máxima verosimilitud. El término  $\log f(Y, Z | \theta)$  es la función si hubiésemos observado la muestra completa, y  $f(Z | Y, \theta)$  proporciona la densidad de los datos ausentes, toda vez que conocemos la muestra y el vector de parámetros  $\theta$ .

La función de verosimilitud es  $L(\theta | Y) = L_c(\theta | Y, Z) - \log f(Z | Y, \theta)$ . El algoritmo EM es un procedimiento iterativo para encontrar el estimador de máxima verosimilitud (MV) de  $\theta$ , utilizando la función  $L_c(\theta | Y, Z)$ . La aplicación del algoritmo se logra ejecutando los pasos siguientes:

(a) Partiendo de un estimador inicial,  $\hat{\theta}^{(i)}$  en la primera iteración ( $i=1$ ) se calcula la esperanza matemática de las funciones de los valores ausentes que aparecen en la función de verosimilitud completa,  $L_c(\theta | Y, Z)$ , con respecto a la distribución de  $Z$  dados los valores de  $\hat{\theta}^{(i)}$  y los datos observados  $Y$ . Sea  $L_e^*(\theta | Y) = E_{Z/\hat{\theta}^{(i)}} [L_c(\theta | Y, Z)]$  el resultado de la operación que se denomina el paso **E** (cálculo del valor esperado) del algoritmo. Cuando  $L_c(\theta | Y, Z)$  sea una función lineal de  $Z$ , los valores ausentes se sustituyen por los valores esperados dado el vector de parámetros  $\theta$ .

(b) Se maximiza la función  $L_c(\theta | Y, Z)$  con respecto a  $\theta$ . Este es el paso **M** de maximización del algoritmo. El paso M equivale a maximizar la verosimilitud completa donde se han sustituido las observaciones faltantes por estimadores.

Sea  $\hat{\theta}^{(i+1)}$  el estimador obtenido en el paso M. Con este valor se vuelve a ejecutar el paso E, y se itera hasta lograr la convergencia; es decir, hasta que la diferencia  $\|\hat{\theta}^{(i+1)} - \hat{\theta}^{(i)}\|$  sea suficientemente pequeña. Se demuestra que este algoritmo maximiza la  $L(\theta | Y)$  (Dempster, Laird y Rubin, 1977), y la verosimilitud aumenta en cada iteración hasta lograr la convergencia.

Fuente: Elaboración propia de los autores.

Los valores son imputados mediante un proceso iterativo y en cada iteración se incorpora más información, y el procedimiento converge cuando los valores de la matriz de covarianza son prácticamente los mismos que los obtenidos en la iteración anterior. En general, el algoritmo converge rápido pero dependerá del porcentaje de datos faltantes que se observen en la muestra analizada.<sup>37</sup>

<sup>37</sup> Este método se encuentra incorporado en el paquete SPSS en el módulo MVA (*Missing value Analysis*). En von Hippel (2004), se critica el algoritmo utilizado por el SPSS, y el autor indica que no se especifica la manera en que se genera el término aleatorio para cada dato imputado que se incluye en el nuevo archivo de datos. Según von Hippel "El método EM genera estimadores asintóticamente insesgados, pero los resultados se limitan a producir valores puntuales (sin errores estándar) de la media, la varianza y la covarianza. El procedimiento MVA imputa también valores utilizando el algoritmo EM, pero genera estimadores de medias, varianza y covarianzas sin variación residual, lo cual puede conducir a estimadores sesgados" (von Hippel, 2004 p.160). El autor también afirma que el método de imputación por regresión del SPSS genera resultados sesgados, debido a que el procedimiento estima los parámetros del modelo utilizando sólo los datos con información (*pairwise*).

## IX. Imputación múltiple

---

### A. ¿Qué es imputación múltiple (IM)?

Imputar significa sustituir observaciones, ya sea porque se carece de información (*missing values*) o porque se detecta que algunos de los valores recolectados no se corresponden con el comportamiento esperado (*outliers*). En esta situación, es común que se desee reponer las observaciones y se decida aplicar algún método de sustitución de datos y de imputación.

No obstante, utilizar algún procedimiento inapropiado puede generar más problemas de los que resuelve, introduciendo sesgos en el valor de los estimadores y en su error estándar, al tiempo que podría distorsionar la potencia de las pruebas de hipótesis (Little y Rubin, 1987), lo que sugiere reflexionar acerca de la mejor manera de obtener estimadores que generen inferencia válida a partir de datos imputados. En Rubin (1987), se hace esta reflexión y se propone como solución el método de imputación múltiple (IM).

IM utiliza métodos de simulación de Monte Carlo y sustituye los datos faltantes a partir de un número ( $m > 1$ ) de simulaciones que, de acuerdo al autor, se ubica entre 3 y 10. La metodología consta de varias etapas, y en cada simulación se analizan la matriz de datos completos a partir de métodos estadísticos convencionales y posteriormente se combinan los resultados para generar estimadores robustos, su error estándar e intervalos de confianza. En la figura 5 se esquematiza la manera en que operan los métodos de imputación múltiple, y en recuadro 2 se sintetiza la metodología propuesta por Rubin (1987).

## RECUADRO 2 PROCEDIMIENTO DE IMPUTACIÓN MÚLTIPLE

El procedimiento propuesto por Rubin (1987) combina distintos estimadores generados a partir de  $m$  imputaciones. Utilizando la notación de Schafer (1997), considere una variable  $\hat{Q}$  y  $U$  el estimador de su varianza. Después de generar  $m$  conjuntos de datos mediante simulaciones, se tienen  $m$  estimadores de  $\hat{Q}$  y de  $U$

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^m \hat{Q}_i \quad (1)$$

Existen dos componentes de la varianza de  $\bar{Q}$ . La varianza de cada imputación,

$$\bar{U} = \frac{1}{m} \sum_{i=1}^m \hat{U}_i \quad (2)$$

y la varianza entre las imputaciones,

$$B = \frac{1}{m-1} \sum_{i=1}^m (\hat{Q}_i - \bar{Q})^2 \quad (3)$$

Por tanto, la varianza total ( $T$ ) se obtiene sumando las expresiones (2) y (3), corrigiendo el número finito de imputaciones por el valor  $\frac{m+1}{m}$ .

$$T = \bar{U} + \frac{(m+1)}{m} B \quad (4)$$

$B/U$  indica cuánta información corresponde a los datos faltantes, y se estima a partir de  $\frac{\gamma}{1-\gamma}$ , en donde  $\gamma$  representa la fracción de información que se pierde por falta de respuesta. En el caso de que  $\gamma = 0$  se observa que  $B=0$ .

El intervalo de confianza se obtiene por medio de:

$$\bar{Q} \pm t_{gl} \sqrt{T} \quad (5)$$

y los grados de libertad de  $t$  se calculan como,

$$gl = (m-1) \left(1 + \frac{1}{r^2}\right) \quad (6)$$

Con

$$r = \left(1 + \frac{1}{m}\right) \frac{B}{\bar{U}} \quad (7)$$

Fuente: Elaboración propia de los autores.

A pesar de las bondades que se le atribuyen a este procedimiento, los métodos de IM no deben ser considerados de manera acrítica como la mejor opción estadística para la sustitución de datos. Cada situación es diferente, y dependiendo de la variable que se analice, del porcentaje de no respuesta y de su patrón de comportamiento, es probable que se presenten situaciones en las que alguno de los métodos descritos en este trabajo entreguen resultados más adecuados que el procedimiento IM (Little, 1986 y Robins, *et al.*1994).

Como fue señalado en el acápite anterior, los estimadores con datos incompletos se pueden obtener por máxima verosimilitud a partir del algoritmo EM, y según Schafer (1999), los estimadores pueden resultar más eficientes que los que se obtendrían con IM, debido a que no requiere de simulaciones ni se depende de un modelo estadístico o econométrico.<sup>38</sup>

<sup>38</sup> En Schafer (1999), se indica que disponiendo de tiempo y recursos suficientes, es posible encontrar algoritmos que generen mejores estimadores que el método IM para cierto tipo de problemas.



## B. Procedimiento de imputación múltiple (IM)

El planteamiento propuesto por Rubin se describe a continuación y se esquematiza en la figura 7:

Sea  $Q$  una variable aleatoria y supongamos que se desean estimar la media, la varianza o su coeficiente de correlación con otras variables. Considere  $X$  la matriz de datos disponibles la cual se puede descomponer como  $X=(X_{obs}, X_{mis})$ , y  $(X_{obs}, X_{mis})$  es valor del estimador de  $Q$  que se genera a partir de los datos y  $U=U(X_{obs}, X_{mis})$  el error estándar de  $\hat{e}$ . Para el conjunto de datos completos se sabe que  $(\hat{Q}-Q)/\sqrt{U} \sim N(0,1)$ , lo que sugiere estandarizar los datos.

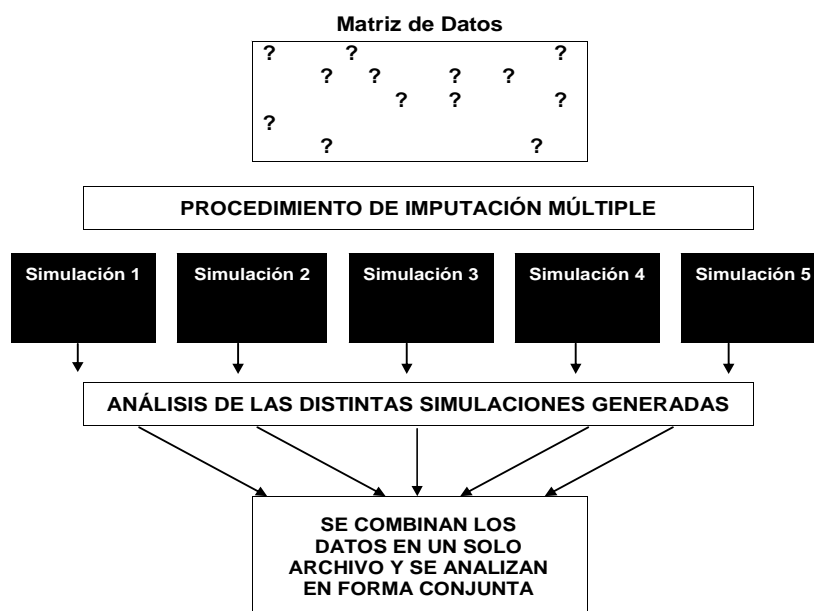
En ausencia de datos faltantes, suponga que se dispone de  $m>1$  simulaciones independientes de datos imputados  $X(1)_{mis}, \dots, X(m)_{mis}$ , que permite calcular el valor del estimador  $\hat{Q}^{(i)} = \hat{Q}(X_{obs}, X^{(i)}_{mis})$  y sus respectivos errores  $U^{(i)} = U(X_{obs}, X^{(i)}_{mis})$  ( $i = 1, \dots, m$ ). El estimador de  $Q$  es el promedio de los estimadores:  $\bar{Q} = m^{-1} \sum \hat{Q}^{(i)}$ .

El error estándar de  $\bar{Q}$  se calcula a partir de la varianza entre las distintas imputaciones  $B = (m-1)^{-1} \sum (\hat{Q}^{(i)} - \bar{Q})^2$  y debido a que la varianza de cada una de las imputaciones es  $\bar{U} = m^{-1} \sum U^{(i)}$ , el estimador de la varianza total es  $T = (1+m^{-1})B + \bar{U}$ .

Las pruebas de hipótesis y los intervalos de confianza se construyen a partir de una aproximación a la  $t$  de Student por medio de:  $(\bar{Q} - Q) / \sqrt{T} \sim t_v$ , donde los grados de libertad se determinan por medio de:  $v = (m-1)[\bar{U} / (1+m^{-1})B]^2$

Rubin propuso la medida  $r = (1+m^{-1})B / \bar{U}$  para determinar el incremento relativo de la varianza debido a la presencia de datos faltantes, y la tasa de datos faltantes se aproxima a  $\lambda = r / (1+r)$  o  $\lambda = [(r+2)/(v+3)] / (1+r)$ .

**FIGURA 7  
IMPUTACIÓN MÚLTIPLE**



Fuente: Elaboración propia de los autores.

Esta metodología es intuitiva y fácil de entender y su aplicación asume los siguientes supuestos. (i) El patrón de datos faltantes es aleatorio (MAR), lo cual significa que la probabilidad de que existan datos omitidos en la variable X dependen de otras variables pero no de X. (ii) Se requiere que el modelo (estadístico o econométrico) utilizado para generar los datos imputados sea apropiado; es decir, que exista correlación alta entre la variable a imputar y el vector de covariables que se utilizará para modelar los datos que se utilizarán como sustitutos. (iii) Finalmente, también es necesario que el modelo de análisis guarde relación con el que se utilizó para efectuar el procedimiento de imputación.<sup>39</sup> Todas estas condiciones se describen en Rubin (1987) y Rubin (1996).

El problema, como lo explica (Allison, 2000), es que la práctica resulta común que los supuestos aludidos no se cumplan, ya que por diversas razones los patrones de datos faltantes no se asimilan a un proceso MAR. Desafortunadamente, señala el autor, no hay otras alternativas viables, a pesar de que siempre existe la posibilidad de aplicar procedimientos *ad hoc* para que generen resultados adecuados.

La validez del método se sustenta en la manera en que se obtienen las imputaciones, ya que no es posible derivar conclusiones apropiadas si las imputaciones se generan de manera arbitraria. Los estimadores finales se calculan como un promedio de los valores sustituidos, lo cual garantiza que se mantenga la variabilidad en los valores imputados.

## C. Consideraciones acerca del procedimiento IM

En Schafer (1999), se da respuesta a las interrogantes que surgen a los usuarios del procedimiento IM, algunas de las cuales se analizan a continuación.

Es genuino preguntarse por qué se debe aplicar IM, si eliminar observaciones es un procedimiento más fácil. En este trabajo se han analizado los sesgos que se generan en el proceso de inferencia cuando la falta de respuesta es importante y se decide eliminar observaciones.

Al eliminar registros de la base de datos, se asume que las observaciones completas tienen la misma distribución de probabilidad que las omitidas; es decir, que los datos faltantes representan una submuestra aleatoria de la muestra total. No obstante, y dado que no es posible corroborar si la información faltante tiene un comportamiento particular, se mantendrá la duda de cómo se comporta el grupo de datos sin información respecto a la distribución de las variables de interés.

Otra interrogante que surge es por qué se deben generar varias imputaciones. En el caso de que la falta de respuesta no sea importante,<sup>40</sup> es probable que una imputación sea suficiente para obtener estimadores insesgados.

Esto se logra si se comprueba que la imputación no ha modificado las características estadísticas de la variable de análisis —respecto a parámetros de referencia conocidos—, y se demuestra que no se han generado cambios relevantes en la forma de su distribución —medidas de tendencia central, dispersión, asimetría y *kurtosis*—. En estas circunstancias es posible concluir que la imputación simple resultó adecuada, y no es necesario generar simulaciones adicionales.

La objeción que frecuentemente se le hace a esta forma de proceder, es que a partir de un algoritmo simple no es posible estimar el error de los estimadores que utilizan valores imputados, lo cual si se puede determinar cuando se aplica el procedimiento IM.

<sup>39</sup> Por ejemplo, si se deseara imputar los salarios de los ocupados para determinar las tasas de retorno a la educación, para la sustitución de los datos faltantes se debieran utilizar los años de escolaridad, la experiencia y su cuadrado, ya que se conoce que estas variables explican un porcentaje importante de la formación de los salarios.

<sup>40</sup> Schafer (1999), indica que cuando la tasa de no respuesta es menor al 5%, los métodos de imputación simple pueden generar resultados adecuados.

En situaciones en que la no respuesta es baja se sugiere aplicar IM, especialmente cuando en el análisis se utilizan técnicas multivariadas. En la actualidad, diversos paquetes han incorporado rutinas con algoritmos de imputación múltiple, pero se sugiere tener cautela y sentido común para no utilizar este procedimiento como “caja negra”, ya que es posible que generen valores inconsistentes que introduzcan sesgos en los resultados.

Se afirma que el método IM es capaz de generar resultados robustos con un número pequeño de iteraciones. De acuerdo con Rubin (1987, p.114), la eficiencia relativa de  $m$  imputaciones es aproximadamente  $(1 + \lambda/m)^{-1}$ , en donde  $\lambda$  es la tasa de registros sin información. Rubin indica que aún en situaciones en donde se tenga 50% de datos faltantes, el estimador generado a partir de  $m=5$  imputaciones tiene una desviación estándar que es sólo 5% mayor que otra simulación basada en  $m=\infty$  iteraciones, debido a que  $\sqrt{1+0,5/5}=1.049$ . También señala, que para tasas de respuesta inusualmente altas sólo se requiere generar entre 5 y 10 imputaciones.

Muchas personas se sorprenden con estos resultados, por lo que en la figura 8 se muestra la relación que existe entre la eficiencia de los estimadores y el número de imputaciones que los genera en donde  $\lambda$  es la tasa de valores faltantes.

**FIGURA 8**  
**EFICIENCIA DEL PROCESO IM**

Imputaciones	Tasa de valores faltantes ( $\lambda$ )				
	0,1	0,3	0,5	0,7	0,9
$m$	0,1	0,3	0,5	0,7	0,9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

Fuente: Elaboración propia de los autores.

Se comprueba que en situaciones en que el nivel de valores omitidos es superior al 50%, el número de imputaciones que se requieren para obtener estimaciones adecuadas no es muy grande.

Por ejemplo, si se registra una tasa de no respuesta del 50% ( $\lambda=.5$ ), con  $m=3$  imputaciones la eficiencia del método IM es 86%, en tanto que si el número de simulaciones se incrementa a  $m=5$  la eficiencia aumenta a 91%.

A pesar de que se afirma que la IM es eficiente ante tasas de no respuesta importantes, se sugiere asumir esta afirmación con cautela. Suponga que como resultado del trabajo de campo en una encuesta de hogares se comprueba que más del 50% de los datos de la variable ingreso no tiene información. En este contexto, y antes de aplicar el método de imputación múltiple (IM) es necesario cuestionar la validez del estudio —incluso analizar la posibilidad de prescindir de esa encuesta—, y se debe tener consciencia de que sustituir la información faltante, significa imputarle valores a más de la mitad de los entrevistados. Es decir, la mitad de los datos que no fueron recopilados en el terreno, serán sustituidos por valores provenientes de un modelo estadístico.

Si el estudio se efectuó para apoyar el diseño de políticas públicas, es pertinente preguntarse qué tipo de decisiones se pueden adoptar si se sabe que sólo la mitad de la información corresponde a datos proporcionados por la población objetivo, y el resto proviene de estimaciones generadas por método estadístico sustentando en un algoritmo de simulación.

Para los trabajos académicos situaciones como la descrita puede resultar adecuada para concluir una de investigación. No obstante, cuando de las conclusiones del estudio depende la asignación de fondos públicos para la inversión social, se debe ser riguroso y estar consciente de las malas decisiones que se pueden adoptar, ante información modelada que se aleja de la realidad.

Un procedimiento que compite con los métodos de IM es el procedimiento de máxima verosimilitud (MV) que utiliza el algoritmo EM. Ambas propuestas aplican métodos numéricos y simulaciones de Monte Carlo, y se demuestra que para tamaños de muestra grandes generan resultados similares. No obstante, también se ha comprobado que en el caso de muestras pequeñas la técnica IM produce resultados más robustos que el método de MV (Schafer, 1999).

Algunos investigadores se preguntan si los métodos IM guardan relación con las cadenas de Markov *Markov Chain Monte Carlo* (MCMC). En ese sentido, se debe aclarar que MCMC es una colección de procesos de simulación generados por métodos de selección aleatoria mediante cadenas de Markov, y es uno de los procedimientos que se considera más adecuados para generar imputaciones cuando se está en presencia de problemas de estimación no triviales.

MCMC utiliza simulación paramétrica generando muestras aleatorias a partir de métodos bayesianos, y en el método IM este procedimiento se aplica para generar  $m$  selecciones independientes de valores faltantes, las cuales se utilizan en la etapa de inferencia.

Debido a que es común que el patrón de datos omitidos esté condicionado por las características de la población de interés, surge la inquietud de qué hacer cuando la no respuesta no se puede ignorar (MNAR) (i.e. no es MCAR ni MAR).

Schafer (1999), afirma que en la aplicación del paradigma IM no es necesario suponer que el patrón de observaciones faltantes debe ser ignorado, y señala que los procedimientos desarrollados por Rubin pueden ser aplicados a cualquier tipo de modelos de simulación. También reconoce que el tema está en desarrollo, y que resolver satisfactoriamente este tipo de inquietudes aún representa un importante desafío.

## **X. Imputación de datos en encuestas complejas**

---

Los métodos de imputación hacen supuestos acerca de la manera en que se distribuyen los datos faltantes, pero en ningún caso hacen referencia al mecanismo que se aplicó en la selección de las unidades de observación. Se asume, de manera errónea, que los datos provienen de una muestra aleatoria y que todas las unidades tienen la misma probabilidad de selección.

Las encuestas de hogares se ejecutan a partir de diseños de muestra complejos, y distintos autores han cuestionado la validez de los métodos de imputación múltiple IM (Binder, 1996; Binder y Weimin, 1996).

