UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
FACPYA
FACULTAD DE CONTADURÍA PÚBLICA Y ADMINISTRACIÓN
VinculaTégica EFAN

Imputación de datos faltantes en las variables de ingreso en la Encuesta Nacional sobre Salud y Envejecimiento en México.

(Imputation of missing data in the income variables in the National Survey on Health and Aging in Mexico)

Guillermo Andrés Villagra Fuentes[*1]

[1] *Universidad Autónoma de Nuevo León – Facultad de Contaduría Pública y Administración (México), guillermo.villagrafn@uanl.edu.mx*

* *Autor de Contacto*

Cómo citar: Villagra-Fuentes, G. A. Imputación de datos faltantes en las variables de ingreso en la Encuesta Nacional sobre Salud y Envejecimiento en México. *Vinculatégica EFAN*, *11*(3), 141–161. https://doi.org/10.29105/vtga11.3-1099

Información revisada por arbitraje tipo doble par ciego.

Recibido: 9 de mayo 2024

Aceptado: 20 de mayo 2024

Publicado: 30 de mayo 2025

Resumen

La presencia de datos faltantes, conocidos como Missing Values o missing data, es una situación habitual a la que se enfrentan tanto investigadores como tomadores de decisiones. Este estudio no es una excepción, ya que se basa en la Encuesta Nacional sobre Salud y Envejecimiento en México (ENASEM), la cual es longitudinal y está dirigida a personas mayores de 50 años, lo que hace que la presencia de valores faltantes sea evidente. Para este estudio en particular, se ha decidido abordar especialmente los valores faltantes en tres áreas principales: variables de ingreso, gasto y activos. La propuesta consiste en emplear el método de imputaciones múltiples bajo el supuesto de Missing at Random (MAR). De un total de variables faltantes de 28,892, se logró imputar el 100% de estas. Se observó que la mayor concentración de valores faltantes se encontraba en la ronda 2001, disminuyendo en las rondas siguientes. En cuanto a las secciones de la encuesta, se encontró que la que presentaba la mayor cantidad de valores faltantes, los cuales fueron imputados, era la de activos, con un 67%, seguida por la de ingresos con un 19% y la de gastos con un 13%.

Palabras clave: *imputación múltiple, personas mayores, ingresos.*
Códigos JEL: *C15, I31, I32*

Abstract

The presence of missing data, also known as Missing Values or missing data, is a common situation faced by both researchers and decision-makers. This study is no exception, as it is based on the National Survey on Health and Aging in Mexico (ENASEM), which is longitudinal and targeted at individuals over 50 years old, making the presence of missing values evident. For this particular study, special attention has been given to missing values in three main areas: income, expenditure, and assets variables. The proposal involves employing the method of multiple imputations under the assumption of Missing at Random (MAR). Out of a total of 28,892 missing variables, 100% of these were successfully imputed. It was observed that the highest concentration of missing values was found in the 2001 round, decreasing in subsequent rounds. Regarding the survey sections, it was found that the one with the highest percentage of missing values, which were imputed, was the assets section, with 67%, followed by the income section with 19%, and the expenditure section with 13%.

Key words: *Multiple imputation, elderly individuals, income.*
JEL Codes: *C15, I31, I32*

Introducción

Los investigadores y los tomadores de decisiones se enfrentan constantemente a la presencia de datos faltantes. En cualquier base de datos, es común encontrar información incompleta, conocida como *Missing Values o missing data*, que consiste en un conjunto de valores faltantes por diversas razones desconocidas. En las encuestas a hogares o individuos directamente, la falta de respuesta puede atribuirse a causas como la fatiga del encuestado, el desconocimiento de la información solicitada, el rechazo a proporcionar información sobre temas sensibles, la negativa a participar en la investigación o problemas relacionados con el marco de muestreo.

En general, pueden clasificarse en dos grandes grupos: los datos faltantes por unidad (cuando el entrevistado no responde) y otros en los que los ítems del cuestionario no son comprendidos o están mal formulados, lo que también puede llevar a que no se respondan. En encuestas longitudinales, la ausencia de datos puede deberse a la falta de localización de la persona o a que las unidades abandonen el estudio, lo que se conoce como "muerte de datos".

La Encuesta Nacional sobre Salud y Envejecimiento en México (ENASEM) es un estudio longitudinal que se inició en el año 2001, con entrevistas de seguimiento realizadas en 2003, 2012, 2015, 2018 y 2021. Desde su primer levantamiento hace más de 20 años, una de las consideraciones más importantes ha sido la continuidad de las personas de la muestra original, al mismo tiempo que ha agregado nuevas muestras en 2012 y 2018. La ENASEM se destaca como una fuente integral de información sobre diversos aspectos de la vida de mujeres y hombres de 50 años o más, abarcando áreas sociodemográficas, económicas, de salud física y mental, estilo de vida y uso del tiempo.

Durante las últimas décadas, se han desarrollado algoritmos de imputación múltiple que poseen mejores propiedades estadísticas que las opciones tradicionales de eliminación de datos (*listwise*), emparejamiento de observaciones (*pairwise*), método de medias o el *hot-deck*. Estos algoritmos pueden aplicarse utilizando paquetes estadísticos. Más adelante se desarrolló el método de Máxima Verosimilitud con Información Completa (FIML) (Arbuckle, 1996). Aunque no se considera estrictamente una técnica de imputación, FIML agrupa distintos conjuntos de valores o patrones donde la información está parcialmente completa. Este método construye "parches" para completar los datos faltantes utilizando la información completa de otras rondas de la encuesta. Algunos investigadores sugieren que FIML puede reducir los sesgos y la variabilidad en mayor medida que los métodos de imputación de datos faltantes de Rubín (Enders, 2001).

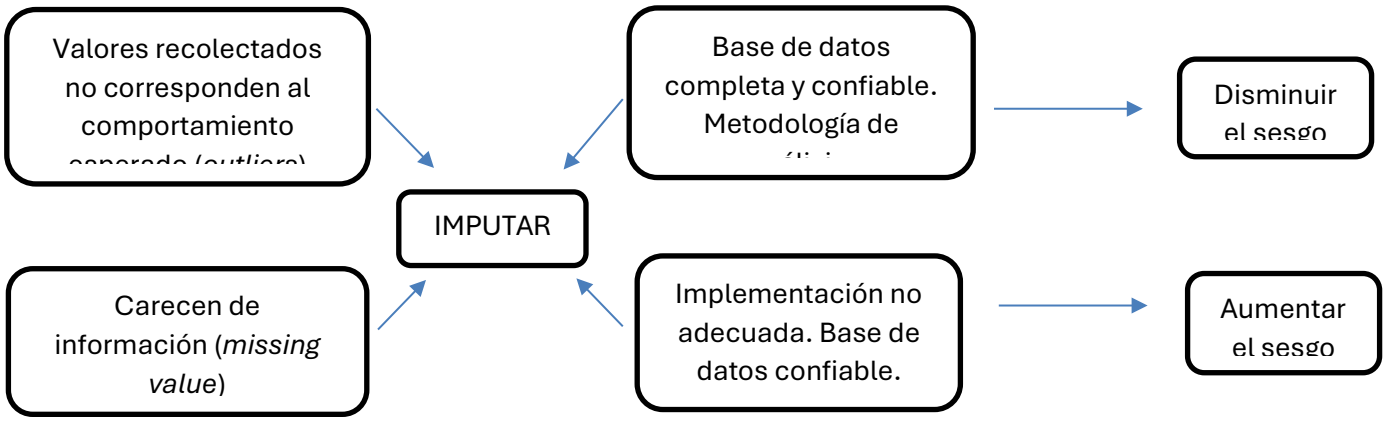
El método de imputación múltiple propuesto en este estudio es el Patrón de Pérdida de Datos Faltantes, conocido como MAR. Aunque esta técnica es la misma que la implementada por ENASEM, la diferencia más significativa radica en su aplicación práctica. Mientras que ENASEM

empleó un método de regresión secuencial con un sistema basado en SAS, utilizando la rutina del software IVEware desarrollado en la Universidad de Michigan (Raghunathan, 2000; Raghunathan, 2001), la información detallada sobre los pasos seguidos es general. Se limitan a describir la técnica utilizada y el software empleado, sin ofrecer un análisis detallado del proceso de imputación.

