
Desarrollo y Sociedad
ISSN: 0120-3584
revistadesarrolloysociedad@uniandes.edu.co
Universidad de Los Andes
Colombia

Restrepo Estrada, María Isabel; Marín Diazaraque, Juan Miguel
Imputación de ingresos en la Gran Encuesta Integrada de Hogares (GEIH) de 2010
Desarrollo y Sociedad, núm. 70, 2012, pp. 219-243
Universidad de Los Andes
Bogotá, Colombia

# Imputación de ingresos en la Gran Encuesta Integrada de Hogares (GEIH) de 2010*

## *Income Imputation in the 2010 Great Integrated Household Survey (GEIH)*

María Isabel Restrepo Estrada **
Juan Miguel Marín Diazaraque ***

## Resumen

Este trabajo presenta el problema del manejo de encuestas con datos faltantes, y para hacerle frente reseña una técnica conocida como *imputación*. Además se implementan algunas metodologías en la imputación de los ingresos y las ganancias de la Gran Encuesta Integrada de Hogares de Colombia de 2010. En ese sentido se evaluaron siete métodos para el total de la muestra y por grupos de estratos de la vivienda: la eliminación del caso, la imputación por media no condicionada, imputación por regresión estocástica, el *hot-deck,* el *hot-deck* con regresión, la imputación múltiple normal multivariada y la imputación múltiple con ecuaciones encadenadas. Se concluye que al no contar con porcentajes altos de no respuesta y dado que es posible que los datos

* Se agradecen los valiosos comentarios y sugerencias de Andrés Felipe García Suaza y de los dos evaluadores anónimos.

** Profesora del departamento de Economía de la Universidad de Antioquia. Correo electrónico: *mirestrepo@economicas.udea.edu.co*.

*** Profesor del departamento de Estadística, Universidad Carlos III de Madrid. Correo electrónico: *jmmarin@est-econ.uc3m.es*.

Este artículo fue recibido el 10 de septiembre de 2012; modificado el 30 de octubre de 2012 y, finalmente, aceptado el 9 de noviembre de 2012.

faltantes sigan un patrón que pueda ignorarse, los resultados de los métodos aplicados son relativamente similares.

*Palabras clave*: datos faltantes, imputación simple, imputación múltiple.

*Clasificación JEL*: C49, C81.

## Abstract

The aim of this paper is to present the issue in managing surveys with missing data. In order to address this problem, it is reviewed a technique known as imputation. Some methodologies on the income imputation in the 2010 Great Integrated Household Survey (GEIH 2010) were implemented. Seven methods for the total sample and groups of housing strata were evaluated: listwise deletion, unconditional mean imputation, stochastic regression imputation, hot-deck imputation, regression hot deck imputation, multivariate normal imputation, and multiple imputation by chained equations. It was discovered that the results of the applied methods are relatively similar, since non-response percentages are not high and the missing data may follow an ignorable pattern.

*Key words*: Missing data, single imputation, multiple imputation.

*JEL classification*: C49, C81.

## Introducción

El manejo de datos incompletos es un problema frecuente que se presenta en muchos tipos de conjuntos de datos, especialmente en encuestas de hogares. En las encuestas, la situación de datos incompletos generada por patrones de no respuesta puede aparecer de dos formas: en primer lugar, la no respuesta de toda la unidad, cuando el hogar no desea participar en la encuesta o no puede ser localizado por los encuestadores. En segundo lugar, la no respuesta de un ítem y sucede cuando la familia no desea responder a todas las preguntas hechas por el encuestador, bien sea por falta de comprensión de la pregunta, por falta de conocimiento de la respuesta o la renuencia a revelar cierta información (Barceló, 2008; Haziza, 2009).