Se reconoce que en todas las encuestas existe ausencia de datos, y lo común es buscar algún procedimiento que permita completar la información. Todos los procedimientos existentes hacen énfasis en la necesidad de analizar el patrón de datos faltantes, pero pasan por alto el hecho de que las unidades de observación tienen probabilidades de selección diferentes.

Aún en situaciones en que la falta de respuesta sea baja, se sugiere analizar el valor los ponderadores asociados a los datos faltantes. Es probable que unos cuantos hogares de la muestra tengan una representación importante en la población, y un criterio de imputación mal aplicado puede introducir sesgos difíciles de identificar y evaluar.

En los diseños de muestra complejos, la elección de las observaciones depende de la manera en que se estratificó y se conglomeró el marco de muestreo, así como del vector de ponderaciones asociado a las distintas unidades en muestra.

El método de imputación simple (IM) y el algoritmo de máxima verosimilitud EM no consideran este hecho —tampoco el resto de los métodos analizados—, y en el caso del procedimiento IM las  $m > 1$  simulaciones se generan asumiendo que se está en presencia de una muestra aleatoria, en la que todas las observaciones tienen la misma probabilidad de selección y el mismo factor de expansión.

En Binder y Weimin (1996), se demuestra que bajo el supuesto de que los datos se generen a partir de un diseño aleatorio simple y sin reemplazo, los procesos de imputación simple y múltiple generan estimadores adecuados para la media y totales en el caso de que se apliquen métodos *bayesianos (bootstrap)*, pero esto no se cumple cuando se utiliza un criterio determinístico como la imputación aleatoria o el método de promedios.

En el mismo trabajo, se analizan los métodos de imputación por subgrupos —como el *hot-deck*—, y se distinguen las observaciones con respuesta y sin respuesta y se asume que la no respuesta es independiente en cada subgrupo (Sarnald, Swenson y Wretman, 1992). Se concluye que en esquemas de selección aleatoria es posible obtener una adecuada estimación de la varianza de los estimadores con datos imputados.

No obstante, en la medida que el diseño involucre estratificación y conglomeración para elegir las unidades que formarán parte de la muestra, las expresiones que se deben aplicar para estimar la varianza se complican. Por lo tanto, Binder (1996), conjetura que los procedimientos de imputación múltiple no resultan adecuados en diseños complejos en los que existen al menos dos etapas de selección y conglomeración y las probabilidades de selección son desiguales.

En los procedimientos de imputación por subgrupos (*hot-deck* o medias), se debe tener presente que los donantes y receptores pueden pertenecer a un mismo grupo, lo que no significa que provengan del mismo conglomerado o estrato de muestreo, y es probable que tengan distintas probabilidades de selección.

Ante esta situación, surge la pregunta de cómo se afectan los estimadores ante un diseño de muestra polietápico en donde las unidades de selección tienen probabilidades desiguales y son elegidas en varias etapas que involucran la estratificación del marco de muestreo y la formación de conglomerados —compactos o dispersos— con las unidades de segunda etapa.

Esta pregunta no se aborda en ninguno de los métodos descritos, y no forma parte de las preocupaciones del método IM. Fay (1993), señala que en la abundante literatura que existe sobre el tema de IM nunca se ha aclarado cuáles son las situaciones y supuestos bajo los cuales la aplicación del método genera resultados satisfactorios. El autor señala, “el capítulo 1 del libro de Rubin (1987), promete mucho y daría la impresión de que se ha inventado una herramienta de uso universal”, y en Fay (1991; 1992), se presentan contraejemplos en los que se demuestran las limitaciones de los métodos de imputación múltiple.<sup>41</sup>

En Montaquila y Jernigan (1997), se desarrollan expresiones para calcular la varianza para esquemas de muestreo aleatorio simple y aleatorio estratificado, cuando el procedimiento de imputación utilizado es el *hot-deck*. Sin embargo, en ambos casos se asume que las observaciones tienen la misma probabilidad de selección.

---

<sup>41</sup> En “Multiple Imputation After 18+ Years”, Rubin (1996), revisita el tema y se hace cargo de las críticas recibidas. Inicia su exposición diciendo que IM, como cualquier otro método estadístico, algunas veces no es adecuado y en muchas otras situaciones entrega resultados apropiados. El autor sostiene que para los usuarios finales que desconocen los detalles técnicos de cómo fueron generados los datos y las causas en las que se generó la falta de respuesta, IM es el procedimiento más adecuado para sustituir datos faltantes. Analiza los ejemplos y las críticas de Fay, y retoma argumentos de trabajos anteriores para defender su postura. No obstante, en ningún momento se hace cargo de los comentarios de Binder (1996), en el sentido de que en los diseños complejos —estratificados, de conglomerados y con probabilidades desiguales—, los procedimientos propuestos por Rubin no siempre resultan adecuados, y a la necesidad de considerar el diseño de muestra en la generación de los valores imputados.

Debido a que en la práctica lo habitual es disponer de información proveniente de encuestas complejas, se aconseja actuar con cautela ya que persiste el desafío de desarrollar algoritmos de imputación robustos que tengan en cuenta el diseño de la muestra y las probabilidades de selección de las observaciones.

En la figura 9 se presenta una tipología que resume las principales características de los métodos de imputación que se utilizan con mayor frecuencia. Se indican los supuestos en que se sustenta cada método, con relación al patrón observado en los datos faltantes, la forma en que se aplica el procedimiento, así como algunas de sus principales ventajas y desventajas. Finalmente, se da cuenta de su disponibilidad en los paquetes estadísticos utilizados para la elaboración de este trabajo.

**FIGURA 9**  
**TIPOLOGÍA DE LOS MÉTODOS PARA LA IMPUTACIÓN DE DATOS**

Método	Supuestos y patrón de datos faltante	Aplicación	Ventajas	Desventajas	SPSS	SAS	STATA
<b>Determinístico</b>	Patrón de datos faltantes condicionado (MNAR)	Los datos faltantes se imputan conforme a las reglas definidas en los catálogos de crítica y codificación	Los datos se imputan con la información del registro que se analiza, lo cual garantiza consistencia.	Las situaciones que ocurren en la realidad superan la imaginación del personal encargado de preparar criterios ad-hoc. No es posible generar criterios determinísticos para todas las variables.	*	*	*
<b>Repetición</b>	Patrón de datos faltantes condicionado (MNAR)	En las encuestas de hogares continuas se utiliza información pasada de la misma unidad de observación.	Se hace uso de información de las mismas unidades de observación.				
<b>Datos completos (Listwise)</b>	Patrón de datos faltantes MCAR o MAR. Los datos eliminados son una submuestra aleatoria de la muestra.	Se utilizan sólo registros con información.	Fácil de aplicar. Es la opción que por defecto aplican los paquetes SPSS, SAS, STATA.	Se reduce el tamaño de la muestra. Se debilita la significancia estadística de las pruebas.	X	X	X
<b>Datos disponibles (Pairwise)</b>	Patrón de datos faltantes MCAR.	Se utilizan las observaciones que tienen información para todas las variables.	Fácil de aplicar y trabaja con toda la información.	Se distorsiona la relación entre variables.	X		
<b>Medias no condicionadas</b>	Patrón de datos faltantes MCAR.	Las observaciones faltantes se reemplazan por el valor medio de la variable de análisis.	Fácil de entender y aplicar.	Genera estimadores sesgados. Subestima la varianza de los estimadores.	X	X	
<b>Medias por subgrupos</b>	Patrón de datos faltantes MCAR.	Se divide la base de datos en subgrupos utilizando variables correlacionadas. Las observaciones faltantes en el subgrupo de interés se reemplazan por el valor medio de la variable.		Genera estimadores sesgados. Subestima la varianza de los estimadores.	X	X	
<b>Variables binarias</b>	Patrón de datos faltantes MCAR.	Se forman variables binarias para identificar las observaciones con datos faltantes.		Introducen distorsiones en la interpretación de los parámetros de los modelos de regresión.	X		
<b>Reponderación</b>	Patrón de datos faltantes MCAR.	Se forman grupos, y al interior de los mismos se eliminan las observaciones sin datos. Se ponderan las observaciones utilizando la información de la muestra completa. Los factores de reponderación se pueden obtener por modelos.	Las observaciones que permanecen en muestra y los ponderadores ajustados estiman correctamente los totales por celda	Cuando la falta de respuesta por celda es muy alta, es posible que se introduzcan sesgos en el valor de los estimadores y su varianza			
<b>Hotdeck (condicionado a covariables)</b>	Patrón de datos faltantes MAR.	Se divide la base de datos en subgrupos utilizando variables correlacionadas. Los valores faltantes se sustituyen con la información de un registro con información similar en las covariables.	Los donantes y receptores pertenecen a un mismo subgrupo. El número de donantes se condiciona con el uso de covariables.	No siempre es fácil definir un criterio de distancia o similitud entre los posibles donantes y receptores.			X
<b>Hotdeck con regresión (condicionado a covariables)</b>	Patrón de datos faltantes MCAR. Se requiere especificar un modelo, en donde las covariables estén altamente correlacionadas con la variable a imputar.	Se divide la base de datos en subgrupos utilizando variables correlacionadas. Los valores faltantes se sustituyen con el valor medio estimado por la regresión efectuada en el subgrupo de interés.	Fácil de aplicar.	Todas las observaciones pertenecientes al subgrupo tienen el mismo valor imputado. Se subestima la varianza y se introducen sesgos en la correlación. No siempre es fácil encontrar un modelo adecuado para la variable de interés. Se subestima el error estándar del estimador.			X
<b>Regresión</b>	Patrón de datos faltantes MCAR. Se requiere especificar un modelo, en donde las covariables estén altamente correlacionadas con la variable a imputar.			Se subestima el error estándar del estimador. Se subestima la varianza.	X		X
<b>Regresión por subgrupos</b>	Patrón de datos faltantes MCAR. Se requiere especificar un modelo, en donde las covariables estén altamente correlacionadas con la variable a imputar.	Se divide la base de datos en subgrupos utilizando variables correlacionadas. Los valores faltantes se sustituyen con el valor medio estimado por la regresión efectuada en el subgrupo de interés.	Fácil de aplicar.	Genera estimadores sesgados. Subestima la varianza de los estimadores. Genera sesgos de correlación. No siempre es fácil encontrar un modelo adecuado para la variable de interés. Se subestima el error estándar del estimador	X		X
<b>Regresión aleatoria</b>	Patrón de datos faltantes MAR. Se requiere especificar un modelo en donde las covariables estén altamente correlacionadas con la variable a imputar.			No siempre es fácil encontrar un modelo adecuado para la variable de interés. Se subestima el error estándar del estimador. Se subestima la varianza.			X
<b>Máxima Verosimilitud (EM)</b>	Patrón de datos faltantes MAR.		Genera estimaciones robustas basadas en la muestra observada. No efectúa simulaciones.	No siempre están disponibles. Hay que programar el algoritmo que se desea aplicar.	X	X	
<b>Imputación simple</b>	Patrón de datos faltantes MAR. Se requiere especificar un modelo en donde las covariables estén altamente correlacionadas con la variable a imputar.		Utiliza procedimientos estadísticamente robustos.	Requiere que los datos faltantes sigan el patrón MAR. Ocurren situaciones en que los supuestos del método no se cumplen. No es posible conocer el error estándar de los estimadores, ya que sólo efectúa un interacción. No siempre es fácil encontrar un modelo adecuado para la variable de interés.			X
<b>Imputación múltiple</b>	Patrón de datos faltantes MAR. Se requiere especificar un modelo, en donde las covariables estén altamente correlacionadas con la variable a imputar. El modelo que se utiliza para imputar, debe ser el mismo o muy similar al que se utilizará en el análisis secundario de datos.	Se generan m subconjuntos de datos imputados por medio de simulaciones. Se combina en forma apropiada a fin de obtener estimadores robustos.	Muy intuitivo y fácil de entender. Utiliza procedimientos estadísticos robustos. Genera distintas opciones de datos imputados y las combina en forma adecuada. Permite calcular el error estándar de los estimadores.	Requiere que los datos faltantes sigan el patrón MAR. Ocurren situaciones en que los supuestos del método no se cumplen. No siempre es fácil encontrar un modelo adecuado para la variable de interés. Se requiere de paquetes estadísticos de cómputo que contengan algoritmos de cálculo para este propósito. Si no se utilizan con precaución pueden funcionar como cajas negras. Se le deja la responsabilidad a un procedimiento estadístico.		X	X

Fuente: Elaboración propia de los autores.

\* Programando.

Todos los procedimientos incluidos en la tipología (con excepción del método deductivo), confirman que asumen que los datos omitidos siguen un patrón MAR o MCAR, y ninguno de ellos tiene en cuenta el diseño de muestra ni la probabilidad de selección de las observaciones.



## XI. Efectos de la imputación en la pobreza y la desigualdad

---

### A. Objetivos del estudio

La CEPAL genera periódicamente estimaciones de pobreza y desigualdad, utilizando información oficial recopilada a partir de las encuestas de hogares que efectúan las Oficinas de Estadística de América Latina y el Caribe.

Se comprueba que la calidad de los datos es heterogénea entre países, y no es común que las bases de datos se acompañe con documentación que permita a los usuarios conocer las causas que generaron la falta de respuesta, las soluciones que se implementaron y las evaluaciones que se llevaron a cabo para conocer sus implicaciones en la confiabilidad estadística de los estimadores.

En este sentido, es posible afirmar que los datos se imputan tantas veces como usuarios existen, lo cual mantiene latente el peligro de que se generen contradicciones entre los resultados oficiales y las estimaciones independientes generadas con la misma base de datos.

El objetivo de este trabajo es evaluar la sensibilidad de los indicadores de pobreza y desigualdad a distintos procedimientos de sustitución de datos omitidos. Debido a que la incidencia de la pobreza está correlacionada con el ingreso de las familias, se postula que cualquier alteración del ingreso *per capita* incide en el nivel de los indicadores, lo que sugiere elegir un método de imputación robusto que, además de preservar la forma de la distribución del ingreso, evite introducir sesgos en los resultados.

## B. Características de los datos

Los datos utilizados en este trabajo provienen de la Encuesta Permanente de Hogares (EPH), realizada por el Instituto Nacional de Estadística y Censos de Argentina (INDEC) en el 2004.

La EPH es una encuesta de hogares continua que tiene el propósito de recabar información sobre características relevantes de los hogares urbanos de Argentina. En ese sentido, además de recopilar información sobre variables sociodemográficas, la encuesta registra características sobresalientes asociadas a las condiciones de ocupación de la población activa (empleo y desempleo), e indaga sobre el origen de distintos tipos de ingreso recibidos por los miembros del hogar que participan en el mercado de trabajo.

La selección de las viviendas que forman parte de la muestra se lleva a cabo a partir de la aplicación de un diseño de muestreo polietápico, estratificado y por conglomerados, y el tamaño de muestra involucra la selección de 27,303 viviendas lo que permite generar estimaciones confiables para 28 aglomerados urbanos.

La encuesta pone especial énfasis en captar información para las personas que están en condiciones de participar en el mercado de trabajo. En ese sentido, para los propósitos de la EPH, la población en edad de trabajar (PET) está formada por las personas de 10 y más años, y partir de distintos criterios se analiza su condición de actividad, su categoría ocupacional y la rama de actividad económica en la que los ocupados desarrollan su trabajo, entre otras variables de interés.

Para los propósitos de esta investigación, y reconociendo la importancia que las remuneraciones al trabajo tienen en la formación del ingreso familiar, se consideró necesario corregir las omisiones que se observaron en los sueldos y salarios percibidos por los ocupados, las ganancias netas reportadas por las personas que trabajan por su cuenta y las que poseen negocios propios, así como el monto de las jubilaciones y pensiones recibidas por los miembros del hogar que presentan esa condición.

## C. Metodología

La identificación de la falta de respuesta se logró al verificar que los asalariados, los patrones y los trabajadores por cuenta propia que se declararon ocupados, tenían el valor cero en la variable sueldos y salarios y ganancias, respectivamente. Esta misma situación se presentó en el caso de los miembros del hogar que clasificaron como jubilados o pensionados. El dígito cero fue sustituido por el carácter “.”, que los paquetes estadísticos utilizados le asignan a los datos omitidos.

La primera etapa del trabajo consistió en cuantificar la falta de respuesta, y observar su distribución en la muestra. Posteriormente, se efectuaron pruebas estadísticas para validar si los datos faltantes mostraban algún patrón de comportamiento, con el propósito de determinar si la distribución de la no respuesta se podía asociar con un comportamiento aleatorio.

Para la sustitución de los datos faltantes se aplicaron los siguientes ocho procedimientos de imputación simple y múltiple: *listwise*, medias condicionadas, *hot-deck*, *hot-deck* con regresión, regresión condicionada, máxima verosimilitud, imputación simple y dos algoritmos de imputación múltiple.<sup>42</sup>

---

<sup>42</sup> Con excepción del algoritmo EM —que se aplicó con el paquete SPSS—, el resto de las técnicas se implementaron con STATA. El primer método de imputación múltiple se aplicó utilizando el comando MUVIS, en tanto que el segundo corresponde al comando ice desarrollado por Royston (2005), el cual hace uso del comando UVIS de STATA, que realiza imputación simple y que también fue utilizado en este trabajo.

En virtud de que la mayoría de los métodos aplicados se sustentan en técnicas de regresión, como fase previa a la imputación de los datos fue necesario especificar modelos que relacionan las variables de interés (sueldos y salarios, ganancias y jubilaciones y pensiones) con las características de las personas. Toda vez que se corroboró la significancia estadística de los parámetros generados por las especificaciones ajustadas, se introdujeron como información auxiliar en los procedimientos que lo requieren y se procedió a imputar los datos faltantes. Utilizando la matriz de datos completos se generaron estimadores para el ingreso total y promedio —y su error estándar— para las tres variables estudiadas. Posteriormente, se reconstruyó el ingreso *per capita* del hogar utilizando los datos generados por los ocho procedimientos de imputación que se comparan.

Con la información anterior se estimó el ingreso *per capita* del hogar y se calcularon las tasas de indigencia y pobreza, así como diversos indicadores que frecuentemente se utilizan para evaluar la concentración del ingreso (índices de Gini, Atkinson ( $\epsilon = 2$ ) y Theil), con el propósito de conocer la sensibilidad de los resultados a los procedimientos de imputación utilizados.

La aplicación de técnicas que utilizan métodos de regresión requiere la especificación de una forma funcional que relacione la variable que se desea imputar, con un vector de covariables que expliquen su trayectoria. En ese sentido, debido a que se comprobó que los ingresos reportados en las variables de interés se correlacionaron positivamente con la escolaridad, la experiencia y el sexo de las personas, estas variables se utilizaron en las especificaciones propuestas por Mincer —ecuaciones de capital humano— para modelar el comportamiento de los sueldos, salarios, las ganancias netas y el monto proveniente de las jubilaciones y pensiones.<sup>43</sup> Con el propósito de reducir los sesgos de estimación se formaron subgrupos de observaciones utilizando la condición de ocupación, la categoría ocupacional, la rama de actividad agrupada en cuatro estratos y el sexo de las personas. En el caso de los jubilados y pensionados, además de que se verificó que su condición de actividad coincidiera con esa categoría, la imputación se llevó a cabo teniendo en cuenta el sexo de las personas.

## D. Resultados

### 1. Tasa de no respuesta

En el cuadro 1 se observa que el 14,8% de los asalariados (4.048 personas) no reportaron ingresos por sueldos y salarios (*sysorimi*), y se advierte que la falta de respuesta fue mayor en el caso de los hombres (17,2%), (ver el cuadro 2).<sup>44</sup> El porcentaje de omisión más importante se reportó en las ganancias provenientes del trabajo independiente (*ganorimi*). En esta variable, el 27,6% de los trabajadores por cuenta propia y los patrones no informaron sus ingresos (véase cuadro 1), y la concentración de datos faltantes no muestra diferencias por género (véase cuadro 2).

No obstante, el hecho de que uno de cada cuatro ocupados —patrones y cuenta propia— no haya reportado ingresos, significa que en la matriz de datos completos alrededor del 28% de las observaciones contendrán información generada por un modelo estadístico y no de datos observados.<sup>45</sup> Esta situación debe considerarse en el análisis y es recomendable utilizar información exógena para verificar que la imputación no haya producido sesgos en los datos.<sup>46</sup>

<sup>43</sup> Las ecuaciones utilizadas en los distintos procedimientos de imputación que utilizan modelos de regresión fueron:

$$l_{sys} = a_0 + a_1 * exp + a_2 * exp^2 + a_3 * niveduc + a_4 * sexo + u$$

$$l_{gan} = a_0 + a_1 * exp + a_2 * exp^2 + a_3 * niveduc + a_4 * sexo + u$$

$$l_{yjub} = a_0 + a_1 * sexo + a_2 * edad + u.$$

Como es convencional en este tipo de estudio, la experiencia se aproximó restándole a la edad los años de estudio de las personas y el valor 6, que corresponde a la edad de ingreso al ciclo de educación básica.

<sup>44</sup> Los promedios que se presentan para las distintas fuentes de ingreso corresponde a resultados sin ponderar (valores muestrales).