Método

Imputar datos implica reemplazar observaciones, ya sea porque algunos valores recolectados no concuerdan con el comportamiento esperado (*outliers*) o porque carecen de información (*missing value*). La ventaja de la imputación radica en la obtención de una base de datos completa, lo que permite llevar a cabo un análisis metodológico y reducir así el sesgo (Figura 1). Sin embargo, si la técnica o su implementación no son adecuadas, es posible que el sesgo aumente, lo que resultaría en una base de datos no confiable.

Figura 1. *¿Por qué se tiene que imputar?*



Participantes

Se comenzó a desarrollar la imputación especialmente en tres áreas: variables de ingreso, gasto y activos. El estudio se llevó a cabo utilizando la base de datos maestra que comprende las cuatro rondas consecutivas. La ronda de 2001 consta de 9862 observaciones, de las cuales solo se consideraron las principales. Además, se eliminaron las encuestas completadas por terceros sobre personas fallecidas y nuevas muestras, ya que esto genera un problema de corte en los datos. Se restaron 227 observaciones debido a la presencia de valores faltantes en las cuatro rondas, 24 debido a la falta de

información sobre el estado civil y 56 porque no se informó la localidad, lo que resultó en un total de 9555 observaciones en 2001, 8584 en 2003, 5909 en 2012 y 5323 en 2015 (Tabla 1).

Tabla 1 *Observaciones Trabajadas por Ronda*

Rondas ENASEM	Observaciones
2001	9555
2003	8584
2012	5909
2015	5323

Técnica e Instrumento

En la década de los 70, la imputación de datos implicaba identificar y reemplazar registros de información. En 1976, Rubin propuso un marco conceptual para el análisis de datos faltantes basado en métodos de inferencia estadística. Posteriormente, se desarrollaron algoritmos Expectation Maximization (EM) con estimadores robustos a partir del método de máxima verosimilitud (Dempster, Laird y Rubin, 1977), en los cuales los datos faltantes se consideran variables aleatorias y se imputan sin ajustar el modelo. Rubin (1987) introdujo el concepto de imputación múltiple, aplicando simulaciones de Monte Carlo con m > 1 simulaciones.

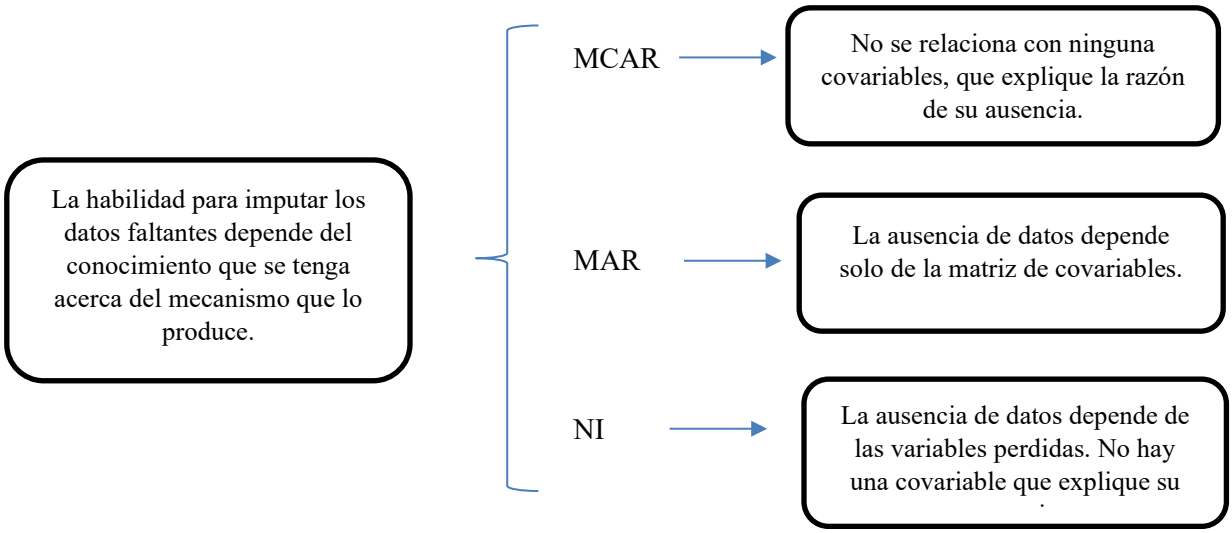
Al iniciar el proceso de imputación, es crucial considerar el patrón de pérdida de los datos faltantes, ya que esto puede influir en el método de imputación. Este patrón puede ser completamente aleatorio (MCAR, Missing Completely At Random), aleatorio (MAR, Missing At Random), o No Ignorable (NI) (Rubin, 1987). MCAR se refiere a cuando la ausencia de información no está relacionada con ninguna variable presente en la matriz de datos, incluidos los valores faltantes, y no está relacionada ni con los datos observados ni con los ausentes. En el caso de MAR, la ausencia de datos no depende de los valores faltantes en sí, sino que puede explicarse a partir de la matriz de datos y las covariables. Las covariables son variables que ayudan a predecir el resultado de un estudio y son esenciales en la imputación múltiple para proporcionar información necesaria a la variable a imputar. En nuestro estudio, consideramos covariables con una alta correlación con las variables y baja correlación entre sí. En cuanto a NI, la ausencia de datos depende de la variable perdida (Figura 2).

En la década de los 70, la imputación de datos significaba identificar y sustituir los registros de información. En 1976 Rubin propuso un marco conceptual para el análisis de datos faltantes sustentando en métodos de inferencia estadística. Posteriormente, se permitió generar algoritmos

Expectation Maximization (EM), con estimadores robustos a partir del método de máxima verosimilitud (Dempster, Laird, y Rubin, 1977), en donde los datos faltantes son tomados como variables aleatorias y los datos imputados se generan sin necesidad de ajustar el modelo. Más tarde Rubin 1987, introdujo el concepto de imputación múltiple, donde se aplican simulaciones de Monte Carlo a partir m>1 simulaciones.

Uno de los puntos a considerar al comenzar a imputar es el patrón de pérdida de los datos faltantes, ya que esto puede influir en el método de imputación. Este patrón puede darse de manera completamente aleatoria (MCAR, *Missing Completely At Random*), de manera aleatoria (MAR, *Missing At Random*) o No Ignorable (NI) (Rubin, 1987). MCAR se refiere a cuando la ausencia de información no está relacionada con ninguna variable presente en la matriz de datos, incluidos los valores faltantes, y no está relacionada ni con los datos observados ni con los ausentes. En el caso de MAR, la ausencia de datos no depende de los valores faltantes en sí, sino que puede explicarse a partir de la matriz de datos y las covariables (Las covariables son variables que ayudan a predecir el resultado de un estudio. En la imputación múltiple es esencial para otorgar la información necesaria a la variable a imputar. En nuestro estudio consideramos covariables que tengan una correlación alta con las variables y pequeña entre sí), es decir, los valores faltantes pueden recuperarse a partir de los datos observables. En cuanto a NI, la ausencia de datos depende de la variable perdida (Figura 2).