Una consecuencia obvia de la primera modalidad de falta de respuesta en las encuestas es que el tamaño real de la muestra es inferior al previsto, y en este caso se pueden distinguir dos estrategias comúnmente usadas: 1) descartar todos los individuos con datos faltantes y 2) imputar los valores faltantes con el fin de producir una matriz de datos completa. Aunque la mejor solución para tratar con datos incompletos es evitar el problema mediante una planificación cuidadosa y consciente de la recolección de datos (Acock, 2005; Olinsky, Chen y Harlow, 2003), la estrategia (2), conocida como *imputación,* es el método más común para manejar este problema (Andridge y Little, 2010).

*Imputación* es la estimación de ítems faltantes en las respuestas de una encuesta, es decir, la imputación "completa una respuesta incompleta" (Sande, 1982). Algunos de los métodos de imputación más usados son: la eliminación de casos, la imputación por media condicionada y no condicionada, la imputación por regresión, el método *hot-deck* y el método de máxima verosimilitud, entre otros.

La imputación de datos faltantes surge como una alternativa que, más allá de buscar valores imputados plausibles, lo que busca es evitar perder información cuando se realizan análisis solo con casos completos, preservando las características de la distribución de los datos y las relaciones entre variables. Sin embargo, algunos autores señalan que ciertos métodos de imputación tales como los de imputación de la media y la eliminación de casos e imputación no estocástica, no son adecuados. Esto se debe a que no preservan la distribución de los datos completos, es decir, la distribución conjunta de los datos observados y faltantes (Barceló, 2008).

El objetivo de este trabajo consiste en presentar el problema del manejo de encuestas con datos faltantes y reseñar las técnicas de imputación con mayor implementación en la literatura y, además, evaluar algunos de estos métodos en la imputación de ingresos y ganancias de los ocupados —asalariados y cuenta propia— en la GEIH de Colombia para 2010. El trabajo se divide en tres secciones. En la primera sección se presenta el marco teórico del problema de datos faltantes y algunos métodos de imputación. En la segunda se ofrece la aplicación de algunos métodos a las variables de ingresos y ganancias y, finalmente, en la última sección se exponen las conclusiones finales de este estudio.

## I. Marco teórico

El problema de datos faltantes es muy común en Estadística, de manera que una variable con datos faltantes puede representarse como $Y = \{Y_{obs}, Y_{mis}\}$. En ese sentido, la aplicación de métodos basados en datos completos, sin considerar el mecanismo de datos faltantes, puede llevar a conclusiones erróneas. Esto se debe a que la validez y eficiencia de métodos basados en datos completos no puede garantizarse cuando los datos son incompletos (Rubin, 1996; Zhang, 2003). De acuerdo con Sande (1982) y Barceló (2008), la supresión de casos incompletos usualmente incrementa la ineficiencia de los estadísticos al reducir el tamaño de la muestra. Por otro lado, Horton y Lipsitz (2001) y Olinsky *et al.* (2003) afirman que la parte utilizable del conjunto de datos es susceptible a posibles sesgos si las personas que no respondieron son sistemáticamente diferentes de las que sí lo hicieron.

Determinar la aproximación analítica adecuada para tratar el problema de datos incompletos es una de las mayores inquietudes en el análisis de datos. El desarrollo de métodos estadísticos para tratar este problema ha sido un área activa de investigación en los últimos años (Horton y Lipsitz, 2001). Diversos autores consideran que en lugar de omitir observaciones con datos faltantes es mejor imputar esos datos con valores apropiados. Una vez imputados todos los datos faltantes, el conjunto de datos puede analizarse usando las técnicas estándar para datos completos (Hron, Templ y Filzmoser, 2010).

En general, se pueden distinguir dos tipos de métodos de imputación: el de imputación simple, cuando se asigna un único valor, y el de imputación múltiple, cuando se asignan varios valores. De acuerdo con Barceló (2008), el método de imputación simple permite estimar el parámetro, pero la estimación de la variabilidad se ignora, lo cual produce la subestimación de los errores estándar y los intervalos de confianza. De acuerdo con Walczak y Massart (2001), Rubin y Schenker (1986) y Rubin (1996), esta desventaja puede superarse mediante el mecanismo de imputación múltiple.