<sup>45</sup> Este tipo de situaciones son las que ponen en duda la validez de un método de imputación. En este caso, la falta de respuesta se considera muy alta, lo cual refiere a la posibilidad de prescindir de esta variable. Este tipo de situaciones, necesariamente debiera

**CUADRO 1**  
**RESULTADO DE LOS MODELOS AJUSTADOS**

	N	Media	Desviación estándar	Missing		N° extremos	
				Count	%	Bajo	Alto
<b>Sueldos y salarios</b>							
<i>Sysorimi</i>	23 301	628,0	648,6	4 048	14,8	0	1 239
<i>Sexo/Conduct/Categ</i>	27 349			0	0		
<b>Ganancias netas</b>							
<i>ganorimi</i>	6 466	607,4	797,1	2 464	27,6	0	442
<i>Sexo/Ramar1</i>	8 930			0	0		
<b>Jubilaciones y pensiones</b>							
<i>Juborimi</i>	6 984	483,9	425,4	599	7,9	0	506
<i>Sexo/Conduct</i>	7 583			0	0		

Fuente: Elaboración propia de los autores.

Nota: Número de casos fuera de rango ( $Q1 - 1,5 \cdot IQR$ ,  $Q3 + 1,5 \cdot IQR$ ).

Para analizar la situación descrita es posible hacer uso de información de encuestas anteriores, en donde la falta de información en las ganancias no haya sido tan severa, y analizar, además de las medidas de tendencia central habituales, la dispersión de los datos y la forma de la distribución a partir de coeficientes de asimetría y *kurtosis* y de indicadores de desigualdad.

En el caso de los pensionados se constata que sólo el 8% de los entrevistados (*juborimi*) no reportó el monto de sus ingresos, y de acuerdo al cuadro 2 no existe ningún patrón de comportamiento que dé cuenta que la concentración de las omisiones se encuentra condicionada por el sexo de las personas.

**CUADRO 2**  
**DISTRIBUCIÓN DE DATOS FALTANTES POR SEXO**

	Total	Hombre	Mujer
<b>Sueldos y salarios<sup>a</sup></b>			
Número de observaciones	23 301	12 329	10 972
Porcentaje	85,2	82,8	88,0
% Missing	14,8	17,2	12,0
<b>Ganancias<sup>b</sup></b>			
Número de observaciones	6 466	2 083	4 383
Porcentaje	72,4	72,3	72,5
% Missing	27,6	27,7	27,5
<b>Jubilaciones y pensiones<sup>a</sup></b>			
Número de observaciones	6 984	2 856	4 128
Porcentaje	92,1	92,2	92,7
% Missing	7,9	8,8	7,3

Fuente: Elaboración propia de los autores.

**Notas:** No han sido desplegadas:

<sup>a</sup> Indicador de variables con menos de 1% missing.

<sup>b</sup> Indicador de variables con menos de 5% missing.

llamar la atención de los administradores de la encuesta, ya que sugiere la necesidad de aplicar mecanismos de supervisión más rigurosos, que aseguren el control de las tasas de no respuesta por variable.

<sup>46</sup> Para este propósito, se recomienda utilizar información de otras rondas de la encuesta, para corroborar la robustez estadística de las imputaciones efectuadas en los montos de las ganancias netas reportadas por los patrones y trabajadores por cuenta propia.

Como fase previa a la aplicación de los procedimientos de imputación se verificó que la no respuesta no estuviera asociada con algún patrón atípico (MNAR), con el propósito de estar en condiciones de afirmar que los datos que hacen falta se generaron por un proceso aleatorio (MAR) o completamente aleatorio (MCAR), situación que es deseable para la aplicación de los algoritmos de imputación que se comparan. Además del análisis gráfico del patrón de no respuesta, y con el propósito de disponer de más evidencia para sustentar esta hipótesis, se aplicó la prueba de Little (1988), la cual entrega parámetros estadísticos que permiten determinar si es apropiado asumir que los datos faltantes siguen un patrón MAR o MCAR.

Como hipótesis nula ( $H_0$ ) se establece que el patrón de datos sin respuesta es MCAR, y conforme a lo dispuesto por Little<sup>47</sup> se calcula el valor del estadístico de prueba  $d^2$  —que tiene una distribución  $X^2$  con  $f$  grados de libertad—, y la regla de decisión establece que para valores grandes de  $d^2$  se rechaza  $H_0$ . Los valores de  $d^2$  calculados fueron casi nulos para las tres variables de interés, lo que permite afirmar que las observaciones faltantes siguen una distribución aleatoria en la muestra que se analiza.<sup>48</sup>

Por otra parte, y debido a que se ajustaron modelos que utilizan la escolaridad de los miembros del hogar como covariable, fue necesario analizar el patrón de datos omitidos en esa variable. Se constata que 638 personas (0,7%) no reportaron años de escolaridad, y el resultado de la prueba de hipótesis indica que la falta de respuesta se generó mediante un patrón MCAR. La corrección de la omisión se efectuó a partir de la aplicación del algoritmo de máxima verosimilitud (EM) que se incluye en el paquete SPSS.

Antes de modificar los valores en las variables originales es importante conocer la forma que asume la distribución de frecuencias, por lo que en los gráficos 1 a 3 se presentan los histogramas de las variables con *missing values*, y en los gráficos 1 a 3 del anexo 1 se grafican las distribuciones generadas con las variables imputadas.<sup>49</sup>

## 2. Distribuciones de frecuencias

Los histogramas de las variables originales difieren de la forma que asume la distribución cuando el valor “0” es sustituido por el carácter “.” (*missing values*). A pesar de que a simple vista no es posible percibir las diferencias, un análisis cuidadoso comprueba que el valor de la media no coincide, además de que también se manifiestan cambios en la dispersión de la variable (desviación estándar) y en la forma de la distribución (asimetría y *kurtosis*), (véanse los gráficos 1, 2 y 3).

## 3. Modelos ajustados

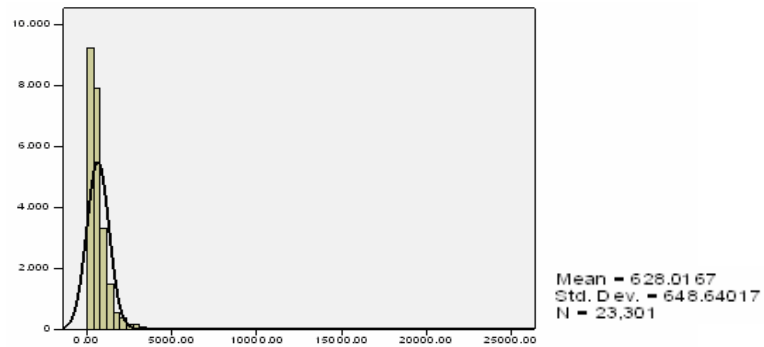
Como requisito previo para la aplicación de los algoritmos de imputación que utilizan modelos de regresión, es necesario ajustar los modelos propuestos y verificar la significancia estadística de los parámetros asociados a las covariables incorporadas en las ecuaciones propuestas. El examen de los resultados que se muestran en el cuadro 3, advierte que en los tres modelos propuestos los parámetros resultaron estadísticamente significativos al 1%, y conforme a lo que se espera en este tipo de especificaciones econométricas, el coeficiente de determinación ( $R^2$ ) asumió valores adecuados.

<sup>47</sup> La prueba de Little está incorporada en el procedimiento MVA del SPSS en la opción EM.

<sup>48</sup> Previo a la imputación de los datos faltantes en la variable años de educación, se aplicó la prueba de Little. El valor del estadístico  $d^2$  fue casi 0, lo cual indica que es posible asumir que las omisiones siguen un patrón completamente aleatorio.

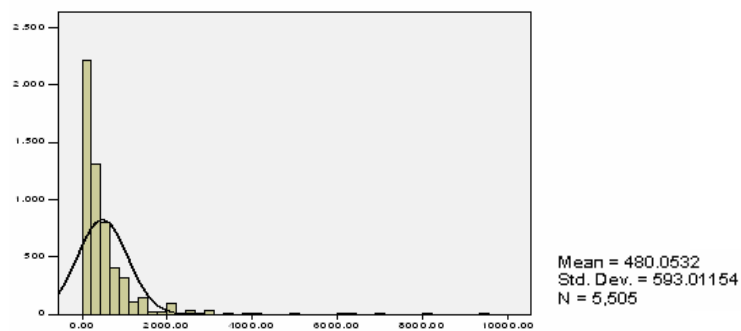
<sup>49</sup> El análisis cuidadoso de los gráficos revela las diferencias importantes que hay en la forma de las distintas distribuciones generadas. Por ejemplo, mientras que la forma del histograma para la variable sueldos y salarios con valores originales da cuenta de una distribución unimodal en la que predomina el valor cero, la que tiene valores *missing* corresponde a una distribución multimodal, en tanto que la variable imputada por el método de promedios muestra una distribución unimodal concentrada en torno a los valores imputados.

**GRÁFICO 1**  
**SUELDOS Y SALARIOS ORIGINALES CON MISSING**  
*(Frecuencia por miles de hogares)*



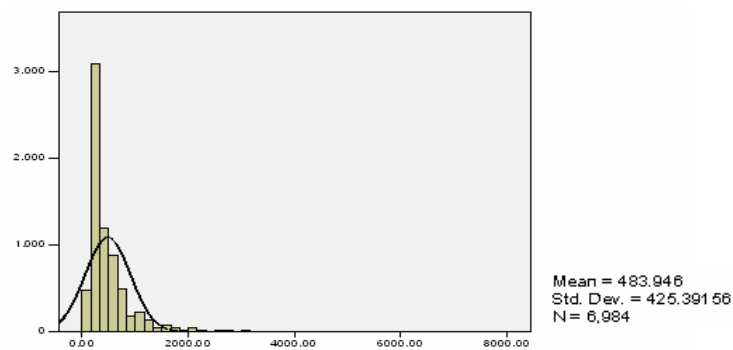
Fuente: Elaboración propia de los autores.

**GRÁFICO 2**  
**GANANCIAS ORIGINALES CON MISSING**  
*(Frecuencia por miles de hogares)*



Fuente: Elaboración propia de los autores.

**GRÁFICO 3**  
**JUBILACIONES Y PENSIONES ORIGINALES CON MISSING**  
*(Frecuencia por miles de hogares)*



Fuente: Elaboración propia de los autores.

**CUADRO 3**  
**RESULTADOS DE LOS MODELOS AJUSTADOS**

<b>Sueldos y salarios</b>						
Source	SS	df	MS		Number of obs = 23301	
Model	4373.39883	4	1093.34971		F( 4, 23296) = 1744.51	
Residual	14600.4938	23296	.62673823		Prob > F = 0.0000	
					R-squared = 0.2305	
					Adj R-squared = 0.2304	
Total	18973.8926	23300	.814330156		Root MSE = .79167	
lsysorimi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
anoseduc	.0877942	.0013713	64.02	0.000	.0851063	.0904821
sexo	.4735554	.0104083	45.50	0.000	.4531544	.4939564
exp	.0517455	.0013448	38.48	0.000	.0491096	.0543814
exp2	-.0007439	.0000256	-29.07	0.000	-.0007941	-.0006938
_cons	4.32035	.0233638	184.92	0.000	4.274555	4.366144
<b>Trabajadores por cuenta propia y patronos</b>						
Source	SS	df	MS		Number of obs = 6466	
Model	1817.2517	4	454.312924		F( 4, 6461) = 486.74	
Residual	6030.52181	6461	.933372823		Prob > F = 0.0000	
					R-squared = 0.2316	
					Adj R-squared = 0.2311	
Total	7847.77351	6465	1.21388608		Root MSE = .96611	
lganorimi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exp	.070817	.0029178	24.27	0.000	.0650972	.0765368
exp2	-.0010402	.0000487	-21.37	0.000	-.0011356	-.0009448
anoseduc	.1044805	.0031216	33.47	0.000	.0983611	.1105999
sexo	.5020693	.0257268	19.52	0.000	.4516363	.5525024
_cons	3.530786	.0575714	61.33	0.000	3.417927	3.643645
<b>Jubilados y pensionados</b>						
Source	SS	df	MS		Number of obs = 6984	
Model	524.615875	4	131.153969		F( 4, 6979) = 381.20	
Residual	2401.14042	6979	.344052217		Prob > F = 0.0000	
					R-squared = 0.1793	
					Adj R-squared = 0.1788	
Total	2925.7563	6983	.418982715		Root MSE = .58656	
ljuborimi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
anoseduc	.0487992	.0018868	25.86	0.000	.0451006	.0524979
sexo	.2473725	.0143167	17.28	0.000	.2193075	.2754376
exp	.0317077	.0022286	14.23	0.000	.0273339	.0360763
exp2	-.0003	.0000227	-13.25	0.000	-.0003444	-.0002556

Fuente: Elaboración propia de los autores.

#### 4. Ingreso promedio y *per capita* del hogar

Los ocho métodos de imputación propuestos se aplicaron para sustituir las omisiones detectadas en la variable sueldos y salarios, y los estimadores de totales, promedios y sus errores estándar se presentan en el cuadro 4.

Los promedios se ubicaron entre 546,3 (datos originales) y 662,7 para la opción que elimina las observaciones con *missing values* (*listwise*). El promedio estimado con los datos originales es el más bajo de todos los valores comparados, debido a que los registros con valor “0” en la variable de análisis se involucran en el cálculo, lo cual explica que se obtenga un menor registro. Por su parte, el procedimiento *listwise* elimina las observaciones que tienen el carácter “.”, por lo que el ingreso medio se calcula sólo para la submuestra con información completa, lo que explica el incremento que se percibe en el valor del estimador.

En la figura 10 se muestra el ordenamiento de los estimadores, lo que permite observar el rango de variación entre los promedios imputados. En la parte izquierda del paréntesis inferior, se reporta el porcentaje de observaciones que el método de imputación les asignó un ingreso inferior a dos veces el error estándar del promedio, en tanto que el lado derecho corresponde al porcentaje de registros con valores imputados superiores a dos veces el error estadístico.

**CUADRO 4**  
**SUELDOS Y SALARIOS IMPUTADOS POR DISTINTOS PROCEDIMIENTOS**

Método de imputación	Total	Error estándar	Promedio	Error estándar
Original (ORIG)	3 811 866 578	5,62E+07	546,3	4,49
Listwise <sup>a</sup>	3 811 866 578	5,81E+07	662,7	4,58
Medias condicionadas <sup>b</sup>	4 608 836 053	5,55E+07	660,5	3,44
Hot-deck (HDD) <sup>c</sup>	4 584 487 415	5,01E+07	657,4	3,72
Hot-deck con regresión (HDR) <sup>d</sup>	4 446 271 182	5,01E+07	637,2	3,92
Regresión condicionada (R) <sup>e</sup>	4 426 107 789	5,29E+07	634,3	3,83
Máxima verosimilitud (MV) <sup>f</sup>	4 598 244 445	5,37E+07	659,4	3,23
Imputación simple (IS) <sup>g</sup>	4 622 756 868	6,10E+07	662,5	4,07
Imputación múltiple (IM) <sup>h</sup>	4 445 294 105	5,51E+07	637,1	3,43
Imputación múltiple (ICE) <sup>i</sup>	4 404 730 384	5,61E+07	631,3	3,75

Fuente: Elaboración propia de los autores.

Nota: El error estándar de totales y promedios se calculó a partir de la aplicación del procedimiento *bootstrap* con 100 réplicas.

<sup>a</sup> Eliminación de observaciones incompletas (*Listwise*).

<sup>b</sup> Imputación por medias condicionadas.

<sup>c</sup> Imputación por *hot-deck* con distancias: Promedio de 5 imputaciones con STATA.

<sup>d</sup> Imputación por *hot-deck* con regresión condicionada (STATA).

<sup>e</sup> Imputación por regresión condicionada: comando *impute* de STATA.

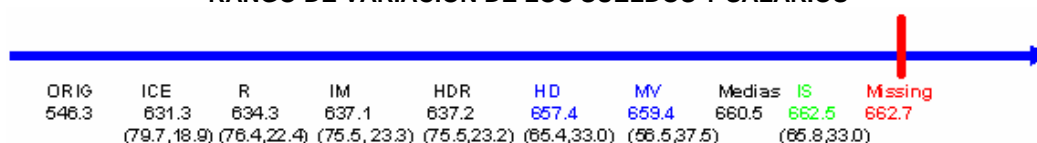
<sup>f</sup> Máxima verosimilitud: Algoritmo EM del paquete SPSS.

<sup>g</sup> Imputación simple: Comando UVIS de STATA.

<sup>h</sup> Imputación múltiple: Comando MUVIS de STATA.

<sup>i</sup> Imputación múltiple: Comando ice de STATA.

**FIGURA 10**  
**RANGO DE VARIACIÓN DE LOS SUELDOS Y SALARIOS**



Fuente: Elaboración propia de los autores.

Lo anterior significa, por ejemplo, que como resultado de la aplicación del método de máxima verosimilitud (MV), el 56,5% de los datos omitidos se sustituyó con un guarismo que resultó 6,46 veces inferior al promedio estimado (659,4).<sup>50</sup>

Si se ubica como punto de referencia el valor estimado por el método de medias (660,5), se observa que, con excepción del algoritmo de imputación simple (IS) (662,5), en el resto de los procedimientos el promedio estimado de sueldos y salarios fue menor al valor de referencia.

Debido a la importancia de los sueldos y salarios en la formación del ingreso *per capita* —y de la ubicación de las familias en distintos niveles de pobreza—, cabe esperar que la tasa de pobreza estimada por cualquiera de los métodos comparados, sea mayor de la que se obtendría utilizando el ingreso *per capita* estimado a partir del método de medias.

Esta aseveración se refuerza al observar que con excepción de los métodos HDD y MV, el porcentaje de observaciones imputadas con un valor inferior a dos veces el error estándar superó el

<sup>50</sup> En el cuadro 2 del anexo 1 se presenta los datos que avalan esta afirmación.



75%, y en el caso del método ICE llegó a casi al 80% (véase el cuadro 2 del anexo1). La teoría sugiere que el método de promedios introduce sesgos en el valor del estimador y en su varianza, lo que se constata a partir de los resultados obtenidos. En lo que corresponde a los procedimientos de MV y de imputación simple y múltiple, los valores se sustituyen en forma aleatoria, por lo que cabría esperar que no se genere sesgos en la asignación de los valores imputados.<sup>51</sup>

La ubicación de la media y el error estándar inciden en la forma de distribución (asimetría y *kurtosis*), lo que se constata a partir de los resultados que se muestran en el cuadro 1 del anexo 1. Valores grandes en el coeficiente de *kurtosis* advierten de la existencia de un número importante de observaciones concentradas alrededor de la media, lo que es característico de distribuciones con “colas pesadas” —como la que se observa para la variable ingreso— que presenta baja concentración de valores extremos. En este contexto, los coeficientes generados por el método de promedios y de máxima verosimilitud muestran los valores más grandes (238,01 y 240,53).

Los estimadores de totales y promedios obtenidos para las ganancias imputadas aparecen en el cuadro 5. Los valores generados por el método de medias y el de máxima verosimilitud son prácticamente iguales, y se distinguen por presentar los promedios más altos (543,7 y 543,7 respectivamente) entre las observaciones que se comparan.

**CUADRO 5**  
**GANANCIAS IMPUTADAS POR DISTINTOS PROCEDIMIENTOS**

Método de imputación	Total	Error estándar	Promedio	Error estándar
Original (ORIG)	781 801 441	1,96E+07	386,1	6,41
Listwise	741 801 441	2,56E+07	543,9	8,64
Medias condicionadas	1 044 701 246	1,87E+07	686,1	6,01
Hot-deck (HDD)	1 011 016 798	2,02E+07	656,7	5,43
Hot-deck con regresión (HDR)	939 467 765	2,02E+07	527,5	6,13
Regresión condicionada (R)	934 000 056	2,10E+07	593,0	6,06
Máxima verosimilitud (MV)	1 044 026 595	1,95E+07	664,5	6,62
Imputación simple (IS)	1 039 917 889	1,76E+07	669,8	8,34
Imputación múltiple (IM)	925 017 263	2,00E+07	599,6	6,62
Imputación múltiple (ICE)	917 782 612	1,82E+07	581,6	6,45

Fuente: Elaboración propia de los autores.

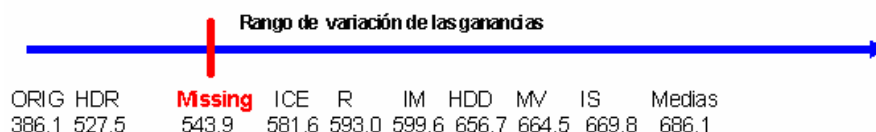
En este caso, el sesgo que introduce el método de medias en la varianza del estimador se percibe al comparar los valores del error estándar generado por los diversos procedimientos de imputación. En efecto, el método de medias produce el menor error estándar (6,01), siguiendo en orden de magnitud el error que se genera a partir de la aplicación del método de regresión.

Por otra parte, es importante hacer notar que el ordenamiento generado para la variable ganancias difiere del observado en los sueldos y salarios. Esto constata que la eficacia de los procedimientos depende de la variable de análisis, de la tasa de no respuesta y de su distribución en la muestra, y permite afirmar que si una técnica de imputación resultó adecuada para una variable, no significa que su uso se debe generalizar sin analizar las condiciones en que se generó la falta de respuesta en otras variables de interés.