Figura 2. *Patrones de prueba de datos faltantes*



La imputación múltiple es una técnica estadística utilizada para analizar bases de datos

incompletas, es decir, aquellas en las que algunas entradas tienen valores faltantes. Esta técnica consiste en reemplazar cada uno de los valores faltantes por dos o más valores posibles, idea propuesta por Rubín en 1977. El proceso de imputación múltiple implica tres pasos:

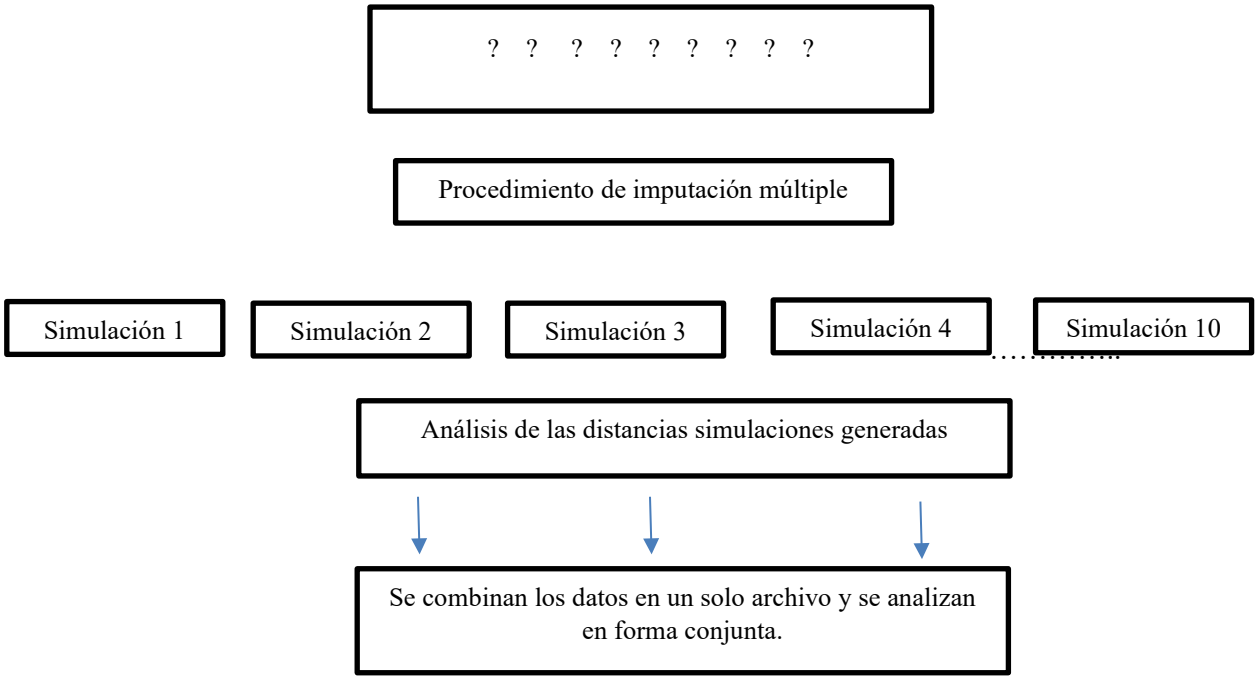
Imputación: Se generan m valores para cada entrada faltante a partir de una distribución, lo que resulta en la creación de m bases de datos completas.

Análisis: Cada conjunto de bases de datos completas se analiza de forma separada, lo que implica realizar m análisis.

Resultados: Los resultados de cada análisis se integran en un resultado final utilizando alguna regla preestablecida para combinar los m resultados.

La propuesta inicial de Rubín no abordaba la formulación para calcular las estimaciones combinadas. Sin embargo, en 1987, partiendo de los tres pasos mencionados, propuso una solución que empleaba métodos de simulación de Monte Carlo y sustituía los datos faltantes mediante un número de simulaciones que, según el autor, oscilaban entre 3 y 10. (Figura 3).

Figura 3. Imputación Múltiple



Nota: Elaboración propia en base a los autores.

A esta metodología que establece las fórmulas obtenidas a partir de las combinaciones de las

estimaciones de cada imputación, resumiendo así todos los modelos ajustados en uno solo, se le llama las Reglas de Rubín. Estas reglas constituyen el procedimiento teórico desarrollado por Rubín para explicar el proceso de imputación múltiple, en el cual propone que los datos omitidos sean reemplazados por múltiples realizaciones aleatorias. Este proceso se conoce como Imputación Múltiple y su sustento teórico se encuentra en la estadística bayesiana, la cual utiliza la información de la muestra para realizar inferencias respecto de los parámetros. La idea principal detrás de la imputación múltiple es que utiliza toda la información disponible, en contraposición a métodos que descartan los datos con valores faltantes, ya que al final esta información puede ser muy útil para el análisis posterior.

Según como se explica en Vargas y Valdés 2018, la imputación múltiple de Rubín ocupa simulaciones de Monte Carlo. Consiste en aumentar los datos faltantes, que contiene el paso 1 y 2. En el paso 1 se obtiene una muestra aleatoria de las observaciones a partir de las distribuciones marginal en la primera iteración:

$$Y_{falt}^{(t+1)} \sim p\left(Y_{falt} | Y_{obs}, \theta^{(t)}\right) \qquad (1)$$

En el paso 2, se obtiene una muestra aleatoria de parámetros de la distribución marginal que incorpora los valores observados e iniciales de los valores faltantes en el paso 1, en la primera iteración:

$$\theta^{(t+1)} \sim p\left(\theta | Y_{obs}, Y_{falt}^{(t+2)}\right) \qquad (2)$$

El paso 1 y 2 de la primera iteración proporciona los valores iniciales $\{Y_{falt,}^{(0)}, \theta^{(0)}\}$, así para todas las iteraciones posteriores, para crear una cadena de Markov con valores $\{Y_{falt,}^{(1)}, \theta^{(1)}; Y_{falt,}^{(2)}, \theta^{(2)}; \ldots.\}$ que convergen en distribución a $p(\theta, Yfalt|Yobs)$, con la finalidad de crear una cadena de observaciones completas $\{Y_{falt,}^{(1)}, Y_{falt,}^{(2)}, \ldots. Y_{falt,}^{(mt)}\}$, la cual es equivalente de correr m cadenas independientes.