Sin embargo, los métodos de imputación están sujetos al mecanismo que generó la no respuesta. En general se pueden identificar tres mecanismos por

los cuales existen datos faltantes: el completamente aleatorio (MCAR)[1], el aleatorio (MAR)[2] y el no aleatorio (MNAR)[3].

Se dice que el patrón de datos faltantes es MCAR si la probabilidad de un elemento faltante es independiente tanto de los datos observados como de los no observados. En este caso es posible hacer inferencia correcta acerca de los parámetros poblacionales considerando solamente una submuestra de los que respondieron.

Sin embargo, en muchos casos prácticos —y en la mayoría de los métodos de imputación— se usa un supuesto más débil sobre el patrón de datos faltantes conocido como MAR (Nicoletti y Peracchi, 2006). Se dice que el patrón de datos faltantes es MAR si la probabilidad de que un dato falte depende solamente de los datos observados. Finalmente, se dice que es MNAR si la probabilidad de que un elemento sea faltante depende del valor no observado de los elementos faltantes. Debemos considerar que este mecanismo no puede ignorarse —NI— si los datos son, por ejemplo, censurados o si la distribución de estos es asimétrica, sugiriendo alguna clase de censura (Walczak y Massart, 2001).

Usualmente los datos contendrán muy poca información que nos permita decidir si los datos faltantes son MCAR, MAR o MNAR. Los investigadores deben hacer todo lo posible para averiguar por qué algunas observaciones faltan mientras que otras no. De acuerdo con Horton y Lipsitz (2001) y Baraldi y Enders (2010), solo es posible realizar pruebas para verificar el supuesto MCAR, aunque las pruebas que existen tienden a ser deficientes. En contraste, sin información adicional no es posible realizar pruebas para verificar los supuestos MAR y MNAR, particularmente en el último caso, pues este depende de datos no observados.

En la mayoría de los métodos de imputación se supone que el patrón de falta de datos es aleatorio (MAR); sin embargo, ese supuesto usualmente no es válido, pues la falta de respuesta suele estar asociada a las características de las personas y de los hogares (Barceló, 2008). Por ejemplo, muchos autores consideran que las familias de altos ingresos son las más renuentes a reportar esta

1 *Missing Completely At Random.*

2 *Missing At Random.*

3 *Missing Not At Random.*

información. Así y todo, Baraldi y Enders (2010) sostienen que agregar variables auxiliares puede aumentar la probabilidad de satisfacer el supuesto MAR; además, estas variables tienen el potencial de mejorar la calidad de las estimaciones resultantes.

Existe poca literatura en la estimación de parámetros en situaciones en las que el mecanismo que causa los datos faltantes no es aleatorio, es decir, casos en los que la probabilidad de no respuesta para la variable de interés depende del valor de esa variable (Greenlees, Reece y Zieschang, 1982). No obstante, se puede considerar que la variable ingreso sigue un patrón MAR si, en promedio, las personas que no respondieron tienen un ingreso similar a las que sí lo hicieron (Barceló, 2008).

## A. Mecanismos de imputación

Aunque en los últimos años se han propuesto métodos de imputación alternativos, aquí se presentan los métodos más usados (Olinsky *et al.*, 2003). Dentro de los nuevos métodos de imputación se encuentran el de imputación por cuantiles, el cual puede aplicarse incluso si no se dispone de información auxiliar (Muñoz y Rueda, 2009); los modelos neurodifusos, de análisis de componentes principales y de algoritmos genéticos (Hlalele, Nelwamondo y Marwala, 2009), entre otros.