<sup>51</sup> Se podría argumentar que los tres métodos generan un error estándar similar, y por tanto cualquiera de ellos se puede utilizar. Sin embargo, las propiedades teóricas de los procedimientos han sido analizadas, y se demuestra que el método de promedios genera estimadores sesgados.

Si se decidiera, por ejemplo, imputar las ganancias con el método ICE porque en la variable sueldos y salarios el valor estimado por esta técnica se ubicó a la derecha del punto de referencia (medias), la figura 11 demuestra que esta situación ha cambiado, ya que el procedimiento ICE estima uno de los promedios de ganancias más bajos entre los datos comparados.

**FIGURA 11**  
**RANGO DE VARIACIÓN DE LAS JUBILACIONES Y PENSIONES**



Fuente: Elaboración propia de los autores.

En el cuadro 6 se presentan los resultados que se obtuvieron en la variable jubilaciones y pensiones. Se corrobora que el método de imputación simple generó el promedio más elevado (485,3), en tanto que el estimador más bajo lo asignó el procedimiento de imputación múltiple (ICE).

**CUADRO 6**  
**JUBILACIONES Y PENSIONES IMPUTADAS POR DISTINTOS PROCEDIMIENTOS**

Método de imputación	Total	Error estándar	Promedio	Error estándar
Original (ORIG)	812 190 300	2,62E+07	431,1	5,23
Listwise (1)	812 190 300	1,88E+07	475,8	5,17
Medias condicionadas	897 583 451	2,50E+07	476,5	4,50
Hot-deck (HDD)	898 423 065	2,78E+07	476,9	4,60
Hot-deck con regresión (HDR)	896 823 872	2,56E+07	473,8	4,44
Regresión condicionada (R)	892 463 834	2,48E+07	473,7	4,65
Máxima verosimilitud (MV)	896 211 122	2,80E+07	475,7	4,80
Imputación simple (IS)	914 291 691	2,29E+07	485,3	4,72
Imputación múltiple (IM)	892 436 881	2,72E+07	473,7	4,29
Imputación múltiple (ICE)	885 575 398	2,99E+07	470,1	4,50

Fuente: Elaboración propia de los autores.

La figura 12 nuevamente pone en evidencia que el ordenamiento que se genera depende de la variable de análisis y de la tasa de no respuesta. En este caso, el límite superior del rango de variación lo ocupó el valor estimado por el método (IS) —en la figura 10 fue el método *missing* y en la 11 el procedimiento de medias—, y además se redujo la distancia entre los valores extremos (54,2 pesos).<sup>52</sup> Asimismo, los errores más pequeños se aprecian en los valores generados por los procedimientos IS, Medias e ICE.

La imputación de los sueldos y salarios, las ganancias y jubilaciones y pensiones, es una fase intermedia para aproximarse al ingreso *per capita* de las familias, ya que este es el indicador se contrasta con los valores de las líneas de indigencia y pobreza.

<sup>52</sup> En la variable sueldos y salarios el rango de variación fue 116,4 pesos, en tanto que en las ganancias la diferencia entre los extremos fue de 157,8 unidades.

**FIGURA 12**  
**RANGO DE VARIACIÓN DE LAS JUBILACIONES Y PENSIONES**



Fuente: Elaboración propia de los autores.

Conforme a los resultados que se muestran en el cuadro 7, se percibe todos los algoritmos comparados estimaron un ingreso *per capita* inferior al obtenido por el método de medias (valor de referencia). Por su parte, los procedimientos *hot-deck* y de máxima verosimilitud produjeron los promedios más altos con 538,9 y 541,5 pesos respectivamente.

**CUADRO 7**  
**INGRESO PER CAPITA Y SU ERROR ESTÁNDAR**

Método de imputación	Ingreso <i>per capita</i>	Error estándar
Medias condicionadas	726,8	3,20
<i>Hot-deck</i> (HDD)	538,9	3,59
<i>Hot-deck</i> con regresión (HDR)	509,8	3,67
Regresión condicionada (R)	515,6	3,20
Máxima verosimilitud (MV)	541,5	3,68
Imputación simple (IS)	530,6	3,52
Imputación múltiple (IM)	508,7	3,20
Imputación múltiple (ICE)	505,5	3,47

Fuente: Elaboración propia de los autores.

A partir de las cifras del cuadro 7, cabría esperar que todos los métodos que generan promedio de ingreso *per capita* por debajo de 587,3 (valor de referencia), produzcan tasas de pobreza superiores, y ocurra lo contrario cuando el estimador admita un valor superior.

Como fue señalado, el promedio de sueldos y salarios estimado por el método de medias fue superior al de máxima verosimilitud y al *hot-deck*. No obstante, cuando se analiza el ingreso *per capita* la relación se invierte, debido a que en ambos casos un porcentaje elevado de los datos imputados (43% y 34,4%), se les asignó un valor superior al que le correspondió por el método de medias.

En lo que se refiere a la dispersión del ingreso *per capita* se observa que los errores estándar se ubicaron entre 3,20 y 3,68 unidades, correspondiendo el valor menor a los métodos de medias condicionadas, regresión e imputación múltiple, en tanto que el error de mayor magnitud se observó en el procedimiento de imputación simple.

El sesgo que introduce el método de medias se refleja en el estimador de la varianza del ingreso *per capita*. Se puede argumentar que, desde el punto de vista del error estadístico, el estimador de medias es similar al que se genera por regresión e imputación simple; sin embargo, en los dos últimos casos el algoritmo asigna en forma estocástica los valores imputados, mientras que el procedimiento de medias reduce artificialmente la varianza del estimador.

En el cuadro 8 se presentan indicadores de indigencia, pobreza y desigualdad, toda vez que los valores faltantes fueron reemplazados con distintos procedimientos de imputación. Los datos demuestran que el método de imputación aplicado afecta la incidencia de la pobreza y el nivel de la concentración del ingreso.

**CUADRO 8**  
**ESTIMACIONES DE POBREZA Y DESIGUALDAD CON DISTINTOS MÉTODOS DE IMPUTACIÓN**

Método de imputación	% de Personas en		Índices de		
	Indigencia	Pobreza	Gini	Atkinson ( $\epsilon=2$ )	Theil
Medias condicionadas	11,1	29,4	0,531	0,644	0,634
<i>Hot-deck</i> (HDD)	10,6	28,3	0,527	0,644	0,615
<i>Hot-deck</i> con regresión (HDR)	12,0	32,7	0,536	0,637	0,661
Regresión condicionada (R)	12,1	32,7	0,543	0,646	0,666
Máxima verosimilitud (MV)	10,3	27,5	0,522	0,641	0,607
Imputación simple (IS)	12,8	32,8	0,551	0,663	0,673
Imputación múltiple (IM)	12,1	32,9	0,536	0,636	0,663
Imputación múltiple (ICE)	12,0	33,0	0,536	0,634	0,666

Fuente: Elaboración propia de los autores.

En el cuadro 9 se muestra el ordenamiento de los métodos conforme al error estándar estimado en cada una de las variables de interés. El primer lugar se asigna al procedimiento que generó el error más bajo y la posición ocho se reserva para el caso contrario. Se observa que el método de medias se ubica en la segunda y tercera posición en las tres variables imputadas, generando el error estándar más bajo en el ingreso *per capita*. Por su parte, el procedimiento ICE se mantuvo entre la cuarta y quinta posición y en el ingreso *per capita* ocupó el cuarto lugar.

**CUADRO 9**  
**ORDENAMIENTO DE LOS MÉTODOS DE IMPUTACIÓN**  
**SEGÚN EL ERROR ESTÁNDAR ESTIMADO EN LA VARIABLE IMPUTADA**

Método	Sueldos y salarios	Ganancias	Jubilaciones y pensiones	<i>Per capita</i>
Medias	3	2	3	1
HDD	4	1	5	6
HDR	6	4	2	7
R	5	3	6	2
MV	1	6	8	8
IS	7	8	7	5
IM	2	7	1	3
ICE	4	5	4	4

Fuente: Elaboración propia de los autores.

## 5. Incidencia de la pobreza

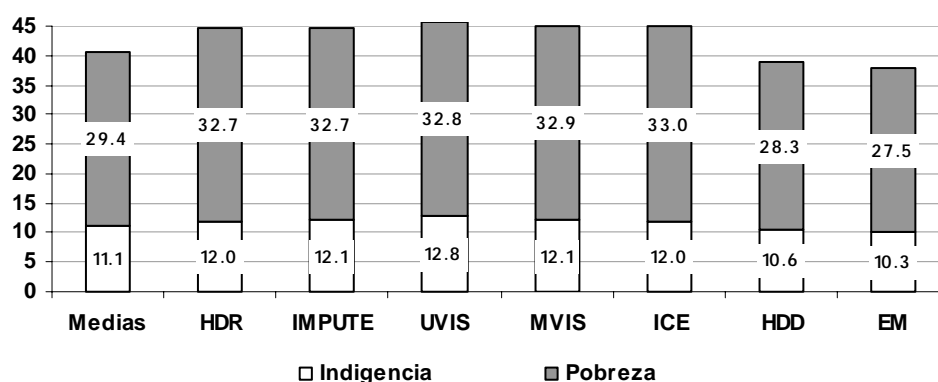
La importancia de elegir el método de imputación que posea las mejores propiedades estadísticas y entregue los resultados más robustos, se combina con el interés de disponer de un buen estimador del ingreso *per capita* que sea de utilidad para cuantificar la magnitud, profundidad y severidad de la pobreza. En este sentido, la sensibilidad de los indicadores de pobreza a los distintos métodos de imputación se presenta en el cuadro 10 y el gráfico 4.

**CUADRO 10**  
**ERRORES DE MUESTREO DE LAS TASAS DE INDIGENCIA Y POBREZA**

Método	Indigencia	Bootstrap al 99% (100) repeticiones	Error estándar	z	Pobreza	Bootstrap al 99% (100) repeticiones	Error estándar	z
Medias	0,111	0,108 - 0,114	0,00156		0,294	0,291 - 0,298	0,00188	
HDR	0,120	0,117 - 0,123	0,00143	4,3	0,327	0,322 - 0,332	0,00245	10,5
IMPUTE	0,121	0,118 - 0,124	0,00147	4,5	0,327	0,323 - 0,332	0,00213	11,6
UVIS	0,128	0,125 - 0,131	0,00155	7,6	0,328	0,324 - 0,332	0,00199	12,3
MVIS	0,121	0,118 - 0,123	0,00145	4,5	0,329	0,324 - 0,333	0,00226	11,6
ICE	0,120	0,117 - 0,123	0,00153	4,0	0,330	0,326 - 0,334	0,00207	12,7
HDD	0,106	0,103 - 0,109	0,00147	2,5	0,283	0,279 - 0,287	0,00202	4,1
EM	0,103	0,100 - 0,105	0,00143	4,0	0,275	0,270 - 0,279	0,00214	6,7

Fuente: Elaboración propia de los autores.

**GRÁFICO 4**  
**INCIDENCIA DE LA POBREZA POR DISTINTOS MÉTODOS DE IMPUTACIÓN**



Fuente: Elaboración propia de los autores.

Los porcentajes de indigencia y pobreza más bajos se obtuvieron con el algoritmo de máxima verosimilitud (10,3%) y *hot-deck* (10,6%), mientras que el resto de los casos las tasas estimadas fueron superiores al valor generado por el método de medias (11,1%). Por su parte, el algoritmo IS y el *hot-deck* con regresión estimaron 12% de población en pobreza extrema, mientras que ICE e IM generaron porcentajes de 12,1 y 12,8%, respectivamente. La diferencia entre el porcentaje estimado por MV y el método de medias fue de menos de un punto porcentual (0,8), en tanto que a partir del procedimiento ICE se obtuvo una tasa de indigencia 1,7 puntos superior al valor de referencia.

En los estudios que dan cuenta de la magnitud de la pobreza no es usual que la tasa de indigencia y pobreza se acompañe por su error de muestreo, ni se indiquen los límites del intervalo de confianza en que se encuentra el valor del estimador. Este hecho inhibe la posibilidad de afirmar si las diferencias observadas entre dos o más años son estadísticamente significativas.

En el caso que nos ocupa, y con el propósito de corroborar si las diferencias observadas en las tasas de indigencia y pobreza estimadas por los distintos algoritmos son significativas desde el punto de vista estadístico, se efectuaron pruebas de hipótesis fijando como parámetro de referencia el porcentaje de indigentes estimado a partir de medias condicionadas.

Conforme a los resultados del cuadro 10 se demuestra que con excepción del porcentaje estimado con el algoritmo HDD, todos los estimadores de pobreza son estadísticamente distintos al valor de referencia (11,1%) al 1% de significancia, lo que permite concluir que:

“el método de imputación utilizado para corregir las omisiones en distintas variables de ingreso, si tiene repercusiones en el porcentaje y volumen de población indigente y pobre”.

En atención a los resultados presentados, es posible afirmar que los diseñadores de política deben tener precauciones acerca del procedimiento de imputación que aplicarán para subsanar la falta de respuesta, ya que ha quedado en evidencia que el volumen de familias y personas en situación de pobreza es sensible al método elegido.

Si los resultados se desean utilizar para determinar el total de familias que participarán en un programa social, y este depende del monto de los recursos disponibles, es claro que el monto de familias seleccionadas no es independiente del método de imputación aplicado.

Debido a la importancia que tienen los sueldos y salarios en la estructura del ingreso familiar es necesario señalar, en el caso de que la tasa de omisión sea importante, que se debe actuar con cautela en la elección del algoritmo de imputación ya que se confirma que esta variable es la que mayor influencia ejerce en la determinación de la tasa de pobreza.

Para corroborar la hipótesis planteada, en el cuadro 11 se presentan estimaciones de indigencia y pobreza aislando el efecto de cada variable imputada en el porcentaje de personas por categoría de pobreza, utilizando el método de medias como parámetro de referencia. Es decir, se sustituyeron los datos omitidos en la variable sueldos y salarios, manteniendo constantes los valores asignados por el método de promedios a las ganancias y jubilaciones y pensiones.

**CUADRO 11**  
**EFFECTO DE LA IMPUTACIÓN EN LA TASA DE POBREZA POR FUENTE DE INGRESO**

Método	Completa		Aislando salarios		Aislando ganancias		Aislando jubilaciones	
	Indigencia	Pobreza	Indigencia	Pobreza	Indigencia	Pobreza	Indigencia	Pobreza
MV	10,3	27,5	10,3	27,8	11,0	29,1	11,1	29,5
HDD	10,6	28,3	10,6	28,2	11,1	29,6	11,1	29,4
Medias	11,1	29,4						
ICE	12,0	33,0	11,9	31,8	11,2	30,6	11,1	29,5
HDR	12,0	32,7	11,9	31,5	11,3	30,5	11,1	29,5
IS	12,1	32,9	11,9	31,5	11,3	30,7	11,1	29,5
R	12,1	32,7	11,9	31,7	11,3	30,4	11,1	29,5
IM	12,8	32,8	12,4	31,5	11,5	30,6	11,1	29,5

Fuente: Elaboración propia de los autores.

Se comprueba que la tasa de indigencia estimada por el método UVIS (12,8%) se explica fundamentalmente por la imputación efectuada a los sueldos y salarios. Es decir, si se decidiera sustituir solamente esta variable, sin modificar las ganancias y las jubilaciones y pensiones, la indigencia se reduciría en sólo cuatro décimas (se ubicaría en 12,4%).

Una situación similar se observa en el resto de los métodos (con excepción de ICE), lo cual confirma que los sueldos y salarios es la variable de mayor relevancia para explicar la manera en que se clasifican los hogares en los distintos estratos de pobreza.<sup>53</sup>

<sup>53</sup> Esta afirmación se refuerza al observar que en esta variable el número absoluto de casos imputados fue mayor (4,048), en comparación con lo observado en ganancias y jubilaciones y pensiones.

## 6. Distribución el ingreso

En lo que se refiere a los indicadores de desigualdad, el cuadro 12 da cuenta que los valores más altos se obtuvieron a partir de la aplicación de los métodos de regresión e imputación simple. Por otra parte, con el procedimiento *hot-deck* y el de máxima verosimilitud se obtuvieron los niveles más bajos para el coeficiente de Gini, en tanto que el resto de los métodos produjeron estimadores que se ubicaron en el intervalo 0,5310 y 0,5359.

**CUADRO 12**  
**ERRORES DE MUESTREO DE LOS ÍNDICES DE DESIGUALDAD**

Método	Gini	Error estándar	z	Theil	Error estándar	z	Atkinson ( $\epsilon = 2$ )	Error estándar	z
MEDIAS	0,531	0,00445		0,633	0,02770		0,644	0,00465	
HDR	0,536	0,00452	0,74	0,661	0,02975	0,68	0,637	0,00537	1,00
IMPUTE	0,543	0,00496	1,72 <sup>a</sup>	0,666	0,02515	0,87	0,646	0,00431	0,29
UVIS	0,551	0,00490	3,08 <sup>b</sup>	0,673	0,02566	1,03	0,662	0,00480	2,79 <sup>b</sup>
MVIS	0,536	0,00466	0,76	0,663	0,02764	0,76	0,636	0,00500	1,13
ICE	0,536	0,00457	0,76	0,666	0,03111	0,78	0,634	0,00435	1,52
HDD	0,527	0,00429	0,68	0,615	0,02649	0,48	0,644	0,00480	0,08
EM	0,522	0,00450	1,36	0,607	0,02786	0,67	0,641	0,00422	0,45

Fuente: Elaboración propia de los autores.

<sup>a</sup> Estadísticamente significativo al 10%.

<sup>b</sup> Estadísticamente significativo al 1%

Es particularmente importante conocer la sensibilidad de los métodos de imputación a los cambios que se generan en la forma de la distribución del ingreso, y en especial en lo que corresponde a la participación de las familias ubicadas en la parte baja. Atendiendo a esta situación, se calculó el coeficiente de Atkinson para  $\epsilon=2$  y los resultados demuestran que no existen diferencias importantes entre los valores estimados por los distintos algoritmos que se contrastan.

Para corroborar si las diferencias observadas entre métodos son estadísticamente significativas, se aplicaron pruebas de hipótesis utilizando como referencia los valores estimados por el procedimiento de medias para los distintos indicadores de desigualdad que se analizan.

Los resultados del cuadro 12 demuestran que, con excepción del índice de Gini, estimado por el procedimiento de imputación simple, en ninguno de los casos las diferencias observadas en los coeficientes de Gini, Theil y Atkinson ( $\epsilon = 2$ ) son estadísticamente significativas, respecto a los valores generados con el método de medias condicionadas.

El cuadro 13 y el gráfico 5 dan cuenta de las diferencias entre el porcentaje de ingreso que concentran los distintos grupos de familias, y queda en evidencia que el método que generó los mayores cambios fue el algoritmo de máxima verosimilitud.

En efecto, mientras que con UVIS los deciles “1” y “10” retuvieron 0,57 y 50,6%, respectivamente, MV asigna porcentajes de 1,1% y 41,1% para los hogares ubicados en los extremos de la distribución del ingreso, lo cual explica el que esta alternativa de imputación genera los valores más bajos en los tres coeficientes de desigualdad que se comparan.

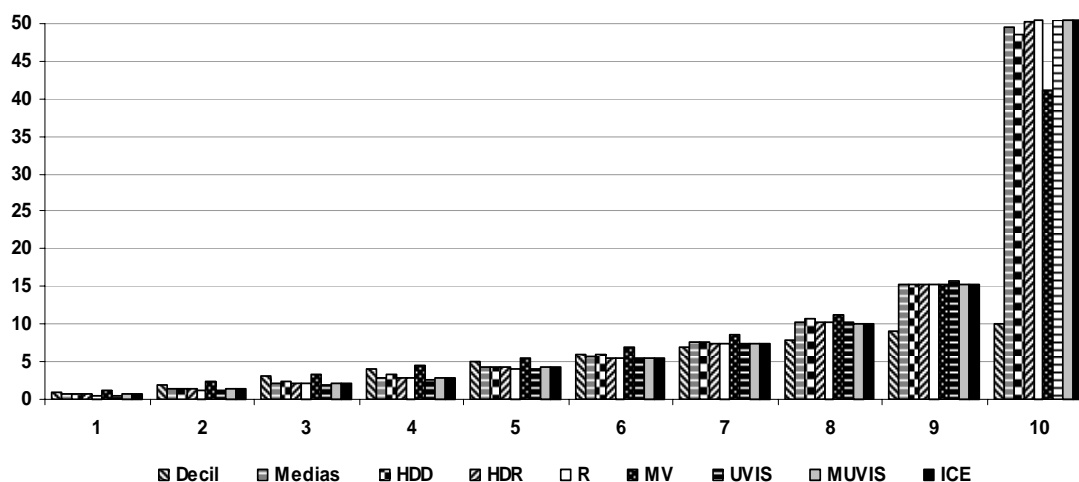
Se observa que el primer decil retiene entre 0,57% (imputación simple) y el 1,12% (MV) del ingreso total, lo cual genera coeficientes de Gini de 0,551 y 0,522, respectivamente, y la prueba de hipótesis demuestra que la diferencia entre esos valores es estadísticamente significativa al 1%.