Un aspecto importante que depende de la cantidad de datos faltantes es probar la

convergencia del proceso de cadenas múltiples de Monte Carlo (MCMC) y el número de iteraciones que se requieren. Se han propuesto varios métodos de convergencia de la distribución conjunta para determinar un valor especifico de m (Ritter y Tanner, 1992; Roberts, 1992). Según Vargas y Valdés 2018, desde un punto vista práctico, se puede ocupar la función de Autocorrelación (ACF) para determinar la convergencia del algoritmo para cada rezago-p de la serie estacional $\{K^{(t)}: t = 1,2, \dots, k\}$; la ACF se define como:

$$\rho_\rho = \frac{Cov(k^{(t)}, k^{(t+p)})}{V(k^{(t)})} \qquad (3)$$

Se gráfica la función de autocorrelación para un valor finito (p) de la muestra en un correlograma, donde se evidencia la posible dependencia lineal entre las iteraciones. Un decaimiento brusco para los valores de p entre 2 y 4 sugiere una independencia serial, lo que indica que el algoritmo converge hacia una solución satisfactoria (Box y Jenkins, 2015). Se utiliza la peor función lineal (WLF) y su función de autocorrelación (ACF) para dos propósitos: primero, verificar que la WLF genere ruido blanco y, segundo, asegurarse de que la ACF de la WLF muestre una disminución en la autocorrelación para valores de p mayores a 4, lo que garantiza una convergencia satisfactoria del proceso.

Procedimiento

Para generar las variables sujetas a imputación, se consideraron tres áreas principales de estudio: ingreso, gasto y activos. Estas variables se derivaron de una base de datos completamente pegada en forma longitudinal, que posteriormente se segmentó por ronda para abordar la imputación de manera independiente. Se seleccionaron 25 variables para la imputación, de las cuales 12 corresponden a ingresos, 6 a gastos y 7 a activos. Estas variables fueron elegidas debido a su mayor grado de respuesta en todas las rondas. Sin embargo, solo se consideraron trece variables finales (Tabla 2). Para la selección de estas variables se tomaron en cuenta las siguientes consideraciones:

a) No deben generar correlación.

b) Deben tener correlaciones muy bajas con las covariables.

c) Deben presentar una correlación significativa entre variables y entre covariables.

Tabla 2. *Variables para imputar*

Variable de ingreso	Variable de gasto	Activos
Sueldo y Aguinaldo	Otras deudas	Valor neto de la casa o departamento
Jubilación	Gastos de visitas medicas	Valor neto del negocio
Viudez	Gasto de dentista	Otros activos
Apoyo familiar	Consumo	Valor neto de ahorros
		Valor del vehículo

Los aspectos generales de la base de datos son los siguientes: todos los datos se actualizaron a los valores de mayo de 2018, y la base de datos se conformó pegando completamente los datos. Solo se consideraron las observaciones principales, excluyendo a los fallecidos y las nuevas muestras.

En cuanto a los aspectos particulares de la base de datos: se introdujeron ceros estructurales en determinadas variables. Por ejemplo, en la pregunta sobre si se contaba con ingresos o no, si la respuesta era negativa, se asignaba un cero estructural. Además, las respuestas de "No sabe" (NS) o "No responde" (NR) se consideraron como valores faltantes (*missing*) (Tabla 3).

Tabla 3 *Consideraciones generales y particulares*

Generales	Particulares
Base pegada completamente	Ceros estructurales
Se considera solo al principal(target)	NR-NS=Missing
No se considera fallecidos, ni nuevas muestras.	

Dado que es posible recuperar los valores faltantes a partir de los datos observados, se introdujeron preguntas de rescate o de intervalos en el proceso de imputación, las cuales se consideraron como covariables. Los rangos de los intervalos se actualizaron para asegurar que los ingresos estén dentro del rango correspondiente, con el objetivo de mejorar la tasa de respuesta de las covariables. Esto se complementa con las respuestas de rescate y la consideración de ceros estructurales, lo que asegura que tanto las covariables de respuesta como las básicas no presenten valores faltantes.

A partir de los trece covariables restantes, se derivó un índice de nivel socioeconómico (INSE) mediante un análisis factorial confirmatorio, el cual también se integró en el proceso de imputación como una única covariable. Este índice se utiliza como una covariable específica para la ronda a ser imputada, pero no se emplea en comparaciones debido a que las covariables no son consistentes en todas las rondas. Por lo tanto, se excluyen aquellas relacionadas con la vivienda, la computadora y el acceso a internet (Anexo1). Los coeficientes de determinación se mantienen en aproximadamente

0.85 en todos los años. Sin embargo, al integrar todas las variables y años para formar un único índice, este coeficiente aumenta a 0.95.

Proceso de imputación

Una vez completado el proceso de ajuste, actualización y selección de variables y covariables, se inicia la etapa de imputación. En primer lugar, con el objetivo de hacer más consistente el supuesto de normalidad, las variables a imputar fueron transformadas utilizando el logaritmo. Esta transformación permitió ajustar el modelo y reducir la asimetría y la curtosis de las distribuciones (Anexo 2). Para abordar los valores cero, se procedió a cambiarlos por uno, y luego se utilizó el comando "lnskew0" para generar los logaritmos naturales de cada variable (por ejemplo: "*lnskew0 lnsalar_01_tot=salar_01_tot*"). Si bien en la mayoría de las variables fue posible realizar esta transformación, aquellas con un alto número de ceros presentaron errores en Stata. En estos casos, aproximadamente la mitad de los ceros fueron reemplazados por uno, y la otra mitad por dos. (Anexo 6)

La primera imputación se llevó a cabo en un formato *long*, pero no pudo completarse debido a la alta correlación entre algunas variables y a la escasez de observaciones en otras. Los cambios más significativos se observaron en las covariables. Inicialmente, se exploraron todas las variables y covariables disponibles, para luego reducir la selección a únicamente el índice de nivel socioeconómico, las covariables de intervalos, las covariables básicas o aquellas covariables con una mayor correlación con las variables a imputar (Anexo 3).

El método de Rubin (Rubin 1987; Vargas y Lorenz 2015), como se detalla en la metodología, emplea simulaciones de Monte Carlo. En este estudio, se generarán 10 imputaciones completas para los datos que se desean imputar. Para garantizar que las simulaciones converjan de manera satisfactoria, se utiliza la peor función lineal (WLF) durante el proceso de imputación de los datos. Esto implica la presencia de ruido blanco y no alguna tendencia o patrones extraños, y que la función de autocorrelación (ACF) de la WLF muestre una disminución en las autocorrelaciones para $p > 4$ (Anexo 4). En las cuatro rondas de este estudio, se cumplen ambas condiciones mencionadas, lo que garantiza una convergencia satisfactoria de los datos. Aunque todas las rondas cumplen las condiciones, la mejor es la ronda 2001.

Una vez seleccionadas las variables a imputar y las covariables que muestran una mejor convergencia, se realiza la imputación en formato *wide* con 10 iteraciones, lo que genera diez resultados para cada variable. Posteriormente, se calcula el promedio de cada variable, reduciéndola

a una única. Una vez finalizado este proceso, se transforman las variables imputadas logarítmicas utilizando el anti-logaritmo para volverlas a su forma original ($\ln(y+k)=x,\ y+k=exp(x), y=\exp(x)-k$). De esta manera, se completa el modelo ajustado y se consolida completamente en un único modelo final utilizando las reglas de Rubin (Rubin, D. B. 1987).

Se logró imputar el 100% de las variables faltantes, que en total ascendían a 28,892. Estas se encontraban mayormente concentradas en la ronda 2001, disminuyendo en las siguientes rondas. En cuanto a las secciones, los activos representan el mayor porcentaje de valores faltantes, con un 67%, seguido por los ingresos con un 19% y los gastos con un 13%. Dentro de las variables de activos, las que concentran más valores faltantes son "otros activos" y el "valor neto de la casa" (Anexo 5).