### 1. Eliminación de la lista o del caso

Es la solución más común y se encuentra incluida en los paquetes estadísticos estándar. Esta técnica excluye un caso completo del análisis de los datos cuando al menos una variable tiene dato faltante. Una ventaja de este método es su fácil aplicación. Sin embargo, la eliminación de casos incompletos puede reducir el tamaño de la matriz de datos en forma drástica si el número de datos faltantes es alto. Esto es fundamental, pues un menor tamaño de la muestra inflará los errores estándar y reducirá el nivel de significación de los estadísticos calculados (Acock, 2005; Olinsky *et al.,* 2003).

Además, con este método se podrían sesgar los resultados, pues se estaría dando por cierto que los datos excluidos tienen las mismas características que los datos con información completa. De acuerdo con Little (1992), parece

entonces razonable buscar maneras de incorporar los casos incompletos en el análisis. Sin embargo, cabe señalar que cuando el patrón de datos faltantes es MCAR, se considera que al ser una submuestra aleatoria de la muestra original, esta conserva sus propiedades.

**2. Sustitución de la media**

Se basa en el hecho de que la media es una aproximación razonable de un valor para una observación aleatoriamente seleccionada de una distribución normal. Para datos faltantes que no son estrictamente aleatorios puede ser una aproximación pobre (Acock, 2005), y usualmente las personas que están en los extremos de la distribución son las que se niegan a responder (ingresos altos). En esta técnica el valor medio de la variable de todos los valores existentes de esa variable se calcula y sustituye por todos los casos con valores faltantes de esa variable. Si los datos son aproximadamente normales y los datos faltantes son MCAR, entonces las estimaciones resultantes de los parámetros no estandarizados serán insesgadas (Olinsky *et al*., 2003).

Una desventaja de esta técnica consiste en que reduce la variabilidad en la variable porque todos los datos faltantes se imputan con un valor constante, e imputar la media de los que respondieron puede desencadenar estimaciones no satisfactorias si los individuos que no responden son diferentes de los que sí lo hicieron (Greenlees *et al.,* 1982; Baraldi y Enders, 2010). Una variante de esta metodología, que se conoce como *sustitución de la media condicionada,* consiste en estimar la media por subgrupos relativamente homogéneos para realizar la imputación.

**3. Imputación por regresión**

Al igual que las técnicas descritas previamente, la imputación por regresión solo genera un único valor imputado. Utilizando información completa se busca predecir los valores de $\hat{y}$ a partir de covariables correlacionadas. Los valores predichos de $\hat{y}$ se usan para imputar los valores faltantes. En este caso, los valores imputados estarán en la recta de regresión. Baraldi y Enders (2010) aseguran que, al igual que la técnica de imputación de medias, esta técnica sesga las medidas de asociación y variabilidad, aunque produce estimaciones insesgadas de la media cuando los datos son MCAR o MAR.

Con el fin de incorporar la variabilidad en el proceso de imputación, algunos investigadores usan una versión modificada denominada *imputación por regresión estocástica*. Se considera que es un proceso estocástico cuando se añade a la regresión un término de error aleatorio. En este caso también se estiman valores predichos para las variables incompletas a las que después se les añade el término de error aleatorio. El valor resultante es el que se usa para imputar. El error es un número aleatorio generado de una distribución normal con media cero y varianza igual a la varianza residual obtenida del proceso de regresión (Baraldi y Enders, 2010).

De acuerdo con Baraldi y Enders (2010), a pesar de que el método de regresión estocástico permite obtener estimaciones comparables a los métodos de máxima verosimilitud y múltiple, este no permite ajustar los errores estándar para compensar el hecho de que los valores imputados son meras aproximaciones de los valores verdaderos. De este modo, los errores estándar son muy pequeños.