**CUADRO 13**  
**DISTRIBUCIÓN DEL INGRESO POR DECIL/DISTINTOS MÉTODOS DE IMPUTACIÓN**

Decil	Medias	HDD	HDR	R	MV	UVIS	MUVIS	ICE
1	0,62	0,64	0,60	0,58	1,12	0,57	0,60	0,60
2	1,34	1,43	1,32	1,28	2,29	1,23	1,33	1,34
3	2,17	2,29	2,15	2,10	3,37	2,02	2,15	2,17
4	2,90	3,23	2,85	2,80	4,43	2,72	2,84	2,84
5	4,33	4,40	4,24	4,15	5,57	4,12	4,26	4,26
6	5,71	5,96	5,56	5,59	6,93	5,44	5,56	5,53
7	7,69	7,62	7,35	7,38	8,68	7,40	7,38	7,33
8	10,22	10,67	10,27	10,39	11,14	10,19	10,16	10,14
9	15,43	15,20	15,30	15,32	15,40	15,75	15,31	15,31
10	49,58	48,55	50,35	50,41	41,08	50,56	50,41	50,48
Total	100	100	100	100	100	100	100	100

Fuente: Elaboración propia de los autores.

**GRÁFICO 5**  
**ESTRUCTURA DEL INGRESO PER CAPITA POR DECILES**



Fuente: Elaboración propia de los autores.

## E. Discusión

1. Si la selección del método de imputación se sustenta únicamente en criterios estadísticos, es posible concluir que cualquiera de las alternativas analizadas genera distribuciones equivalentes, a pesar de las diferencias que se perciben en los parámetros de tendencia central, dispersión, asimetría y *kurtosis*.

El error estándar de los promedios de ingreso *per capita* se ubicó en un rango de variación bastante estrecho, y la matriz de correlación de Pearson da cuenta que los coeficientes de asociación asumen, en todos los casos, valores superiores al 95%. Asimismo, las pruebas de normalidad y bondad de ajuste dan cuenta que no existen diferencias estadísticamente significativas entre las 8 distribuciones comparadas.



2. Por otra parte, si el objetivo fuera completar los datos faltantes y la elección del método de imputación se justificara únicamente a partir los fundamentos teóricos de los procedimientos aplicados, la elección debería recaer en el método de máxima verosimilitud (MV) o en alguno de los métodos de los algoritmos de imputación múltiple.

La teoría avala que los estimadores generados por ambos procedimientos son robustos, y la sustitución de valores omitidos se realiza en forma estocástica, lo que garantiza que no se introducen sesgos de asignación. Las propiedades estadísticas de los estimadores se sustentan en técnicas bayesianas de probada utilidad, así como en procedimientos estocásticos de cadenas de Markov.

3. Si la preocupación es preservar la forma de la distribución de las variables imputadas, es factible decidirse por el algoritmo *hot-deck* en algunas de sus variantes. Este método de imputación es de larga data, y ha sido extensamente probado en censos y encuestas y por tanto es recomendado por los estadísticos dedicados a la teoría del muestreo.

4. Cuando el objetivo de la imputación se centra en generar estimaciones de pobreza y desigualdad, además de hacer uso de criterios estadísticos, la elección del método debiera sustentarse en la sensibilidad de los indicadores a los ajustes efectuados.

5. Debido a que es común que se aplique el método de promedios, los valores estimados por este procedimiento fueron utilizados como punto de comparación con el resto de las técnicas aplicadas. Lo primero que se constata, es que los índices de pobreza son sensibles al procedimiento que se elija para corregir la omisión en sueldos y salarios, ganancias netas y jubilaciones y pensiones.

6. El efecto del método imputación en los cambios en los ingresos depende de la variable de análisis, de la tasa de no respuesta registrada y de su distribución en la muestra, y se comprueba que el ordenamiento generado es distinto según sea la fuente de ingreso que se analice. El ordenamiento para la variable sueldos y salarios es distinto al observado en ganancias y jubilaciones y pensiones, lo cual demuestra que no posible generalizar la aplicación del método de imputación a cualquier variable.

7. Utilizando como referencia la tasa de pobreza estimada con el método de promedios (11,1%), las pruebas de hipótesis dan cuenta que los porcentajes de indigencia obtenidos por los distintos procedimientos son estadísticamente diferentes al 1%, con excepción al valor producido por el algoritmo *hot-deck* (10,6%). En caso de que la significancia estadística fuera 10%, todos valores comparados serían estadísticamente distintos a los obtenidos por el procedimiento de medias.

8. Desde el punto de vista práctico es muy relevante sustentar la elección de la técnica que se adoptará para la sustitución de datos. Si se optara por el método de promedios, la incidencia de la pobreza se subestimaría respecto a los valores generados por cinco de los métodos aplicados (HDR, R, IS, IM e ICE), en tanto que la elección del algoritmo de máxima verosimilitud o el *hot-deck*, resultaría los más apropiados, ya que reportan las tasas más bajas.

9. En lo que corresponde a la distribución del ingreso no se percibe claramente el efecto de los métodos en el valor de los coeficientes de desigualdad. Las pruebas de hipótesis indican que al 1% de significancia, los valores de los índices de Gini, Theil y Atkinson ( $\epsilon = 2$ ) son estadísticamente iguales. No obstante, esto no debe interpretarse en el sentido de que el algoritmo aplicado no produce diferencias en la forma de la curva de concentración del ingreso.

10. Los métodos que utilizan modelos de regresión generan valores que no alteran significativamente los índices de desigualdad. Sin embargo, cuando los hogares se ordenan de acuerdo a su ingreso se demuestra que el método de máxima verosimilitud redistribuye el ingreso hacia el decil más pobre, reduce la participación del 10% más rico y genera valores más bajos en los índices de Gini, Theil y Atkinson ( $\epsilon = 2$ ).

11. Si los resultados de la encuesta fueran utilizados para identificar a potenciales beneficiarios de la política social, el número de hogares elegibles dependerá del método utilizado para imputar la información omitida.

## XII. Conclusiones

---

1. En este trabajo se han analizado los fundamentos teóricos de distintos procedimientos de imputación dando cuenta de sus bondades y limitaciones. Asimismo, los métodos se han aplicado con un propósito muy preciso, observar su impacto en los indicadores de pobreza y desigualdad.

2. Es posible afirmar que el mejor método de imputación es el que no se aplica, lo que sugiere agotar todo los recursos para minimizar la falta de respuesta total y parcial en una encuesta. La técnica de imputación que genera los mejores resultados y preserva la verosimilitud de los datos, es la que se sustenta en información recabada en el terreno.

3. Todos los métodos estudiados tienen limitaciones y su correcta aplicación depende de la manera en que se comporten los datos faltantes. En la medida de que la falta de respuesta no muestre un patrón aleatorio, la eficacia de todas las metodologías se debilita, aún en procedimientos estadísticamente robustos como imputación múltiple y máxima verosimilitud.

4. Ninguno de los métodos analizados tiene en cuenta la estructura del diseño muestral. En las encuestas complejas, los hogares son seleccionados en diversas etapas que incluyen estratificación y conglomeración del marco de muestreo, lo cual significa probabilidades de selección distintas para los hogares en muestra. Este hecho no puede ser pasado por alto en el análisis de los resultados de la investigación que se esté llevando a cabo.

5. No existe el mejor método de imputación. Cada situación es diferente y la elección del procedimiento de sustitución de datos

depende de la variable de estudio, del porcentaje de datos faltantes, del tipo de encuesta que se analice y del uso que se hará de la información imputada. Por ello, no se aconseja elegir un procedimiento de imputación y aplicarlo en forma generalizada para todas las variables en todas las encuestas.

6. Cada encuesta se realiza bajo circunstancias distintas y está sujeta a diversas fuentes de error —muestrales y no muestrales—. Las condiciones en que se lleva a cabo la capacitación del personal de campo, los problemas asociados al marco de muestreo y la ejecución del trabajo de campo cambian entre encuestas, lo que hace muy difícil que la tasa de no respuesta se repita, a pesar de que se mantenga la calidad en la capacitación del personal de campo y la ejecución de la encuesta en terreno.

7. Se sugiere encarar el análisis de datos sin la elección —*a priori*— de un método de imputación. El análisis exploratorio y la consistencia de la información, darán la pauta para elegir el método que genera los estimadores más eficientes. Lo que funcionó para un estudio, no necesariamente generará buenos resultados en otras investigaciones. En este trabajo ha quedado en evidencia que los métodos de imputación generan ordenamientos distintos de una misma variable.

8. Se reconoce que el objetivo del análisis estadístico es generar estimadores robustos que satisfagan propiedades teóricas de interés; sin embargo, en el estudio de los niveles de vida es imperativo involucrar otros criterios para definir cuál de los métodos de imputación se considera el más adecuado.

9. La sensibilidad de los indicadores de pobreza y desigualdad a la imputación de datos, sugiere la necesidad de elegir el procedimiento que introduzca la menor distorsión posible en la distribución de las variables intervenidas.

10. La literatura recomienda en forma explícita la aplicación de dos procedimientos. El método de imputación por máxima verosimilitud (MV) y el de imputación múltiple (IM). No obstante, no es posible afirmar en forma categórica que el algoritmo de imputación múltiple genera siempre mejores resultados que los métodos simples, y abundan los ejemplos en donde esto se demuestra.

11. En distintos campos de las ciencias sociales se demuestra que el procedimiento *hot-deck* puede ser más eficiente que los métodos de imputación múltiple y de regresión paramétrica, lo que garantiza preservar al máximo la distribución de la variable imputada. Si el proceso de generación de información requiere imputar en distintas bases de datos, este algoritmo puede considerarse como una opción viable que mantiene un equilibrio adecuado entre la teoría y la práctica.

12. Cuando las tasas de no respuesta son muy elevadas en las variables de interés (25% o más), hay que mantener latente la posibilidad de prescindir de la base de datos. Una decisión asumida a partir de un indicador construido con información débil (a partir de imputaciones), podría resultar más onerosa para la hacienda pública que el monto de recursos que se destinaron para ejecutar una encuesta que no generó información confiable.

13. En este trabajo se ha encarado el problema de datos faltantes. Sin embargo, se recomienda una segunda fase que debiera enfocarse al análisis de valores aberrantes o extremos (*outliers*), ya que estos también influyen en la forma de la distribución del ingreso e inciden en las tasas de pobreza y en los indicadores de desigualdad.

14. Considerando que la EPH es una encuesta consolidada, sería interesante utilizar los datos de otra ronda para conocer el efecto de los métodos de imputación en las variables que aquí se han comparado, con el propósito de validar las recomendaciones formuladas en este trabajo.

## Bibliografía

---

- Acock, C., Alan (2005), Working With Missing values, *Journal of Marriage and Family* 67, November.
- Acock, C. A., y D. Demo (1994), *Family diversity and well-being*, Thousand Oaks, C. A. Sage.
- Allison, P. D. (2000), *Multiple Imputation for Missing Data: A Cautionary Tale*, University of Pennsylvania.
- Binder, D. A. (1996), Comment to articles by Rao, Fay and Rubin, *Journal of the American Statistical Association* 91.
- Binder, D. A. y W. Sun (1996), Frequency valid multiple imputation for surveys with complex designs, Business Survey Methods Division, Statistics, Canada.
- CEPAL (2002), Programa para el Mejoramiento de las Encuestas y la Medición de las Condiciones de Vida (MECOVI), Tercer Taller Regional, Santiago de Chile.
- Cohen, J., y P. Cohen (1983), Applied multiple regression/correlation analysis for the behavioral sciences (Sec. ed.), Hillsdale, N. J., Erlbaum.
- Cohen, J., P. Cohen, S. West, y L. Aiken, (2003), Applied multiple regression/correlation analysis for the behavioral sciences (Third ed.), Mahwah, N. J. Erlbaum.
- Cochran, W. G., (1977), Sampling Techniques, Second Edition, John Wiley and Sons, Inc.
- Dempster, A. P., N. M. Lair, y D. B. Rubin (1977), Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39.
- Díaz-Alejandro, Carlos F. (1988), Trade, Development and the World Economy, *Selected Essays*, Andrés Velasco (ed.), Oxford, Basil Blackwell.
- Droesbeke, J. J. y P. Lavallée (1996), La no respuesta en las encuestas. *Metodológica* N° 4.
- Durrant, B. G. (2005), Imputation Methods for Handling Item-Non-response in the Social Science: A Methodological Review, ESRC National Centre for

- Research Methods and Southampton Statistical Science Research Institute, University of Southampton, *NCRM Methods Review Papers*, NCRM/002.
- Fay, R. E. (1993), Valid Inference from imputed survey data, U. S. Bureau of the Census, Washington, D. C.
- \_\_\_ (1992), When are inferences from multiple imputation valid? Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA.
- \_\_\_ (1991), A design-based perspective on missing data variance, Proceedings of the 1991 Annual Research Conference, Washington, D. C., U. S. Bureau of the Census.
- Heitjan, D., y D. B. Rubin (1991), Ignorability and coarse data, *Annals of Statistic* 19.
- Ibrahim, J. G., (1990), Incomplete data in generalized linear models, *Journal of the American Statistical Association*.
- Juster, F. T. y J. P. Smith (1998), Improving the quality of economic data, Lesson from the HRS and AHEAD, *Journal of the American Statistical Association*.
- Kalton, G. y D. Kasprzyk (1986), The treatment of missing survey data, *Survey Methodology*, Vol. 12.
- \_\_\_ (1982), Imputing for Missing Surveys Responses, Proceedings of the Section on Survey Research Methods, American Statistical Association.
- King, G., J. Honaker, A. Joseph, y K. Scheve (2001), Analyzing incomplete political science data, An alternative algorithm for multiple imputation, *American Political Science Review* 95.
- Lehtonen, R. y J. E. Pahkinen (1996), Practical Methods for Design and Analysis of Complex Surveys, Statistics in Practice, Vic Barnett Editor.
- Little, R. J. A. (1995), Modeling the dropout mechanism in repeated-measures studies, *Journal of the American Statistical Association*.
- \_\_\_ (1992), Regression with missing X's, A review, *Journal of the American Statistical Association* 87.
- \_\_\_ (1986), Survey non-response adjustments, *International Statistical Review* 54.
- \_\_\_ (1986), A test of Missing Completely at Random for multivariate data with missing values, *Sociological Methods and Research* 18.
- Little, R. J., y D. Rubin (2002), Statistical analysis with missing data (Sec. Ed.), New York, Wiley.
- \_\_\_ (1987). Statistical analysis with missing data, New York, Wiley.
- Lohr, S. (1999), Sampling Design and Analysis, Duxbury Press.
- Madow, W. G., J. Nisselson y I. Olkin (Eds.) (1983), Incomplete data in sample surveys, vol. 1, Report and case studies, New York, Academic Press.
- Montaquila, M. J. y R. W. Jernigan (1997), Variance estimation in the presence of imputed data, Proceedings of the Survey Research Methods Section of the American Statistical Association.
- Neyman, J. (1937), Outline of a theory of statistical estimation based on the classical theory of probability. Philosophical Transactions of the Royal Society of London, *Series A*.
- Neyman, J. y E. S. Pearson (1933), On the problem of most efficient test of statistical hypotheses. Philosophical Transactions of the Royal Society of London, *Series A*.
- Robins, J. M., A. Rotnitzky y L. P. Zhao (1994), Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association*.
- Roth, P. (1994), Missing data, A conceptual review for applied psychologist, *Personnel Psychology* 47.
- Royston P., (2005), MICE for multiple imputation of *missing values*, MRC Clinical Trials Unit, London, 11th London State User's Meeting.
- Rubin, D. B. (1996), Multiple Imputation after 18+ years (with discussion), *Journal of the American Statistical Association* 91.
- \_\_\_ (1987), Multiple imputation for non-response in surveys. New York, Wiley.
- \_\_\_ (1977), Formalizing subjective notions about the effect of non-respondents in sample surveys, *Journal of the American Statistical Association* 72.
- \_\_\_ (1976), Inference and missing data, *Biometrika* 63.
- Schafer, J. L. (1999), Multiple Imputation, a prime *Statistical Methods in Medical Research* 8.
- \_\_\_ (1997), Analysis of incomplete multivariate data, London, Chapman y Hall.
- Schafer, J. L. y J. W. Graham (2002), Missing Data, Our View of the State of the Art, *Psychological Methods*. vol. 7, No. 2.
- Zeger, S. L., K. Y. Liang y P. S. Albert (1988), Models for longitudinal data: A generalized estimating equation approach, *Biometrics* 44.
- Von Hippel, P. T. (2005), How many imputations are needed? A comment on Hershberger and Fisher (2003), *Structural Equation Modeling* 12.

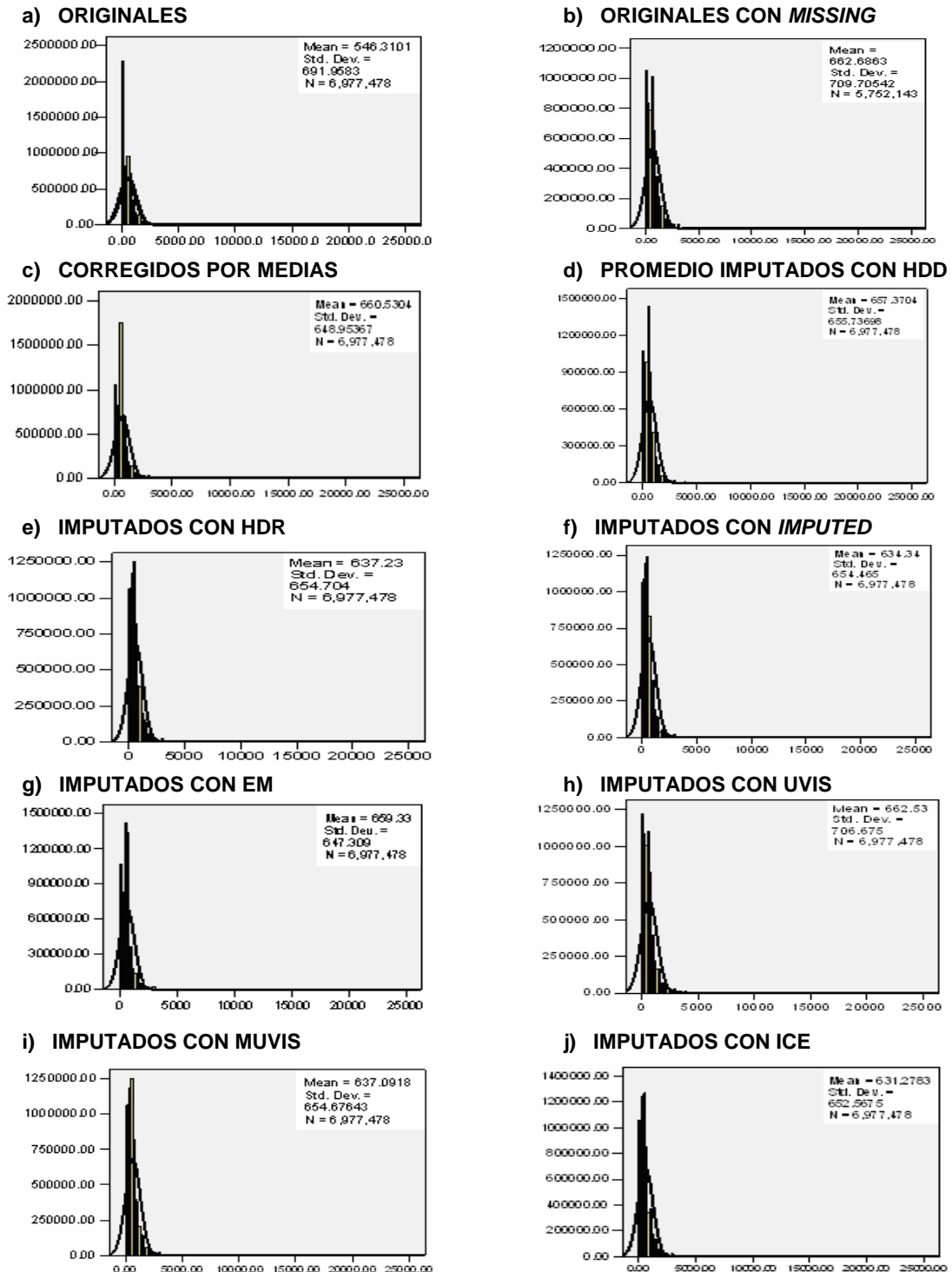
## **Anexos**

---

# Anexo 1

## Estadísticas complementarias

**GRÁFICO A.1**  
**SUELDOS Y SALARIOS**  
*(Frecuencia por miles de hogares)*



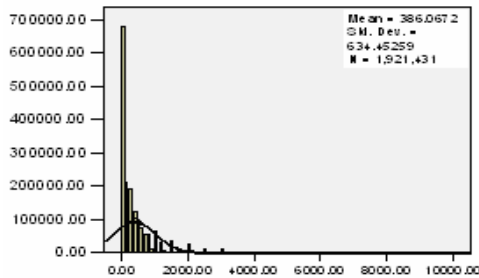
Fuente: Elaboración propia de los autores.