Funciones de distribución y validación del método de imputación

Después de aplicar el método de imputación, se procedió a desarrollar las funciones de distribución acumulada de ingresos y gastos por rondas, previamente transformándolas en logaritmo. Esto nos permite visualizar cómo se distribuyen y cuáles son las tendencias que siguen la imputación de la ENASEM, la imputación de este estudio y las variables sin imputar. Se observa que la distribución de ingresos de la imputación de la ENASEM en las rondas 2001 y 2003 se aparta de la tendencia de la imputación realizada en este estudio y de las variables sin imputar, principalmente debido al ingreso de ayuda de hijos mensual. Por otro lado, las rondas 2012 y 2015 muestran una distribución similar en las tres variables. En cuanto a la distribución acumulada de gastos, se identifica una tendencia común en las cuatro rondas, sin presentar diferencias significativas (Anexo 7).

Resultados

La imputación de los valores faltantes alcanzó el 100% (Anexo 5), lo que significa que no se aplicó ninguna restricción (Cuando las covariables contienen *missing*, la imputación se tiene que forzar y no se imputan todos los valores faltantes.) y se llevó a cabo de manera convencional, considerando las variables y las covariables por separado. Este logro se debió principalmente a que las covariables estaban completas y no presentaban valores faltantes. (Cuando las covariables no presentan valores faltantes, la imputación se logra desarrollarla en forma normal, separando las variables imputar de las covariables.) En el Anexo 5, se encuentran los resultados detallados de la imputación, incluyendo los valores mensuales en pesos de ingresos, gastos y activos para cada ronda, así como las diferencias entre ellos.

Ingresos

Los ingresos que resultaron después de la imputación son: sueldo y aguinaldos, ingreso por jubilación, ingreso por viudez y ayuda de hijos mensuales. En las cuatro rondas de la Enasem, se puede observar que, de las cuatro variables de ingreso a considerar, la ayuda de hijos mensual fue la más significativa en las rondas de 2001 y 2003, seguida por el sueldo y aguinaldos. A partir de la ronda de 2012, los ingresos por jubilación se convirtieron en los más significativos, seguidos por la ayuda de hijos mensual, y luego por los sueldos y aguinaldos. La diferencia entre la imputación del presente estudio y la de la Enasem es pequeña en las rondas de 2001, 2003 y 2015; únicamente hay una diferencia en la ronda de 2012, debido al mayor valor en los ingresos por jubilación.

Gastos

Los gastos que resultaron después de la imputación son: otras deudas, gastos del dentista, gasto total de visitas médicas y consumo. De las cuatro variables de gastos mensuales en las rondas de la Enasem, se puede observar que el consumo es la más importante en 2001 y 2003, mientras que otras deudas y consumo lo son en las rondas de 2012 y 2015. Los cambios más significativos entre las dos imputaciones se dan en las rondas de 2001 y 2012, debido a las mayores diferencias en otras deudas y consumo.

Activos

Los activos que resultaron después de la imputación son: valor neto de la casa, valor neto del negocio, otros activos, valor neto de los ahorros y valor del vehículo. Se puede observar que, de las cinco variables a considerar, el valor neto de la casa y el valor neto del negocio son los más importantes en las cuatro rondas. Los activos de la imputación de la Enasem siempre son más altos en las cuatro rondas, pero la diferencia es mayor en las rondas de 2012 y 2015, principalmente debido al valor neto de la casa.

Discusión

En todas las bases de datos, es inevitable encontrar información incompleta, conocida *como Missing Values o missing data*, que comprende un conjunto de valores faltantes debido a diversas causas desconocidas. Todos estos datos conforman un conjunto de información especial. En general, pueden clasificarse en dos grandes grupos: los datos faltantes por unidad, que se refieren a la falta de respuesta del entrevistado, y aquellos en los que los ítems del cuestionario no son entendidos o están mal formulados, lo que también conduce a la falta de respuesta. En encuestas longitudinales, la ausencia

de datos puede deberse a la ausencia de la persona o a que las unidades abandonen el estudio, lo que se conoce como "muerte de datos". El problema radica en que muy pocos estudios describen cómo manejan los datos faltantes, ya sea eliminándolos o utilizando algún proceso de imputación, y mucho menos informan sobre las consecuencias en la varianza y los estimadores

Los métodos tradicionales como la eliminación de datos (como *listwise*), emparejamiento de observaciones (*pairwise)*, uso de medias o *hot-deck*, métodos de regresión o métodos de imputación por máxima verosimilitud, son los más utilizados, los cuales pueden ser aplicados utilizando software estadístico. Es fundamental recordar que la imputación de datos no debe considerarse como un fin en sí mismo, sino como una herramienta disponible para obtener información completa.

En las últimas décadas, se han desarrollado algoritmos de imputación múltiple que ofrecen mejores propiedades estadísticas que los métodos tradicionales previamente descritos. La imputación múltiple (MI) es una técnica estadística basada en simulaciones que permite analizar datos con valores faltantes. Esta técnica requiere principalmente que el investigador posea la habilidad para imputar datos faltantes teniendo en cuenta el conocimiento que se tenga acerca del mecanismo que los produce, así como comprender completamente la base de datos a estudiar y las covariables que podrían explicarla. La imputación múltiple es una técnica mucho más completa que las tradicionales, pero requiere un conocimiento exhaustivo de la base de datos a estudiar.

Conclusión

Se desarrolló la aplicación de la imputación múltiple en la encuesta ENASEM, con la diferencia de que se explicó detalladamente todos los procedimientos, incluidas las simulaciones y sus convergencias. En comparación, la imputación propia de la ENASEM solo menciona que se aplicó imputación múltiple y que se utilizó un software para desarrollarla, pero no aclara los pasos seguidos, los patrones de covariables utilizados, etc.

La imputación de los valores faltantes alcanzó el 100%, lo que indica que se aplicaron las covariables correctas para explicar los valores faltantes. Este resultado se logró principalmente porque las covariables estaban completas y no presentaban valores faltantes. Cuando las covariables están completas, la imputación se puede llevar a cabo de forma normal, separando las variables a imputar de las covariables. En el artículo se plasmaron los resultados detallados de la imputación, incluyendo los valores mensuales en pesos de ingresos, gastos y activos para cada ronda, así como las diferencias entre ellos.

De las variables imputadas, los ítems de ingresos incluyen sueldos y aguinaldos, ingresos por

jubilación, ingresos por viudez y ayuda de hijos mensuales. Las variables de gastos, después de la imputación, incluyen otras deudas, gastos del dentista, gastos totales de visitas médicas y consumo. Las variables de activos, después de la imputación, incluyen el valor neto de la casa, el valor neto del negocio, otros activos, el valor neto de los ahorros y el valor del vehículo.

Limitaciones

En los diseños de muestra complejos, la selección de las observaciones depende de cómo se estratificó y conglomeró el marco de muestreo, así como del vector de ponderaciones asociado a las distintas unidades muestrales. El método de imputación múltiple no tiene en cuenta este aspecto, asumiendo que las simulaciones se generan bajo la premisa de una muestra aleatoria, donde todas las observaciones tienen la misma probabilidad de ser seleccionadas. Esto fue confirmado por Binder y Weimin (1996), quienes demostraron que, bajo el supuesto de un diseño aleatorio simple y sin reemplazo, los procesos de imputación múltiple producen estimadores adecuados para la media y los totales cuando se utilizan métodos bayesianos (bootstrap). Sin embargo, este resultado no se mantiene cuando se emplea un criterio determinístico, como la imputación aleatoria o el método de promedios.