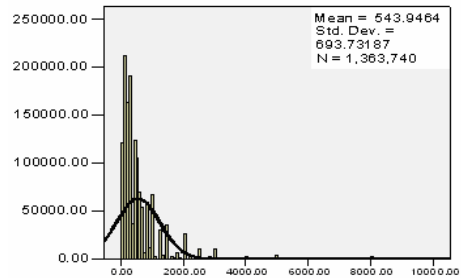
**GRÁFICO A.2  
GANANCIAS**

*Frecuencia por miles de hogares)*

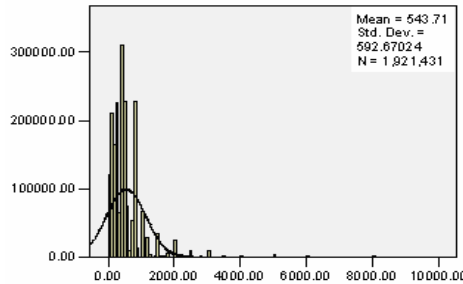
**a) ORIGINALES**



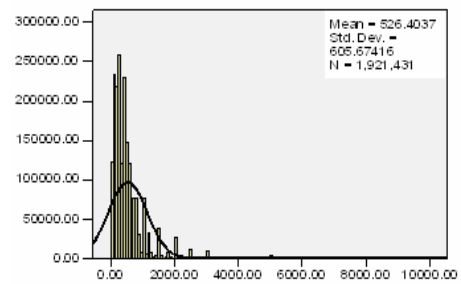
**b) ORIGINALES CON MISSING**



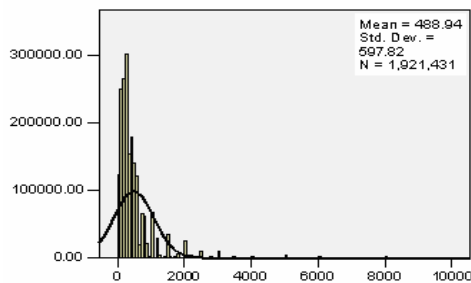
**c) CORREGIDAS POR MEDIAS**



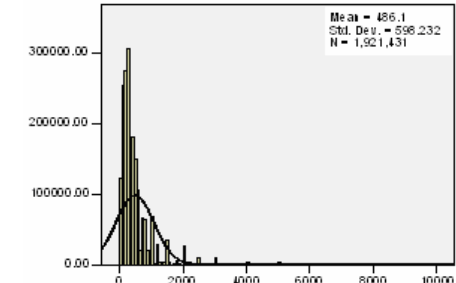
**d) PROMEDIO IMPUTADAS CON HDD**



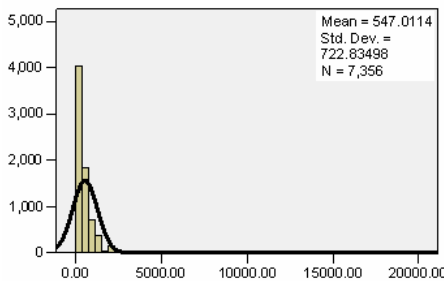
**e) IMPUTADAS CON HDR**



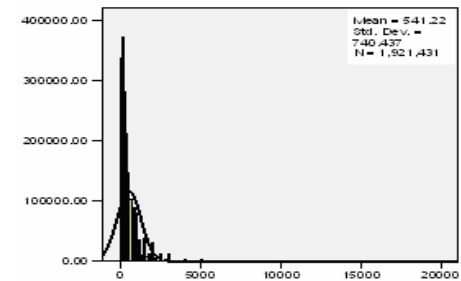
**f) IMPUTADAS CON IMPUTED**



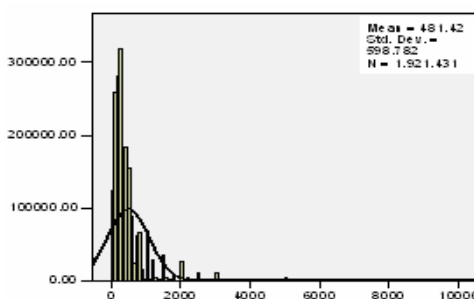
**g) IMPUTADOS CON EM**



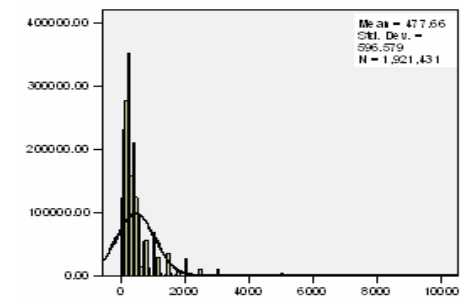
**h) IMPUTADAS CON UVIS**



**i) IMPUTADAS CON MUVIS**

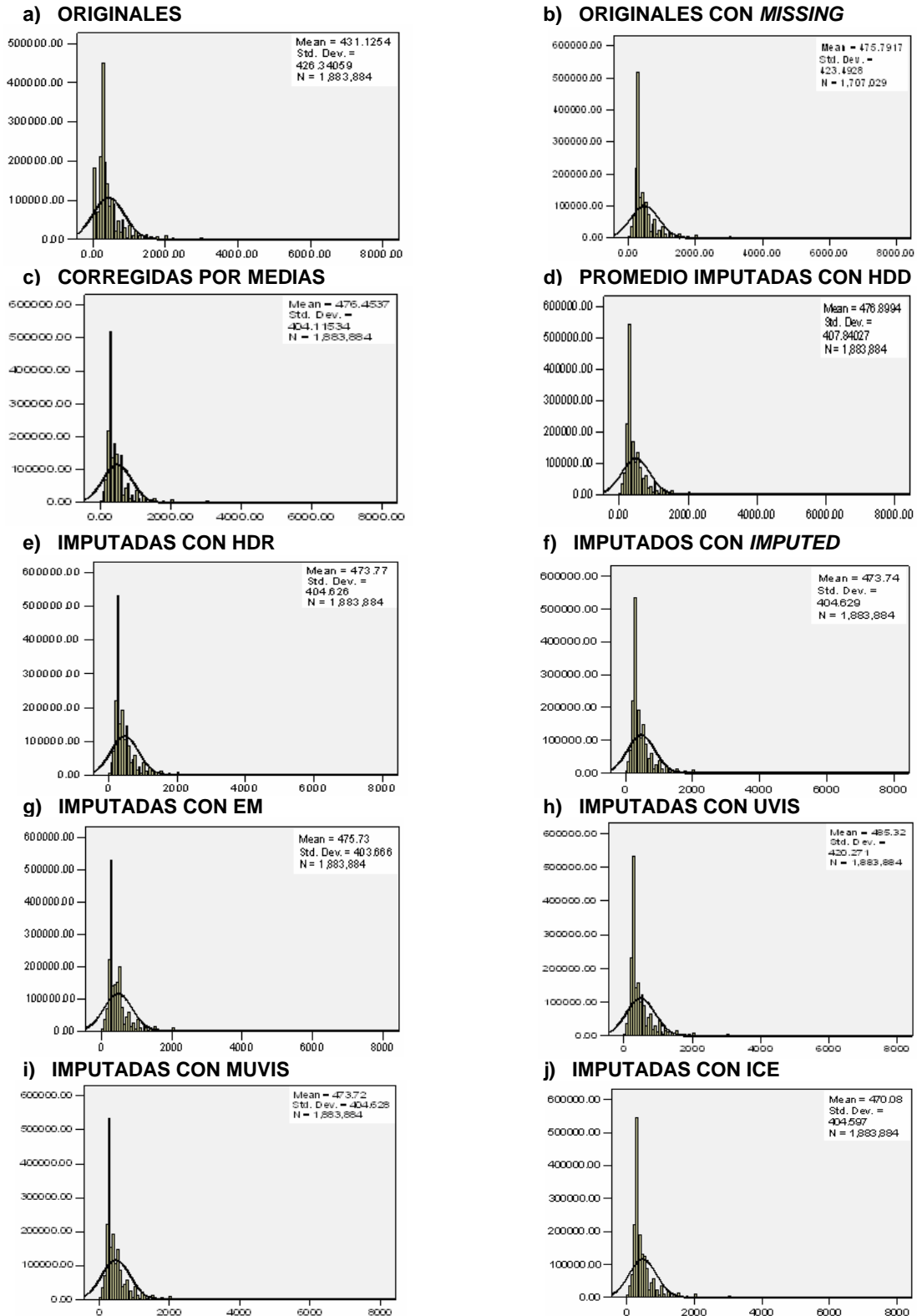


**j) IMPUTADAS CON ICE**



Fuente: Elaboración propia de los autores.

**GRÁFICO A.3**  
**JUBILACIONES Y PENSIONES**  
*(Frecuencia por miles de hogares)*



Fuente: Elaboración propia de los autores.

**CUADRO A.1**  
**MEDIDAS DE ERROR, ASIMETRÍA Y KURTOSIS**

**a) SUELDOS Y SALARIOS**

Método de imputación	Promedio	Desviación estándar	Skewnes	Kurtosis
Original (ORIG)	546,58	692,04	7,87	188,66
Listwise	662,69	709,73	8,38	201,35
Medias condicionadas	660,53	649,10	9,04	238,01
Hot-deck (HDD)	657,39	655,90	8,80	228,39
Hot-deck con regresión (HDR)	637,23	654,85	8,91	230,93
Regresión condicionada (R)	634,34	654,61	8,93	231,43
Máxima verosimilitud (MV)	659,37	647,46	9,11	240,53
Imputación simple (IS)	662,53	706,81	7,55	172,68
Imputación múltiple (IM)	637,09	654,82	8,91	230,98
Imputación múltiple (ICE)	631,28	652,72	9,02	234,33

**b) GANANCIAS**

Método de imputación	Promedio	Desviación estándar	Skewnes	Kurtosis
Original (ORIG)	386,07	634,45	4,67	39,03
Listwise	543,94	693,73	4,41	34,10
Medias condicionadas	686,06	592,67	5,02	45,45
Hot-deck (HDD)	656,65	605,67	4,83	42,37
Hot-deck con regresión (HDR)	527,47	597,82	5,14	45,75
Regresión condicionada (R)	593,03	598,23	5,14	45,72
Máxima verosimilitud (MV)	664,50	592,80	5,02	42,44
Imputación simple (IS)	669,82	740,43	6,82	103,31
Imputación múltiple (IM)	599,58	598,78	5,15	45,72
Imputación múltiple (ICE)	581,62	596,58	5,23	46,52

**c) JUBILACIONES Y PENSIONES**

Método de imputación	Promedio	Desviación estándar	Skewnes	Kurtosis
Original (ORIG)	431,12	426,34	4,47	45,10
Listwise	475,79	423,49	4,82	49,32
Medias condicionadas	476,45	404,11	5,02	54,14
Hot-deck (HDD)	476,89	407,84	4,90	52,08
Hot-deck con regresión (HDR)	473,77	404,63	5,02	53,99
Regresión condicionada (R)	473,74	404,63	5,02	53,99
Máxima verosimilitud (MV)	475,73	403,67	5,04	54,43
Imputación simple (IS)	485,32	420,27	4,58	46,06
Imputación múltiple (IM)	473,72	404,63	5,02	53,99
Imputación múltiple (ICE)	470,08	404,60	5,05	54,19

Fuente: Elaboración propia de los autores.

**CUADRO A.2**  
**ESTIMADORES Y SU ERROR ESTÁNDAR POR DISTINTOS MÉTODOS DE IMPUTACIÓN**

**a) SUELDOS Y SALARIOS**

	Promedio	Error Estándar	+/- 2 desv.std.	Límite inferior	Límite superior
Original	546	4,49	8,98	537	555
Listwise	663	4,58	9,16	654	672
Medias	661	3,44	6,88	654	667
HDD	657	3,72	7,44	650	665
HDR	637	3,92	7,84	629	645
IMPUTE	634	3,83	7,66	627	642
EM	659	3,23	6,46	653	666
UVIS	663	4,07	8,14	654	671
MVIS	637	3,43	6,86	630	644
ICE	631	3,75	7,50	624	639

**b) GANANCIAS**

	Promedio	Error Estándar	+/- 2 desv.std.	Límite inferior	Límite superior
Original	386	6,41	12,82	373	399
Listwise	544	8,64	17,28	527	561
Medias	544	6,01	12,02	532	556
HDD	526	5,43	10,86	516	537
HDR	489	6,13	12,26	477	501
IMPUTE	486	6,06	12,12	474	498
EM	544	6,62	13,24	530	557
UVIS	541	8,34	16,68	525	558
MVIS	481	6,62	13,24	468	495
ICE	478	6,45	12,90	465	491

**c) JUBILACIONES Y PENSIONES**

	Promedio	Error Estándar	+/- 2 desv.std.	Límite inferior	Límite superior
Original	431	5,23	10,46	421	442
Listwise	476	5,17	10,34	465	486
Medias	476	4,50	9,00	467	485
HDD	477	4,60	9,20	468	486
HDR	474	4,44	8,88	465	483
IMPUTE	474	4,65	9,30	464	483
EM	476	4,80	9,60	466	485
UVIS	485	4,72	9,44	476	495
MVIS	474	4,29	8,58	465	482
ICE	470	4,50	9,00	461	479

Fuente: Elaboración propia de los autores.

**CUADRO A.3**  
**ÍNDICES DE DESIGUALDAD PARA DISTINTOS PROCEDIMIENTOS DE IMPUTACIÓN**

Indicadores	Medias	HDR	R	IS	IM	ICE	HDD	MV
Desviación media relativa	0,384	0,388	0,395	0,404	0,388	0,388	0,381	0,376
Coefficiente de variación	2,249	2,379	2,337	2,285	2,388	2,403	2,178	2,166
Coefficiente de variación <sup>2</sup>	5,059	5,659	5,464	5,222	5,700	5,776	4,743	4,693
Varianza de logaritmos	0,953	0,928	0,959	1,019	0,926	0,918	0,958	0,946
Desviación estándar de logaritmos	0,976	0,963	0,979	1,009	0,962	0,958	0,979	0,973
Gini	0,531	0,536	0,543	0,551	0,536	0,536	0,527	0,522
Mehran	0,658	0,660	0,667	0,679	0,659	0,659	0,656	0,651
Piesch	0,468	0,474	0,480	0,488	0,474	0,475	0,462	0,458
Kakwani	0,235	0,239	0,245	0,252	0,240	0,240	0,232	0,228
GE (-1,0)	0,904	0,877	0,911	0,981	0,874	0,867	0,906	0,893
GE (0)/desviación media de logs.	0,522	0,526	0,540	0,562	0,526	0,525	0,515	0,507
GE (1,0) Theil	0,633	0,661	0,666	0,673	0,663	0,666	0,615	0,607
GE (2,0)	2,530	2,829	2,732	2,611	2,850	2,888	2,371	2,346
Atkinson (0,5)	0,243	0,249	0,253	0,259	0,249	0,250	0,239	0,236
Atkinson (1,0)	0,406	0,409	0,417	0,430	0,409	0,409	0,403	0,398
Atkinson (1,5)	0,534	0,532	0,542	0,558	0,532	0,530	0,533	0,528
Atkinson (2,0)	0,644	0,637	0,646	0,662	0,636	0,634	0,644	0,641
Atkinson (2,5)	0,746	0,736	0,743	0,756	0,735	0,734	0,749	0,747
Atkinson (3,0)	0,849	0,840	0,844	0,852	0,840	0,838	0,853	0,853
Atkinson (3,5)	0,924	0,918	0,921	0,924	0,918	0,917	0,926	0,926
Atkinson (4,0)	0,958	0,956	0,957	0,958	0,955	0,955	0,960	0,960

Fuente: Elaboración propia de los autores.

## Anexo 2

### Paquetes estadísticos para efectuar imputaciones

Durante la década de los ochenta, la utilización de algoritmos de imputación —simple y múltiple— estaba limitada a los investigadores que disponían de programas elaborados para tal fin. En la actualidad, y como se muestra en el cuadro A.4, existen distintas aplicaciones comerciales y de acceso libre que pueden ser utilizadas.

**CUADRO A.4**  
**PAQUETES DISPONIBLES PARA IMPUTACIÓN DE DATOS**

#### a) LIBRE

Paquete	Acceso	Imputación		Máxima verosimilitud
		Simple	Múltiple	
Amelia	<a href="http://gking.harvard.edu/amelia/">http://gking.harvard.edu/amelia/</a>		Sí	
CAT	<a href="http://www.stat.psu.edu/~jls/misoftwa.html#aut">http://www.stat.psu.edu/~jls/misoftwa.html#aut</a>		Sí	
EMCOV	<a href="http://methcenter.psu.edu/downloads/EMCO.html">http://methcenter.psu.edu/downloads/EMCO.html</a>	Sí		
NORM	<a href="http://www.stat.psu.edu/~jls/misoftwa.html#aut">http://www.stat.psu.edu/~jls/misoftwa.html#aut</a>	Sí	Sí	
MICE	<a href="http://www.multiple-imputation.com">http://www.multiple-imputation.com</a>		Sí	
MIXED	Gratis con R, se vende con S-Plus <a href="http://www.stat.psu.edu/~jls/misoftwa.html#aut">http://www.stat.psu.edu/~jls/misoftwa.html#aut</a>	Sí	Sí	
MX	<a href="http://www.vcu.edu/mx/">http://www.vcu.edu/mx/</a>			Sí
PAN	Gratis con R, se vende con S-Plus <a href="http://www.stat.psu.edu/~jls/misoftwa.html#aut">http://www.stat.psu.edu/~jls/misoftwa.html#aut</a>	Sí	Sí	

#### b) VENTA

Paquete	Acceso	Imputación		Máxima verosimilitud
		Simple	Múltiple	
AMOS	<a href="http://www.spss.com">http://www.spss.com</a>			Sí
EQS	<a href="http://www.mvsoft.com/">http://www.mvsoft.com/</a>			Sí
HLM	<a href="http://www.ssicentral.com/hlm/index.html">http://www.ssicentral.com/hlm/index.html</a>			Sí
LISREL	<a href="http://www.ssicentral.com/lisrel/mainlis.html">http://www.ssicentral.com/lisrel/mainlis.html</a>			Sí
Mplus	<a href="http://www.statmodel.com">http://www.statmodel.com</a>		Sí	Sí
SAS	<a href="http://www.sas.com">http://www.sas.com</a>		Sí	
SOLAS	<a href="http://www.statsol.ie/solas/imputationtechnics.htm">http://www.statsol.ie/solas/imputationtechnics.htm</a>	Sí	Sí	
S-Plus	<a href="http://www.stat.psu.edu/~jls/misoftwa.html#aut">http://www.stat.psu.edu/~jls/misoftwa.html#aut</a> pan, cat mixed, y otras opciones	Sí	Sí	
SPSS	<a href="http://www.spss.com">http://www.spss.com</a> (módulo opcional)	Sí		
STATA	<a href="http://www.stata.com">http://www.stata.com</a>	Sí	Sí	

Fuente: Working with Missing values, Alan C. Acock, 2005, *Journal of Marriage and Family* 67, noviembre.

En este trabajo se utilizaron los programas SPSS, SAS y STATA. Las capacidades de cada uno de ellos son distintas en lo que concierne a los procedimientos para imputación de datos. El programa que se considera más completo es el STATA 9, ya que incorpora una gran variedad de algoritmos que permiten efectuar sustituciones de datos mediante distintas opciones metodológicas.

El SAS incorpora (el STATA no) el algoritmo EM que utiliza métodos de máxima verosimilitud, además de los comandos PROC MI y PROC ANALYSE para efectuar imputación múltiple y para combinar los archivos de datos que se generan a partir de las distintas simulaciones que se llevan a cabo. Por su parte, la opción MVA (*missing value analysis*) del SPSS incorpora los

métodos tradicionales (*listwise*, *pairwise* y de medias) y tiene la posibilidad de efectuar imputaciones simples por medio de regresión y del algoritmo de máxima verosimilitud EM.<sup>54</sup> Adicionalmente, esta rutina tiene incorporada la opción de la prueba de Little, para llevar a cabo docimasia de hipótesis acerca del patrón aleatorio de los datos.

Los procedimientos de imputación múltiple incorporados en los paquetes SAS y STATA, están basados en los algoritmos desarrollados por Schafer tal como se muestra en el cuadro A.4.

STATA incorpora comandos que permiten conocer la manera en que se distribuyen los datos faltantes (MVPATTERNS), así como opciones de imputación para los procedimientos *hot-deck*, *hot-deck* por regresión, imputación por regresión (*IMPUTE*), imputación simple (IVIS), imputación múltiple (MVIS), y en la versión 9 se incluye la rutina desarrollada por Royston (2005), (MICE) que, además de ser más eficiente, permite combinar los archivos de datos que se generan. También dispone de distintas herramientas para combinar (MIJOIN, MISPLITE y MICOMBINE) los archivos que se forman por la aplicación de los comandos mencionados.

---

<sup>54</sup> Cuando se trabaja con series de tiempo existen 5 distintas opciones para imputar datos faltantes mediante la opción TRANSFORM. Las opciones disponibles son: promedios, promedios de vecinos más cercanos, mediana de vecinos cercanos, interpolación lineal y tendencias lineales.

## Anexo 3

### Ejemplos de uso de los comandos para imputación en STATA

Por considerarlo de interés para los usuarios de bases de datos que enfrentan la necesidad de sustituir datos faltantes, a continuación se presenta el código STATA (versión 9) que se utilizó en este trabajo.

Se aclara que no fueron aplicadas todas las opciones incluidas en los comandos, por lo que en los manuales y en la ayuda contenida en el paquete se encontrará más información acerca de las múltiples capacidades que incorporan los algoritmos de imputación mencionados.

Como fase inicial, se confirma la necesidad de restringir el análisis a las submuestra de interés. Por ejemplo, en el caso de los asalariados (*sysorimi*) se trabajó con el grupo de observaciones que cumple con los siguientes requisitos: personas ocupadas (*conduct=1*) que se clasificaron como asalariados (*categ=3*) ocupados, de ambos sexos y con nivel de educación y rama de actividad dentro de los rangos definidos para el análisis (*niveduc<99* y *ramar1>0* y *ramar1<99*). Es importante recordar que en STATA los datos omitidos (*missing values*) se identifican por el carácter “.”.

#### CUADRO A.5 AJUSTE DE MODELOS DE REGRESIÓN PARA IMPUTAR DATOS

Algunos de los métodos de imputación aplicados utilizan técnicas de regresión para modelar la variable de respuesta, relacionado la variable de interés con un grupo de covariables correlacionadas. En este trabajo, las ecuaciones de Mincer que se utilizaron para modelar el comportamiento de las tres variables de ingresos analizadas (sueldos y salarios, ganancias y jubilaciones y pensiones), y los modelos se ajustaron con la sintaxis que se muestra a continuación.

Se debe tener claro que los modelos de regresión utilizados imputan valores en logaritmos, por lo que es necesario efectuar las transformaciones correspondientes para que los archivos contengan datos en las unidades originales.

#### a) EMPLEADOS (SUELDOS Y SALARIOS)

```
. reg lsysorimi exp exp2 anoseduc sexo if (conduct==1 & categ==3) & /*
*/niveduc<99 & (ramar1>0 & ramar1<99)
```

Source	SS	df	MS	Number of obs = 23299		
Model	4372.89671	4	1093.22418	F( 4, 23294) = 1744.25		
Residual	14599.6995	23294	.626757941	Prob > F = 0.0000		
Total	18972.5962	23298	.814344415	R-squared = 0.2305		
				Adj R-squared = 0.2304		
				Root MSE = .79168		
lsysorimi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exp	.0517435	.0013448	38.48	0.000	.0491075	.0543795
exp2	-.0007438	.0000256	-29.06	0.000	-.000794	-.0006937
anoseduc	.0877892	.0013713	64.02	0.000	.0851013	.0904771
sexo	.4735434	.0104089	45.49	0.000	.4531412	.4939457
_cons	4.320448	.0233645	184.92	0.000	4.274652	4.366244

**b) CUENTA PROPIA Y PATRONES (GANANCIAS)**

```
.reg lganorimi exp exp2 anosedu sexo if conduct==1 & /*
*/(categ==1 | categ==2) & anoseduc<99 & (ramar1>0 & ramar1<99)
```

Source	SS	df	MS	Number of obs =	6466
Model	1817.2517	4	454.312924	F( 4, 6461) =	486.74
Residual	6030.52181	6461	.933372823	Prob > F =	0.0000
				R-squared =	0.2316
				Adj R-squared =	0.2311
Total	7847.77351	6465	1.21388608	Root MSE =	.96611

lganorimi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
exp	.070817	.0029178	24.27	0.000	.0650972 .0765368
exp2	-.0010402	.0000487	-21.37	0.000	-.0011356 -.0009448
anoseduc	.1044805	.0031216	33.47	0.000	.0983611 .1105999
sexo	.5020693	.0257268	19.52	0.000	.4516363 .5525024
_cons	3.530786	.0575714	61.33	0.000	3.417927 3.643645

**c) JUBILADOS Y PENSIONADOS**

```
. reg ljuborimi anosedu exp exp2 sexo if conduct==4
```

Source	SS	df	MS	Number of obs =	6984
Model	524.615875	4	131.153969	F( 4, 6979) =	381.20
Residual	2401.14042	6979	.344052217	Prob > F =	0.0000
				R-squared =	0.1793
				Adj R-squared =	0.1788
Total	2925.7563	6983	.418982715	Root MSE =	.58656

ljuborimi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
anoseduc	.0487992	.0018868	25.86	0.000	.0451006 .0524979
exp	.0317077	.0022286	14.23	0.000	.027339 .0360763
exp2	-.0003	.0000227	-13.25	0.000	-.0003444 -.0002556
sexo	.2473725	.0143167	17.28	0.000	.2193075 .2754376
_cons	4.615811	.0588244	78.47	0.000	4.500498 4.731125

Fuente: Elaboración propia de los autores.

### CUADRO A.6 PATRÓN DE DATOS FALTANTES

Antes de decidirse por la aplicación de alguno de los procedimientos de imputación disponibles, es necesario analizar la manera en que se distribuyen en la muestra los datos faltantes, y corroborar si es posible concluir que la información omitida sigue un patrón aleatorio (MCAR o MAR)<sup>55</sup>.

En este sentido, además de analizar gráficamente la distribución de los datos omitidos en los subgrupos formados, se sugiere efectuar pruebas de hipótesis para certificar si se cumplen las hipótesis básicas en la que se sustentan los procedimientos de imputación aplicados. Esta posibilidad, sin embargo, no está disponible en las rutinas de STATA, por lo que fue necesario efectuar el análisis con el paquete SPSS.

**a) EMPLEADOS (SUELDOS Y SALARIOS)**

```
mvpatterns sysorimi anosedu ramar1 sexo if (conduct==1 & categ==3) & /*
*/anoseduc<99 & (ramar1>0 & ramar1<99), nodrop
```

<sup>55</sup> La información generada mediante el comando *mvpatterns* de STATA, se complementó con los resultados generados por el SPSS. A diferencia del STATA, SPSS entrega como resultado el estadístico de Little (1986), para efectuar la prueba de hipótesis acerca del patrón MCAR de los datos faltantes.



```
-----
Variable | type      obs   mv   variable label
-----
sysorimi | int       23301 4048  sueldos y salarios originales con missing
anoseduc | byte     27349   0   anos de estudio
ramarl   | byte     27349   0   rama de actividad en 4 categorias
sexo     | byte     27349   0   sexo
-----
Patterns of missing values
+-----+
|_pattern  _mv  _freq|
+-----+
|      +++  0  23301|
|     .+++  1  4048|
+-----+
```

#### b) CUENTA PROPIA Y PATRONES (GANANCIAS)

```
mvpatterns ganorimi anosedu ramarl sexo if conduct==1 & /*
(categ==1 | categ==2) & /*
*/anoseduc<99 & (ramarl>0 & ramarl<99), nodrop
```

```
-----
Variable | type      obs   mv   variable label
-----
ganorimi | int       6466 2464  ganancias originales
anoseduc | byte     8930   0   anos de estudio
ramarl   | byte     8930   0   rama de actividad en 4 categorias
sexo     | byte     8930   0   sexo
-----
Patterns of missing values
+-----+
|_pattern  _mv  _freq|
+-----+
|      +++  0  6466|
|     .+++  1  2464|
+-----+
```

#### c) JUBILADOS Y PENSIONADOS

```
mvpatterns juborimi anosedu sexo if conduct==4, nodrop
```

```
-----
Variable | type      obs   mv   variable label
-----
juborimi | int       6984 599  jubilaciones originales
anoseduc | byte     7583   0   anos de estudio
sexo     | byte     7583   0   sexo
-----
Patterns of missing values
+-----+
|_pattern  _mv  _freq|
+-----+
|      +++  0  6984|
|     .++  1  599|
+-----+
```

Fuente: Elaboración propia de los autores.

### CUADRO A.7 EL PROCEDIMIENTO *HOT-DECK*

Si se desea efectuar la imputación de las observaciones de la variable sueldos y salarios que no tienen datos, es posible aplicar el procedimiento *hot-deck* utilizando la sintaxis que se muestra a continuación.

Mediante la opción *impute* (#) se indica el número de simulaciones que se desean efectuar —en este caso 5. Como resultado de la sintaxis utilizada, se generan tantos archivos como simulaciones se soliciten en la opción *impute*.

Se debe tener cuidado, debido a que las simulaciones generadas se guardan en un directorio y archivo distintos al que se está trabajando. En este ejemplo, las cinco simulaciones solicitadas quedarán resguardadas en los archivos `imptuhdds1`, ..., `imputdds5`.

Estos archivos se respaldan en el directorio en donde el usuario mantiene en forma regular sus datos STATA. En caso de que existan dudas del lugar en que se alojaron los archivos, es posible teclear el comando `cd` y en pantalla aparecerá la ruta de acceso y el nombre del directorio.

#### a) EMPLEADOS (SUELDOS Y SALARIOS)

```
hotdeck sysorimi if condat==1 & categ==3 & (ramar1>0 & ramar1<99) /*
  */& anoseduc<99 using imputhdds, store by(ramar1 sexo)impute(5)keep/*
  */(id_hogar factorex condat categ sys syscor ramar1 sexo anoseduc parentco)
```

DELETING all matrices...  
Table of the Missing data patterns signifies missing and - is not missing  
Varlist order: sysorimi

pattern	Freq.	Percent	Cum.
*	4,048	4.27	4.27
-	90,724	95.73	100.00
Total	94,772	100.00	

WARNING: When the <command> option is not selected then no analysis is performed on the imputed datasets

#### b) CUENTA PROPIA Y PATRONES (GANANCIAS)

```
hotdeck ganorimi if condat==1 & (categ==1 | categ==2) & (ramar1>0 & /*
  */ramar1<99) & anoseduc<99 using imputhddg, store by(ramar1 sexo) /*
  */imput(5) keep(id_hogar factorex condat categ gan gancor ramar1 sexo /*
  */niveduc anoseduc parentco)
```

DELETING all matrices...  
Table of the Missing data patterns signifies missing and - is not missing  
Varlist order: ganorimi

Pattern	Freq.	Percent	Cum.
*	2,464	27.59	27.59
-	6,466	72.41	100.00
Total	8,930	100.00	

#### c) JUBILADOS Y PENSIONADOS

```
hotdeck juborimi if condat==4 using imputhddj, store by(sexo)imput(5) /*
  */keep(id_hogar factorex condat yjub yjubcor sexo parentco)
```

DELETING all matrices...  
Table of the Missing data patterns signifies missing and - is not missing  
Varlist order: juborimi

pattern	Freq.	Percent	Cum.
*	599	0.63	0.63
-	94,173	99.37	100.00
Total	94,772	100.00	

WARNING: When the <command> option is not selected then no analysis is performed on the imputed datasets  
8930

Fuente: Elaboración propia de los autores.

Nota: El *WARNING* se presenta debido a que no se ajustó ningún modelo en la opción *command*. No debe interpretarse como error de sintaxis.

Si se desean combinar las simulaciones generadas, se recomienda cambiar el nombre de la variable en cada uno de los archivos generados ya que STATA repite el nombre de la variable imputada en cada uno de los archivos creados.

Para lograr este propósito, en la opción *keep* se especifican las variables que se desean conservar en los archivos que contienen los datos imputados. Se sugiere utilizar nombres de variables que permitan identificar plenamente las observaciones imputadas.

**CUADRO A.8  
EL PROCEDIMIENTO *HOT-DECK* CON REGRESIÓN**

Si se conoce que existe una relación de dependencia entre la variable de interés y un vector de covariables, el comando *hot-deck* permite hacer explícita la forma del modelo a partir de la opción *command*, y efectúa la imputación a partir de métodos de regresión.

**a) EMPLEADOS (SUELDOS Y SALARIOS)**

```
hotdeck lsysorimi anoseduc sexo exp exp2 if ((categ==3 & conduct==1) /*
*/& anoseduc<99 & (ramar1>0 & ramar1<99)) using imputhdrs,store by /*
*/(ramar1 sexo) impute(5) keep(id_hogar factorex conduct categ /*
*/anoseduc sys syscor sysorimi ramar1 sexo parentco) com(reg lsysorimi /*
*/anoseduc sexo exp exp2) parms (anoseduc sexo exp exp2 _cons)
```

DELETING all matrices....

Table of the Missing data patterns signifies missing and - is not missing  
Varlist order: lsysorimi anoseduc sexo exp exp2

pattern	Freq.	Percent	Cum.
*--**	50,357	53.13	53.13
*----	21,114	22.28	75.41
-----	23,301	24.59	100.00
Total	94,772	100.00	

Al igual que en el caso anterior, en esta opción se generan tanto archivos como simulaciones sean requeridas. En este ejemplo se solicitan 5 imputaciones y los archivos se guardarán con los nombres *imputhdrs1*, ..., *imputhdrs5*, por lo que esta opción puede ser entendida como un procedimiento de imputación múltiple.

Como resultado de la ejecución de la sintaxis se generan los promedios de los parámetros estimados —de las cinco simulaciones efectuadas—, los cuales deben utilizarse para imputar los datos faltantes.

Observe que en la tabla de salida la columna *Average Coef.* representa los valores de los coeficientes de regresión que se deben aplicar para generar los valores imputados, los cuales se obtuvieron como un promedio simple de las cinco simulaciones efectuadas.

	Number of Obs.	=	27349				
	No. of Imputations	=	5				
	% Lines of Missing Data	=	75.413624 %				
	F( 73.240 ,5)	=	3.46e+05				
	Prob > F	=	0.0000				
Variable	Average Coef.	Between Imp. SE	Within Imp. SE	Total SE	df	t	p-value
anoseduc	0.0881	0.000	0.001	0.001	719.4	67.023	0.000
sexo	0.4683	0.003	0.010	0.010	411.2	46.194	0.000
exp	0.0521	0.001	0.001	0.001	89.8	37.302	0.000
exp2	-0.0007	0.000	0.000	0.000	62.6	-27.359	0.000
_cons	4.3162	0.006	0.022	0.023	495.1	190.923	0.000

```
-----+-----
Variable | [95% Conf. Interval]
-----+-----
anoseduc | 0.0852    0.0911
sexo     | 0.4454    0.4911
exp      | 0.0489    0.0553
exp2     | -0.0008   -0.0007
_cons    | 4.2654    4.3671
-----+-----
```

**b) CUENTA PROPIA Y PATRONES (GANANCIAS)**

```
hotdeck lganorimi anoseduc sexo exp exp2 if (((categ==1 | categ==2) /*
  */& conduct==1) & anoseduc<99 & (ramar1>0 & ramar1<99)) using /*
  */imputhdrj,store by(ramar1 sexo) keep(id hogar factorex conduct /*
  */categ anoseduc gan gancor ramar1 sexo parentco)com(reg lganorimi /*
  */anoseduc sexo exp exp2) parms(anoseduc sexo exp exp2 _cons) impute(5)
```

DELETING all matrices....

Table of the Missing data patterns signifies missing and - is not missing

Varlist order: lganorimi anoseduc sexo exp exp2

```
-----+-----
pattern | Freq.    Percent    Cum.
-----+-----
*----  | 2,464    27.59     27.59
----   | 6,466    72.41     100.00
-----+-----
Total   | 8,930    100.00
-----+-----
8930
```

```
Number of Obs.      = 8930
No. of Imputations = 5
% Lines of Missing Data = 27.592385 %
F( 226.200 ,5)      = 1.64e+05
Prob > F             = 0.0000
```

```
-----+-----
Variable | Average  Between  Within  Total  df    t      p-value
          | Coef.    Imp. SE  Imp. SE  SE
-----+-----
anoseduc | 0.1052   0.002    0.003    0.003  34.9  32.371  0.000
sexo     | 0.5151   0.019    0.022    0.030  16.9  16.937  0.000
exp      | 0.0713   0.002    0.002    0.003  20.2  21.549  0.000
exp2     | -0.0010  0.000    0.000    0.000  14.8  -17.659 0.000
_cons    | 3.5187   0.034    0.049    0.061  29.1  57.438  0.000
-----+-----
```

```
-----+-----
Variable | [95% Conf. Interval]
-----+-----
anoseduc | 0.0976    0.1129
sexo     | 0.4403    0.5899
exp      | 0.0632    0.0793
exp2     | -0.0012   -0.0009
_cons    | 3.3739    3.6635
-----+-----
```

**c) JUBILADOS Y PENSIONADOS**

```
hotdeck ljuborimi exp exp2 anoseduc sexo if conduct==4 using imputhdrj,
  store by(sexo)/*
  */keep(id hogar factorex conduct sexo yjub yjubcor parentco)/*
  */com(reg ljuborimi exp exp2 anoseduc sexo) parms(exp exp2 anoseduc
  sexo _cons) impute(5)
```

DELETING all matrices....

Table of the Missing data patterns signifies missing and - is not missing

Varlist order: ljuborimi exp exp2 anoseduc sexo

```
-----+-----
pattern | Freq.    Percent    Cum.
-----+-----
***--  | 50,357   53.13     53.13
*----  | 37,431   39.50     92.63
----   | 6,984    7.37     100.00
-----+-----
Total   | 94,772   100.00
-----+-----
```

Variable	Average Coef.	Between Imp. SE	Within Imp. SE	Total SE	df	t	p-value
exp	0.0317	0.001	0.002	0.002	927.2	14.260	0.000
exp2	-0.0003	0.000	0.000	0.000	707.0	-13.210	0.000
anoseduc	0.0490	0.000	0.002	0.002	6768.3	26.674	0.000
sexo	0.2512	0.003	0.014	0.014	1155.5	17.740	0.000
_cons	4.6131	0.009	0.057	0.058	4186.2	80.046	0.000

Variable	[95% Conf. Interval]	
exp	0.0267	0.0367
exp2	-0.0004	-0.0002
anoseduc	0.0449	0.0531
sexo	0.2194	0.2830
_cons	4.4839	4.7423

Fuente: Elaboración propia de los autores.

### CUADRO A.9 IMPUTACIÓN POR REGRESIÓN

El método de regresión imputa valores a partir de ajustar un modelo que vincula la variable de interés con un vector de covariables. La variable imputada, en este caso (*lsysimp1*), se guarda en el archivo de trabajo y en los resultados que se generan se especifica el porcentaje y número de observaciones que fueron imputadas.

#### a) EMPLEADOS (SUELDOS Y SALARIOS)

```
impute lsysorimi anoseduc exp exp2 sexo if ramar1>0 & ramar1<99 & /*
  */categ==3 & conduct==1 & anoseduc<99,gen(lsysimp1) 14.80% (4048) /*
  */observations imputed
```

#### b) CUENTA PROPIA Y PATRONES (GANANCIAS)

```
impute lganorimi anoseduc exp exp2 sexo if ramar1>0 & ramar1<99 & /*
  *//(categ==1 | categ==2) & > conduct==1 & anoseduc<99,gen(lganimp1) /*
  */27.59% (2464) observations imputed
```

#### c) JUBILADOS Y PENSIONADOS

```
impute ljuborimi exp exp2 anoseduc sex if conduct==4,gen(ljubimp1) /*
  */7.90% (599) observations imputed
```

Fuente: Elaboración propia de los autores.

### CUADRO A.10 IMPUTACIÓN SIMPLE

Este método imputa valores a partir de un modelo de regresión. El algoritmo genera sólo una simulación y al finalizar la ejecución del comando se reporta el número de observaciones que fueron imputadas.

La variable imputada, en este caso (*lsysiuvis*), se guarda en el archivo de trabajo y se especifica el porcentaje y número de observaciones que fueron imputadas.

**a) EMPLEADOS (SUELDOS Y SALARIOS)**

```

uvis reg lsysorimi anoseduc exp exp2 sexo if ramar1>0 & ramar1<99 & /*
*/categ==3 & conduct==1 & anoseduc<99, gen(lsysiuvis) [imputing by /*
*/drawing from conditional distribution without bootstrap] /*
*/4048 missing observations on lsysorimi imputed from 23301 complete /*
*/observations.

```

**b) CUENTA PROPIA Y PATRONES (GANANCIAS)**

```

uvis reg lganorimi anoseduc exp exp2 sexo if ramar1>0 & ramar1<99 & /*
*/(categ==1 | categ==2) & conduct==1 & anoseduc<99, gen(lganiuvis)
*/[imputing by drawing from conditional distribution without bootstrap]/*
*/2464 missing observations on lganorimi imputed from 6466 complete /*
*/observations.

```

**c) JUBILADOS Y PENSIONADOS**

```

uvis reg ljuborimi anoseduc exp exp2 sexo if conduct==4, gen(ljubiuvis) /*
*/boot [imputing by drawing from conditional distribution with bootstrap] /*
*/599 missing observations on ljuborimi imputed from 6984 complete /*
*/observations.

```

Fuente: Elaboración propia de los autores.

### CUADRO A.11 IMPUTACIÓN MÚLTIPLE

El número de simulaciones requeridas (cinco en este caso) se especifican en la opción *m(#)*, y los datos imputados se guardan en los archivos *imputmviss1*, ..., *imputmviss5*.