Se puede decir que el mejor método de imputación es el que no es necesario aplicar, lo que implica que se deben utilizar todos los recursos disponibles para minimizar la falta de respuesta total y parcial en una encuesta. Todos los métodos de imputación tienen sus limitaciones y su correcta aplicación depende de cómo se comporten los datos faltantes. A medida que la falta de respuesta no sigue un patrón aleatorio, la eficacia de todas las metodologías disminuye, incluso en procedimientos estadísticamente robustos como la imputación múltiple y la máxima verosimilitud. No existe un método de imputación que sea superior en todas las situaciones. Cada caso es único, y la elección del método de sustitución de datos depende de la variable de estudio, el porcentaje de datos faltantes, el tipo de encuesta y el uso que se dará a la información imputada.

Para que una imputación múltiple sea exitosa, es esencial que el investigador tenga un conocimiento detallado de la base de datos en estudio. Además, es crucial que los datos faltantes sigan el patrón MAR (Missing At Random). Sin embargo, hay ocasiones en las que los supuestos del método no se cumplen, y no siempre es sencillo encontrar un modelo adecuado para la variable de interés. Es necesario utilizar paquetes estadísticos que contengan algoritmos específicos para este fin. Si no se manejan con cuidado, estos paquetes pueden operar como cajas negras, trasladando la responsabilidad a un procedimiento estadístico.

Referencias

Arbuckle, J. L., Marcoulides, G. A., & Schumacker, R. E. (1996). Full information estimation in the presence of incomplete data. Advanced structural equation modeling: Issues and techniques. Recuperado el 20 de Marzo de 2023, de: https://books.google.es/books?hl=es&lr=&id=VcHeAQAAQBAJ&oi=fnd&pg=PA243&dq=Arbuckle,+J.+L.+(1996).+Full+information+estimation+in+the+presence+of+incomplete+data.+In+G.+A.+Marcoulides+%26+R.+E.+Schumacker+(Eds.),+Advanced+structural+equation+modeling+(pp.+243-277).+Mahwah,+NJ:+Lawrence+Erlbaum&ots=HDHifc-F0b&sig=Fr54IfOT1Xbfj-Z6es4DmjU_zmc#v=onepage&q&f=false

Binder, D. A. y W. Sun (1996), Frequency valid multiple imputation for surveys with complex designs, Bussines Survey Methods Division, Statistics, Canada. Recuperado el 2 de Febrero del 2024 http://www.asasrms.org/Proceedings/papers/1996_044.pdf

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.. Recuperado el 2 de Febrero de 2024, de: https://books.google.es/books?hl=es&lr=&id=rNt5CgAAQBAJ&oi=fnd&pg=PR7&dq=Box,+G.+E.+P.+y+G.+M.+Jenkins.&ots=DJ94uQj0SE&sig=LTAQonDW3LqpJ0zxrvHieglW79k#v=onepage&q=Box%2C%20G.%20E.%20P.%20y%20G.%20M.%20Jenkins.&f=false

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society: series B (methodological). Recuperado el 15 de marzo del 2023: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, *39*(1), 1-22.

Encuesta Nacional de Salud y Envejecimiento de México (ENASEM) (2001,2003,2012 y 2015). Recuperado el 5 de Junio de 2023, de: https://enasem.org/Home/index_esp.aspx

Enders, C. K. (2001). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. Educational and Psychological Measurement, Recuperado el 20 de Marzo de 2023, de: https://journals.sagepub.com/doi/abs/10.1177/0013164401615001

Imputation_of_Non-Reponse_on_Economic_Variables_in_the_MHAS-ENASEM 2001, Wong y

Espinoza (2004). ENASEM. Recuperado el 6 de Octubre de 2023, de: http://mhasweb.org/Resources/DOCUMENTS/2001/Imputation_of_Non-Reponse_on_Economic_Variables_in_the_MHAS-ENASEM_2001.pdf

Imputation_of_Non-Reponse_on_Economic_Variables_in_the_MHAS-ENASEM 2003, Wong y Espinoza (2004). ENASEM. Recuperado el 10 de Octubre de 2023, de: http://mhasweb.org/Resources/DOCUMENTS/2003/Imputation_of_Non-Reponse_on_Economic_Variables_in_the_MHAS-ENASEM_2003.pdf

Imputation_of_Non-Reponse_on_Economic_Variables_in_the_MHAS-ENASEM 2012, Wong y Espinoza (2014). ENASEM. Recuperado el 10 de Octubre de 2023, de: http://mhasweb.org/Resources/DOCUMENTS/2012/Imputation_of_Non-Reponse_on_Economic_Variables_in_the_MHAS-ENASEM_2012.pdf

Imputation_of_Non-Reponse_on_Economic_Variables_in_the_MHAS-ENASEM 2015, González, Obregon, Orozco, Wong, y Zhang, Espinoza (2017). ENASEM. Recuperado el 12 de Octubre de 2023, de: http://mhasweb.org/Resources/2DOCUMENTS/2015/Imputation_of_Non-Reponse_on_Economic_Variables_in_the_MHAS-ENASEM_2015.pdf

Little, R y Rubin, D. (1987). Statistical Analysis with Missing Data. Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc. New York.. Recuperado el 5 de Junio de 2023, de https://leseprobe.buch.de/images-adb/61/97/61976bf3-cfac-463d-bb88-ca1ddb674cdf.pdf

Raghunathan, T. , Lepkowski, J., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey methodology, Recuperado el 12 de Octubre de 2023, de: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.405.4540&rep=rep1&type=pdf

Ritter, C., & Tanner, M. A. (1992). Facilitating the Gibbs sampler: the Gibbs stopper and the griddy-Gibbs sampler. Journal of the American Statistical Association. Recuperado el 2 de Noviembre de 2023, de: https://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475289

Roberts, G. O. (1992) “Convergence Diagnosis of the Gibbs Sampler”, in Bernardo, J. M.; J. O. Bergen; A. P. Dawid y A. F. M. Smith (eds.). Bayesian Statistics. Oxford University Press. Recuperado, el 1 de marzo de 2024, de: https://global.oup.com/academic/product/bayesian-statistics-4-9780198522669?lang=en&cc=gb

Rubin, D. B. (1976). Inference and missing data. Biometrika. Recuperado el 5 de Diciembre de 2023, de: https://academic.oup.com/biomet/article-abstract/63/3/581/270932

Rubin, D. B. (1987), Multiple imputation for non-response in surveys. New York, Wiley, Recuperado el 2 de Junio de 2023, de: https://books.google.com.mx/books?hl=es&lr=&id=bQBtw6rx_mUC&oi=fnd&pg=PR24&dq=Rubin,+D.+B.+(1987),+Multiple+imputation+for+nonresponse+in+surveys.+New+York,+Wiley&ots=8OtF7N1-eQ&sig=TQB8x1prrdrUg3dd-XnDnPd4w_Q#v=onepage&q=Rubin%2C%20D.%20B.%20(1987)%2C%20Multiple%20imputation%20for%20nonresponse%20in%20surveys.%20New%20York%2C%20Wiley&f=false

Vargas, D., & Lorenz, F. (2015). Inference with Missing Data Using Latent Growth Curves. Revista del Instituto Interamericano de Estadística.