**a) EMPLEADOS (SUELDOS Y SALARIOS)**

```

mvis lsysorimi anoseduc exp exp2 sexo if ramar1>0 & ramar1<99 & /*
*/categ==3 & conduct==1 & anoseduc<99 using imputmviss, m(5) boot /*
*/(lsysorimi) cmd(reg lsysorimi anoseduc exp exp2 sexo) imputing /*
*/1..2..3..4..5..file imputmviss.dta saved

```

Toda vez que se aplicó la sintaxis anterior, es necesario combinar las imputaciones en un solo archivo de trabajo. Para lograr este objetivo, STATA dispone del comando *micombine* que se utiliza para mezclar archivos y generar estimadores combinados de los cinco modelos de regresión ajustados.

Los archivos se guardan en el directorio en donde el usuario resguarda en forma regular sus datos STATA, y en caso de que no se recuerde es posible utilizar el comando *cd* para conocer la ruta que el programa utilizó para guardar los datos con las simulaciones efectuadas. Para aplicar el comando *micombine*, es necesario cambiarse al directorio en donde se guardaron los archivos imputados ya que el comando utiliza los archivos para generar los estimadores combinados.

El valor promedio de los parámetros se muestran en la columna *coef.*, y se utilizan para imputar las observaciones con valores omitidos. Al final de tabla se informa del número total de observaciones que fueron utilizadas para la estimación del modelo combinado.

```

micombine reg lsysorimi anoseduc exp exp2 sexo if ramar1>0 & ramar1<99
*/& categ==3 & conduct==1 & anoseduc<99, gen(lsysimvis) impid(_j)/*
*/Multiple imputation parameter estimates (5 imputations)

```

lsysorimi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
anoseduc	.0877942	.0013713	64.02	0.000	.0851063 .0904821
exp	.0517455	.0013448	38.48	0.000	.0491096 .0543814
exp2	-.0007439	.0000256	-29.07	0.000	-.0007941 -.0006938
sexo	.4735554	.0104083	45.50	0.000	.4531544 .4939564
_cons	4.32035	.0233638	184.92	0.000	4.274555 4.366144

27349 observations.

## b) CUENTA PROPIA Y PATRONES (GANANCIAS)

```
mvis lganorimi anoseduc exp exp2 sexo if ramar1>0 & ramar1<99 & /*
*/(categ==1 | categ==2 & conduct==1 & anoseduc<99 using imputmvisg, m(5) /*
*/boot(lganorimi) cmd(reg lganorimi anoseduc exp exp2 sexo) imputing /*
*/1..2..3..4..5..file imputmvisg.dta saved micombine reg lganorimi /*
*/anoseduc exp exp2 sexo if ramar1>0 & ramar1<99 & (categ==1 | categ==2 & /*
*/conduct==1 & anoseduc<99, gen(lganimvis) impid(_j) /*

*/Multiple imputation parameter estimates (5 imputations)
```

lganorimi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
anoseduc	.1044805	.0031216	33.47	0.000	.0983611 .1105999
exp	.070817	.0029178	24.27	0.000	.0650972 .0765368
exp2	-.0010402	.0000487	-21.37	0.000	-.0011356 -.0009448
sexo	.5020693	.0257268	19.52	0.000	.4516363 .5525024
_cons	3.530786	.0575714	61.33	0.000	3.417927 3.643645

8930 observations.

## c) JUBILADOS Y PENSIONADOS

```
mvis ljuborimi exp exp2 anoseduc sexo if conduct==4 using imputmvisj, m(5) /*
*/boot(ljuborimi) cmd(reg ljuborimi exp exp2 anoseduc sexo) imputing /*
*/1..2..3..4..5..file imputmvisj.dta saved micombine reg ljuborimi exp /*
*/exp2 anoseduc sexo if conduct==4, gen(jubimvis) impid(_j) mvis /*
*/ljuborimi exp exp2 anoseduc sexo if conduct==4 using imputmvisj, m(5)

*/Multiple imputation parameter estimates (5 imputations)
```

ljuborimi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
exp	.0317077	.0022286	14.23	0.000	.027339 .0360763
exp2	-.0003	.0000227	-13.25	0.000	-.0003444 -.0002556
anoseduc	.0487992	.0018868	25.86	0.000	.0451006 .0524979
sexo	.2473725	.0143167	17.28	0.000	.2193075 .2754376
_cons	4.615811	.0588244	78.47	0.000	4.500498 4.731125

7583 observations.

### CUADRO A.12 IMPUTACIÓN MÚLTIPLE (ICE)

Este procedimiento funciona con la misma lógica que el comando MVIS, y su aplicación se realiza utilizando en forma reiterativa el comando UVIS. Es decir, el número de simulaciones que se generan (cinco en este caso) con la opción m(#) se ejecutan con el comando UVIS.

Los archivos se guardan en el directorio en donde el usuario tiene los datos STATA, y al igual que en los casos en que se generan varios archivos de trabajo, en caso de que no se recuerde se puede utilizar el comando *cd* para conocer la ruta que el programa utilizó para guardar los datos con las simulaciones efectuadas.

**a) EMPLEADOS (SUELDOS Y SALARIOS)**

```
ice lsysorimi anoseduc sexo exp exp2 if ramar1>0 & ramar1<99 & categ==3 & /*
  */conduct==1 & anoseduc<99 using imputices1, genmiss(imput) id(sort) /*
  */ m(5) cmd(reg lsysorimi anoseduc sexo exp exp2) boot(lsysorimi) /*
  */cycles(5) match(lsysorimi)
```

#missing values	Freq.	Percent	Cum.
0	23,301	24.59	24.59
1	4,048	4.27	28.86
.	67,423	71.14	100.00
Total	94,772	100.00	

Variable	Command	Prediction equation
lsysorimi	reg lsysorimi anoseduc sexo exp exp2	anoseduc sexo exp exp2 <sup>a</sup>
anoseduc	reg lsysorimi anoseduc sexo exp exp2	<sup>a</sup>
sexo	reg lsysorimi anoseduc sexo exp exp2	<sup>a</sup>
exp	reg lsysorimi anoseduc sexo exp exp2	<sup>a</sup>
exp2	reg lsysorimi anoseduc sexo exp exp2	<sup>a</sup>

Imputing  
 [Only 1 variable to be imputed, therefore no cycling needed.]  
 1..2..3..4..5..file imputices1.dta saved

En esta opción también es necesario combinar los archivos con las simulaciones generadas, con el propósito de disponer de un vector combinado de parámetros estimados.

```
micombine reg lsysorimi anoseduc sexo exp exp2 if ramar1>0 & ramar1<99 & /*
  */categ==3 & > conduct==1 & anoseduc<99, gen(lsysiice) impid(_j) /*
  */Multiple imputation parameter estimates (5 imputations)
```

lsysorimi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
anoseduc	.073122	.0015208	48.08	0.000	.0701411 .0761029
sexo	.4116381	.0105156	39.15	0.000	.391027 .4322492
exp	.0422761	.0013385	31.59	0.000	.0396527 .0448996
exp2	-.0006129	.0000256	-23.94	0.000	-.0006631 -.0005627
_cons	4.611505	.0263845	174.78	0.000	4.55979 4.66322

27349 observations.

**b) CUENTA PROPIA Y PATRONES (GANANCIAS)**

```
ice lganorimi anoseduc exp exp2 sexo if ramar1>0 & ramar1<99 & /*
  */(categ==1 | categ==2 & conduct==1 & anoseduc<99 using imputiceg1, /*
  */genmiss(imput) id(sort) m(5) cmd(reg lganorimi anoseduc exp exp2 sexo) /*
  */boot(lganorimi) cycles(5) match(lganorimi)
```

#missing values	Freq.	Percent	Cum.
0	6,466	6.82	6.82
1	2,464	2.60	9.42
.	85,842	90.58	100.00
Total	94,772	100.00	

Variable	Command	Prediction equation
lganorimi	reg lganorimi anoseduc exp exp2 sexo	anoseduc exp exp2 sexo <sup>a</sup>
anoseduc	reg lganorimi anoseduc exp exp2 sexo	<sup>a</sup>
exp	reg lganorimi anoseduc exp exp2 sexo	<sup>a</sup>
exp2	reg lganorimi anoseduc exp exp2 sexo	<sup>a</sup>
sexo	reg lganorimi anoseduc exp exp2 sexo	<sup>a</sup>

Imputing  
 [Only 1 variable to be imputed, therefore no cycling needed.]  
 1..2..3..4..5..file imputiceg1.dta saved



```
micombine reg lganorimi anoseduc exp exp2 sexo if ramar1>0 & ramar1<99 & /*
*/(categ==1 | categ==2 & conduct==1 & anoseduc<99, gen(lganiice) impid(_j) /*
*/(Multiple imputation parameter estimates (5 imputations)
```

lganorimi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
anoseduc	.0729901	.0033835	21.57	0.000	.0663576	.0796226
exp	.0497697	.0027297	18.23	0.000	.0444189	.0551205
exp2	-.0007243	.000045	-16.08	0.000	-.0008126	-.0006361
sexo	.3642231	.0299028	12.18	0.000	.3056068	.4228394
_cons	4.190009	.0591418	70.85	0.000	4.074078	4.305941

8930 observations.

**c) JUBILADOS Y PENSIONADOS**

```
ice ljuborimi exp exp2 anoseduc sexo if conduct==4 using imputicej1, /*
*/genmiss(imput) id(sort)> m(5) cmd(reg ljuborimi exp exp2 anoseduc sexo) /*
*/boot(ljuborimi) cycles(5) > match(ljuborimi)
```

#missing values	Freq.	Percent	Cum.
0	6,984	7.37	7.37
1	599	0.63	8.00
	87,189	92.00	100.00
Total	94,772	100.00	

Variable	Command	Prediction equation
ljuborimi	reg ljuborimi exp exp2 anoseduc sexo	exp exp2 anoseduc sexo
exp	reg ljuborimi exp exp2 anoseduc sexo	a
exp2	reg ljuborimi exp exp2 anoseduc sexo	a
anoseduc	reg ljuborimi exp exp2 anoseduc sexo	a
sexo	reg ljuborimi exp exp2 anoseduc sexo	a

Imputing  
 [Only 1 variable to be imputed, therefore no cycling needed.]  
 1..2..3..4..5..file imputicej1.dta saved

```
micombine reg ljuborimi exp exp2 anoseduc sexo if conduct==4, /*
*/gen(ljubiice) impid(_j) /*
*/micombine reg ljuborimi exp exp2 anoseduc sexo if conduct==4, /*
*/gen(ljubiice) impid(_j) /*
*/Multiple imputation parameter estimates (5 imputations)
```

ljuborimi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exp	.0301315	.0022611	13.33	0.000	.025699	.0345639
exp2	-.0002842	.000023	-12.36	0.000	-.0003292	-.0002391
anoseduc	.0444216	.0019091	23.27	0.000	.0406791	.0481641
sexo	.2286324	.0140597	16.26	0.000	.2010715	.2561934
_cons	4.690075	.0601613	77.96	0.000	4.572142	4.808008

7583 observations.





NACIONES UNIDAS

Serie

C E P A L

estudios estadísticos y prospectivos

## Números publicados

El listado completo de esta colección, así como las versiones electrónicas en pdf están disponibles en nuestro sitio web: [www.cep.org/publicaciones](http://www.cep.org/publicaciones)

54. Imputación de datos: teoría y práctica, Fernando Medina y Marco Galván (LC/L.2772-P), N° de venta S.07.II. G.109, (US\$ 10.00), julio, 2007.
53. Indicadores de los objetivos de desarrollo del Milenio en América Latina y el Caribe: una comparación entre datos nacionales e internacionales, Simone Cecchini e Irene Azócar (LC/L.2767-P), N° de venta S.07.II. G.103, (US\$ 10.00), julio, 2007.
52. Transversalizando la perspectiva de género en los objetivos de desarrollo del milenio, Daniela Zapata (LC/L.2764-P), N° de venta S.07.II. G.100, (US\$ 10.00), junio 2007.
51. Un sistema de indicadores líderes compuestos para la región de América Latina, Mauricio Gallardo y Michael Pedersen (LC/L.2728-P), N° de venta S.07.II. G.66, (US\$ 10.00), mayo, 2007.
50. Propuesta regional de indicadores complementarios al Objetivo de Desarrollo del Milenio 7: “Garantizar la sostenibilidad del medio ambiente”, Rayén Quiroga Martínez, (LC/L.2746-P), N° de venta S.07.II. G.84, (US\$ 10.00), mayo, 2007.
49. Indicadores líderes compuestos. Resumen de metodologías de referencia para construir un indicador regional en América Latina, Mauricio Gallardo y Michael Pedersen (LC/L.2707-P), N° de venta S.07.II.G.55, (US\$ 10.00), abril, 2007.
48. The millennium development goals: opportunities and challenges for national statistical systems in Latina America and the Caribbean, (LC/L.2673-P), N° de venta E.07.II.G.40, (US\$ 10.00), March, 2007.
47. El consumo aparente de energía fósil en los países latinoamericanos hacia 1925: una propuesta metodológica a partir de las estadísticas de comercio exterior, Mauricio Folchi y María del Mar Rubio (LC/L.2658-P), N° de venta S.07.II.G.9, (US\$ 10.00), enero, 2007
46. El método DEA y su aplicación al estudio del sector energético y las emisiones de CO<sub>2</sub> en América Latina y el Caribe, Andrés Schuschny (LC/L.2657-P), N° de venta S.07.II.G.8, (US\$ 10.00), enero, 2007.
45. Can Latin America Fly? Revising its engines of growth, Hubert Escaith (LC/L.2605-P), N° de venta E.06.II.G.125, (US\$ 10.00), September, 2006.
44. Importaciones y modernización económica en América Latina durante la primera mitad del siglo XX. Las claves de un programa de investigación, Albert Carreras, Mauricio Folchi, André Hofman, Mar Rubio, Xavier Tafunell y César Yáñez (LC/L.2583-P), N° venta S.06.II.G.113, (US\$ 10.00), septiembre, 2006.
43. La medición de los Objetivos de Desarrollo del Milenio en las áreas urbanas de América Latina, Simone Cecchini, Jorge Rodríguez y Daniela Simioni (LC/L.2537-P), N° de venta S.06.II.G.64, (US\$ 10.00), junio, 2006.
42. Latin America and the Caribbean. Projections 2006-2007. Economic Projections Centre, (LC/L.2528-P), Sales Number E.06.II.G.55, (US\$ 10.00), June, 2006.
42. América Latina y el Caribe: proyecciones 2006-2007, Centro de Proyecciones Económicas (LC/L.2528-P), N° venta S.06.II.G.55, (US\$ 10.00), abril, 2006.
41. Propuesta para un compendio Latinoamericano de indicadores sociales, Unidad de Estadísticas Sociales, (LC/L.2471-P), N° de venta S.06.II.G.15, (US\$ 10.00), diciembre 2005.
40. Oportunidades digitales, equidad y pobreza en América Latina: ¿Qué podemos aprender de la evidencia empírica? Simone Cecchini, (LC/L.2459-P), N° de venta S.05.II.G.206, (US\$ 10.00), diciembre 2005.
39. El seguimiento de los objetivos de desarrollo del milenio: oportunidades y retos para los Sistemas Nacionales de Estadística, José L. Cervera Ferri, (LC/L.2458-P), N° de venta S.05.II.G.204, (US\$ 10.00), diciembre, 2005
38. Elementos teóricos del ajuste estacional de series económicas utilizando X-12-ARIMA y TRAMO-SEATS, Francisco G. Villarreal (LC/L.2457-P), N° de venta S.05.II.G.203, (US\$ 10.00), diciembre 2005.
37. Tópicos sobre el Modelo de Insumo-Producto: teoría y aplicaciones, Andrés Ricardo Schuschny, (LC/L.2444-P), N° de venta S.05.II.G.191, (US\$ 10.00), diciembre 2005.
36. Demanda de exportaciones e importaciones de bienes y servicios para Argentina y Chile, Claudio Aravena, (LC/L.2434-P), N° de venta S.05.II.G.180, (US\$ 10.00), diciembre de 2005.

35. Propuesta metodológica para el desarrollo y la elaboración de estadísticas ambientales en países de América Latina y el Caribe, Dharmo Rojas, (LC/L.2398-P), N° de venta S.05.II.G.143, (US\$ 10.00), octubre, 2005.
34. Indicadores sociales en América Latina y el Caribe, Simone Cecchini, (LC/L.2383-P), N° de venta S.05.II.G.127, (US\$ 10.00), septiembre, 2005.
33. El acuerdo de libre comercio Mercosur-Comunidad Andina de Naciones: una evaluación cuantitativa, Daniel Berrettoni y Martín Cicowicz (LC/L.2310-P), N de venta S.05.II.G.59, (US\$ 10.00), abril, 2005.
32. América Latina y el Caribe: proyecciones 2005, Centro de Proyecciones Económicas (CPE), (LC/L.2297-P), N° venta S.05.II.G.45, (US\$ 10.00), abril, 2005.
31. Metodología de proyecciones económicas para América Latina: formulación de proyecciones de corto plazo a partir de la base de datos de coyuntura, Centro de Proyecciones Económicas, (LC/L.2296-P), N° venta S.05.II.G.44, (US\$ 10.00), abril, 2005.
30. Cuentas ambientales: conceptos, metodologías y avances en los países de América Latina y el Caribe, Farid Isa, Marcelo Ortúzar y Rayén Quiroga, (LC/L.2229-P), N° de venta: S.04.II.G.151, (US\$ 10.00), enero, 2005.
29. Crecimiento económico, creación y erosión de empleo: un análisis intersectorial, Gabriel Gutiérrez (LC/L.2199-P), N° venta S.04.II.G.125, (US\$ 10.00), octubre, 2004.
28. Un enfoque contable y estructural al crecimiento y la acumulación en Brasil y México, (1983-2000), (LC/L.2188-P), N° venta S.04.II.G.116, (US\$ 10.00), diciembre, 2004.
27. Proyecciones de América Latina y el Caribe, 2004, Centro de Proyecciones Económicas (LC/L.2144-P), N° venta S.04.II.G.72, (US\$ 10.00), mayo, 2004.
26. Estados Unidos: ¿Una nueva economía, o más de lo mismo?, Gunilla Ryd (LC/L.2043-P), N° venta S.03.II.G.202, (US\$ 10.00), diciembre, 2003.
25. Potential output in Latin America: a standard approach for the 1950-2002 period, André A. Hofman, Heriberto Tapia, (LC/L.-2042P), Sales Number E.03.II.G.205, (US\$ 10.00), December, 2003.
24. El desarrollo económico de América Latina en épocas de globalización-una agenda de investigación, Albert Carreras, André A. Hofman, Xavier Tafunell y César Yáñez, (LC/L.2033-P), N° venta S.03.II.G.197, (US\$ 10.00), diciembre, 2003.
23. Tendencias y extrapolación del crecimiento en América Latina y el Caribe, Hubert Escaith, (LC/L.2031-P), N° venta S.03.II.G.193, (US\$ 10.00), diciembre, 2003.
22. Apertura y cambio estructural de la economía brasileña, Alejandro Vargas, (LC/L.2024-P), N° venta S.03.II.G.188, (US\$ 10.00), diciembre, 2003.
21. Registros Administrativos, calidad de los datos y credibilidad pública: presentación y debate de los temas sustantivos de la segunda reunión de la Conferencia Estadística de las Américas de la CEPAL, Graciela Echegoyen (comp), (LC/L.2007-P), N° venta S.03.II.G.168, (US\$ 10.00), diciembre, 2003.
20. Reseña de programas sociales para la superación de la pobreza en América Latina, Marcia Pardo (LC/L.1906-P), N° venta S.03.II.G.64, (US\$ 10.00), octubre, 2003.
19. Proyecciones de América Latina y el Caribe, 2003, Centro de Proyecciones Económicas (CPE), (LC/L.1886-P), N° venta S.03.II.G.52, (US\$ 10.00), abril, 2003.
18. Países industrializados: un análisis comparativo de las proyecciones 2002-2003, Gunilla Ryd (LC/L.1868-P), N° venta S.03.II.G.39, (US\$ 10.00), marzo, 2003.
17. Países industrializados: resumen de las proyecciones 2001-2002, Gunilla Ryd (LC/L.1702-P), N° venta S.02.II.G.13, (US\$ 10.00), febrero, 2002.
16. Proyecciones latinoamericanas 2001-2002, Alfredo Calcagno, Sandra Manuelito y Gunilla Ryd (LC/L.1688-P), N° venta: S.02.II.G.3, (US\$ 10.00), enero, 2002.
15. La convertibilidad argentina: ¿un antecedente relevante para la dolarización de Ecuador?, Alfredo Calcagno y Sandra Manuelito (LC/L.1559-P), N° venta S.01.II.G.104, (US\$ 10.00), junio, 2001.
15. Argentine convertibility: is it a relevant precedent for the dollarization process in Ecuador, Alfredo Calcagno, Sandra Manuelito (LC/L.1559-P) N° venta E.01.II.G.104, (US\$ 10.00) July, 2001.

- El lector interesado en adquirir números anteriores de esta serie puede solicitarlos dirigiendo su correspondencia a la Unidad de Distribución, CEPAL, Casilla 179-D, Santiago, Chile, Fax (562) 210 2069, correo electrónico: [publications@cepal.org](mailto:publications@cepal.org).

Name:.....
Activity: .....
Address: .....
Postal code, city, country: .....
Tel.:..... Fax: ..... E.mail:.....