Vargas, Valdés (2018) Ajuste estadístico a la distribución del ingreso en el Módulo de Condiciones Socioeconómicas 2015 mediante imputaciones multiples. Recuperado, el 20 de Marzo de 2020, de: https://www.inegi.org.mx/rde/2018/08/27/ajuste-estadistico-a-la-distribucion-del-ingreso-en-modulo-condiciones-socioeconomicas-2015-mediante-imputaciones-multiples/

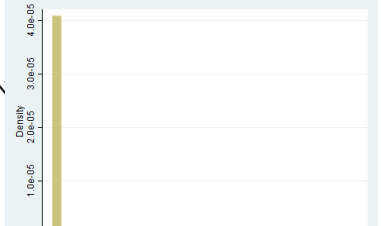
Anexos

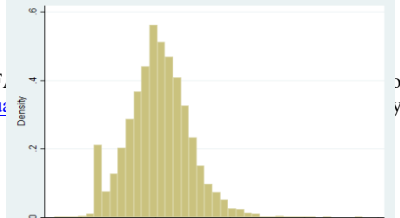
Anexo 1 *Covariables básicas*

Variable	Variable de gasto	Valores	Missing
Sexo	sexo_i	4352	0
	1	5203	0
Edad		9555	0
	2	9555	0
	edad01	9555	0
	edad03	9555	0
N° de hijos	edad12	9543	0
Localidad	edad15	9555	0
	nhijos_i	9555	0
	tam_loc01_03_i	9555	0
Años de estudio	tam_loc12_i	9555	0
	tam_loc15_i		
	esc_i		

Fuente: Elaboración propia en base a Enasem

Anexo 2 *Supuestos de normalidad*



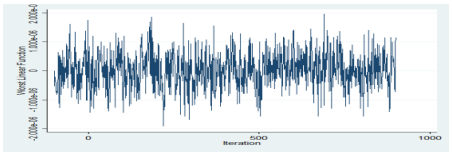


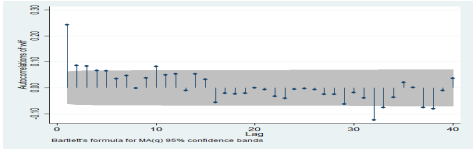
Anexo 3 *Variables imputadas finales*

	2001	2003	2012	2015
Sueldo y aguinaldos	lnsalar_01_tot	lnsalar_03_tot	lnsalar_12_tot	lnsalar_15_tot
Ingreso por jubilación	lnk58_1_0111	lnk58_1_0311	lnk58_1_1211	lnk58_1_1511
Viudez	lnk58_2_0111	lnk58_2_0311	lnk58_2_1211	lnk58_2_1511
Ayuda de hijos	lnayud_mens_011	lnayud_mens_031	lnayud_mens_121	lnayud_mens_151
mensual	lnk83_0111	lnk83_0311	lnk83_1211	lnk83_1511
Otras deudas	lnd9_3_0111	lnd9_3_0311	lnd9_3_1211	lnd9_3_1511
Gasto de dentistas	lnd9_5_0111	lnd9_5_0311	lnd9_5_1211	lnd9_5_1511
Gasto de visitas medicas	lnk85_0111	lnk85_0311	lnk85_1211	lnk85_1511
Consumo	lnj23_0111	lnj23_0311	lnj23_1211	lnj23_1511
Valor Neto de casa	lnk8_1_0111	lnk8_1_0311	lnk8_1_1211	lnk8_1_0111
Valor neto de los negocios	lnk42_0111	lnk42_0311	lnk42_1211	lnk42_1511
Otros activos	lnk31_1_0111	lnk31_1_0311	lnk31_1_1211	lnk31_1_1511
Valor neto de ahorros	lnk40_0111	lnk40_0311	lnk40_1211	lnk40_1511
Valor de vehículos.				

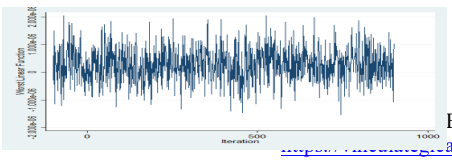
Anexo 4 *Series de tiempo y función de Autocorrelación*

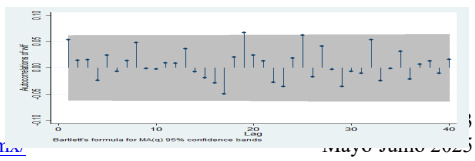
2001



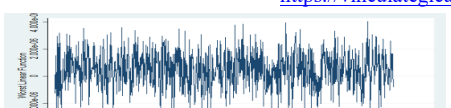


2003





2012



Anexo 5 *Valores faltantes e imputados del estudio*

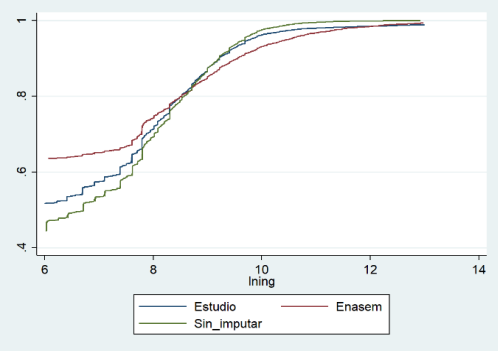
	2001	**Obs.: 9555**		**2003**	**Obs.: 8584**	
Variables a imputar	V.faltantes	Imputados	%	V.faltantes	Imputados	%
Sueldos por jubilación	217	217	2.27	69	69	0.8
Ingresos por jubilación	86	86	0.9	36	36	0.42
Viudez	11	11	0.12	10	10	0.12
Ayuda de hijos mensual	2504	2504	26.21	1058	1058	12.33
Otras deudas	118	118	1.23	89	89	1.04
Gasto del dentista	75	75	0.78	74	74	0.86
Gasto de vistas medicas	201	201	2.10	170	170	1.98
Consumo	871	871	9.12	817	817	9.52
Valor neto de casa	2591	2591	27.12	2614	2614	30.45
Valor neto del negocio	1013	1013	10.60	895	895	10.43
Otros activos	2416	2416	25.29	2026	2026	23.60
Valor neto de ahorros	342	342	3.58	243	243	2.83
Valor del vehículo	342	342	3.58	275	275	3.2
Suma	10787	10787		8386	8376	
	2012	**Obs.: 5909**		**2015**	**Obs.: 5323**	
Variables a imputar	V.faltantes	Imputados	%	V.faltantes	Imputados	%
Sueldos por jubilación	50	50	0.85	33	33	0.62
Ingresos por jubilación	99	99	1.68	102	102	1.92
Viudez	30	30	0.51	45	45	0.85
Ayuda de hijos mensual	784	784	13.27	590	590	11.08
Otras deudas	48	48	0.81	38	38	0.71
Gasto del dentista	69	69	1.17	51	51	0.96
Gasto de vistas medicas	93	93	1.57	71	71	1.33
Consumo	641	641	10.85	390	390	7.33
Valor neto de casa	366	366	6.19	1609	1609	30.23
Valor neto del negocio	346	346	5.86	288	288	5.41
Otros activos	1882	1882	31.85	1288	1288	24.2
Valor neto de ahorros	181	181	3.06	97	97	1.82
Valor del vehículo	354	354	5.99	184	184	3.46
Suma	4943			4786		

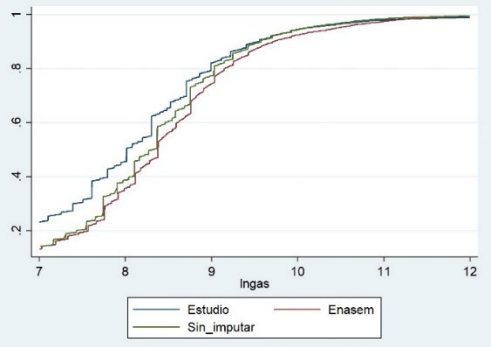
Anexo 6 *Imputación de las cuatro rondas*

```
*
**IMPUTACION 2001*****************
************************************************
************************************************
cd "C:\Users\52818\Documents\DOCTORADO\TESIS DOCTORADO\Avances tesis por semestre\5 - Quinto semestre\Dr. Delfino Vargas\lnskewo, revisazado y desarrllado\Guillermo\imputación 2001"
use cc_cov_con_missing_01.dta, replace

set more off
mi set mlong
mi describe, d

mi misstable patterns lnsalar_01_tot  lnk25_1_0111 lnk58_1_0111 lnk58_2_0111  ///
lnk13_1_0111 lnk83_0111 lnd9_3_0111 lnd9_5_0111 lnk85_0111 lnayud_mens_011 lnj23_0111 lnk8_1_0111 lnk40_0111 lnk42_0111 lnk31_1_0111 ///
sexo_i edad01 nhijos_i esc_i tam_loc01_03_i fs_INSE01 k44b_01_i  k12a1_011_i k26a1_011_i   k59a1_011_i k59a2_011_i k59a3_011_i k59a4_011_i  ///
 k14a1_011_i k84a_011_i d10a3_011_i d10a5_011_i k86a_011_i g21_011_i j24a_011_i ///
    k9a1_011_i k41a_011_i k32a1_011_i k32a2_011_i k32a3_011_i

mi misstable summarize lnsalar_01_tot  lnk25_1_0111 lnk58_1_0111 lnk58_2_0111  ///
lnk13_1_0111 lnk83_0111 lnd9_3_0111 lnd9_5_0111 lnk85_0111 lnayud_mens_011 lnj23_0111 lnk8_1_0111 lnk40_0111 lnk42_0111 lnk31_1_0111 ///
sexo_i edad01 nhijos_i esc_i tam_loc01_03_i fs_INSE01 k44b_01_i  k12a1_011_i k26a1_011_i   k59a1_011_i k59a2_011_i k59a3_011_i  ///
k59a4_011_i k14a1_011_i k84a_011_i d10a3_011_i d10a5_011_i k86a_011_i g21_011_i j24a_011_i ///
k9a1_011_i k41a_011_i k32a1_011_i k32a2_011_i k32a3_011_i

 mi register imputed lnsalar_01_tot  lnk25_1_0111 lnk58_1_0111 lnk58_2_0111  ///
lnk13_1_0111 lnk83_0111 lnd9_3_0111 lnd9_5_0111 lnk85_0111 lnayud_mens_011 lnj23_0111 lnk8_1_0111 lnk40_0111 lnk42_0111 lnk31_1_0111

set seed 9555

mi impute mvn lnsalar_01_tot  lnk25_1_0111 lnk58_1_0111 lnk58_2_0111  ///
lnk13_1_0111 lnk83_0111 lnd9_3_0111 lnd9_5_0111 lnk85_0111 lnayud_mens_011 lnj23_0111 lnk8_1_0111 lnk40_0111 lnk42_0111 lnk31_1_0111 ///
= sexo_i edad01 nhijos_i esc_i tam_loc01_03_i fs_INSE01 k44b_01_i  k12a1_011_i k26a1_011_i   k59a1_011_i k59a2_011_i k59a3_011_i  ///
 k59a4_011_i k14a1_011_i k84a_011_i d10a3_011_i d10a5_011_i k86a_011_i g21_011_i j24a_011_i ///
    k9a1_011_i k41a_011_i k32a1_011_i k32a2_011_i k32a3_011_i, force add(10) savewlf(wlf)

use wlf.dta, clear
tsset iter
tsline wlf, ytitle(Worst Linear Function) xtitle(Iteration)
ac wlf
```

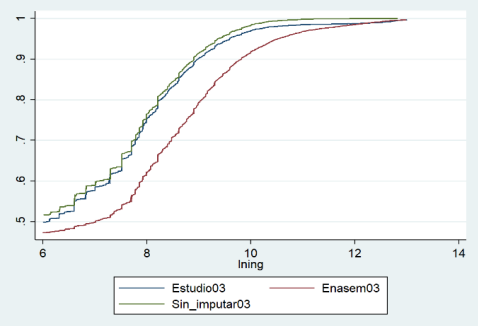
Anexo 7 *Funciones de distribuciones acumuladas de ingreso y gasto por ronda*

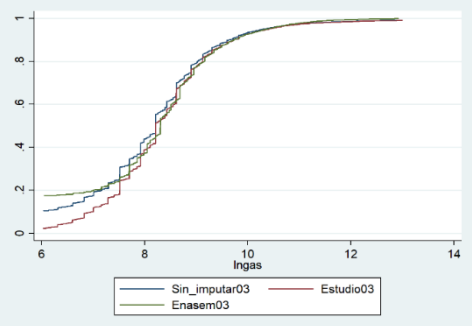
2001



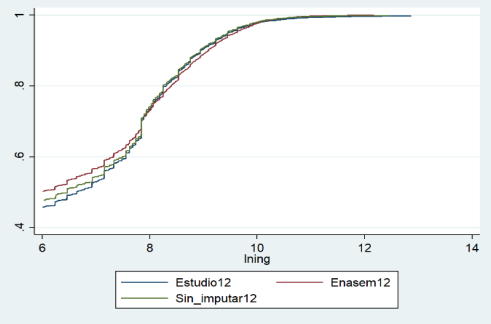


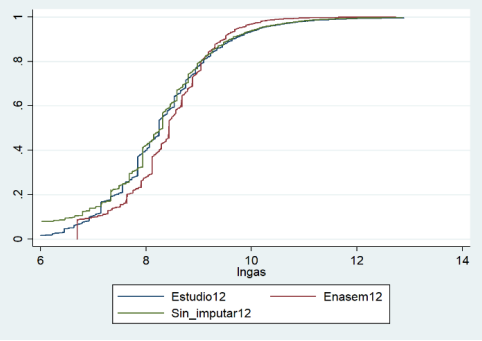
2003



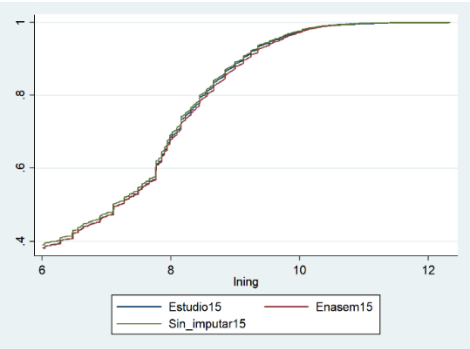


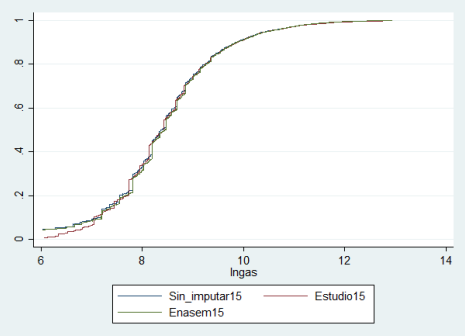
2012





2015





Fuente: Elaboración propia usando ENASEM 2001, 2003, 2012 y 2015.