### 4. Imputación por máxima verosimilitud: algoritmo de maximización de esperanzas (EM)

Se basa en las relaciones observadas entre todas las variables e introduce un grado de error aleatorio para reflejar la incertidumbre de la imputación (Acock, 2005). Cabe señalar que este método ignora el mecanismo que generó la falta de datos. Siguiendo a Walczak y Massart (2001) y Olinsky *et al.* (2003), el algoritmo *EM* es un método iterativo que funciona así: en cada iteración *EM* consiste en un paso *E* y un paso *M*. El paso *M* lleva a cabo una estimación de máxima verosimilitud de los parámetros como si no hubiera datos faltantes. Entonces el paso *E* halla la esperanza condicional de los valores faltantes dados los datos observados y los parámetros actualmente estimados. Estas esperanzas se usan para reemplazar los datos faltantes.

Los valores se imputan iterativamente hasta que la matriz de covarianzas de la iteración actual es bastante similar a la de la precedente (Acock, 2005). Algunos autores sugieren que el algoritmo *EM* provee estimaciones con menos sesgo que la eliminación del caso o la sustitución de la media. Una crítica de este método, y en general de cualquier método de imputación simple, consiste en que los errores estándar no son válidos (Olinsky *et al.*, 2003).

## 5. Imputación múltiple

Este método fue propuesto por Rubin en la década de los ochenta (Rubin, 2003), y en los últimos años se ha convertido en una de las técnicas más populares para manejar el problema de datos faltantes (Robins y Wang, 2000). Provee múltiples conjuntos de datos imputados "m" para cada dato faltante con el fin de reflejar la incertidumbre de los valores imputados (Rubin, 1996). En general, valores de $m = 3$ o $5$ funcionan bastante bien en la práctica (Rubin, 1996).

De acuerdo con Horton y Lipsitz (2001), la imputación múltiple se puede resumir en tres pasos: imputación, análisis y combinación. Para obtener las $m$ imputaciones se usan métodos de simulación. En cada simulación se analiza el conjunto completo de datos usando técnicas estándar de análisis de datos (Rubin, 1996)[4] y posteriormente se combinan los resultados con el fin de obtener estimaciones robustas (Schenker y Taylor, 1996). Baraldi y Enders (2010) presentan una explicación detallada de las fórmulas usadas para combinar los resultados de las $m$ imputaciones.

En el caso de los parámetros estimados se toma la media aritmética de las estimaciones de cada conjunto de datos generado. Sin embargo, en el caso de los errores estándar el proceso es un poco más complicado, pues esto implica calcular los errores estándar de los conjuntos de datos imputados (la varianza dentro de la imputación) y un componente que cuantifica el grado en que las estimaciones varían a lo largo de los conjuntos de datos (la varianza entre imputaciones). Para la varianza dentro de la imputación se calcula la media aritmética de los errores estándar al cuadrado:

$$W = \frac{\Sigma SE_t^2}{m} \tag{1}$$

Donde $t$ es un conjunto de datos imputado particular; por ejemplo, en la aplicación presentada en este trabajo, $t = \{1,2,3,4,5\}$ y $m$ es el número total de conjuntos de datos imputados ($m = 5$). Por otro lado, para la varianza entre imputaciones se usa:

4 Son típicamente los estimadores de datos completos, las matrices de varianzas y covarianzas asociadas y los $p$ valores.

$$B = \frac{\Sigma\left(\hat{\theta}_t - \bar{\theta}\right)^2}{m-1} \tag{2}$$

donde $\hat{\theta}_t$ es el parámetro estimado, por ejemplo la ganancia media, del conjunto de datos imputado $t$ y $\bar{\theta}$ son la media del parámetro estimado en las $m$ imputaciones. Finalmente, el error estándar combinado se calcula a partir de la varianza dentro de la imputación y la varianza entre imputaciones:

$$SE = \sqrt{\left(W + B + \left(B \,/\, m\right)\right)} \tag{3}$$

De esta manera, la imputación múltiple resuelve el problema de subestimación de los errores estándar, que se da en el caso de la imputación simple, incorporando la varianza entre imputaciones en el error estándar. En ese sentido, los errores estándar de la imputación múltiple incorporan el hecho de que los valores imputados son conjeturas falibles acerca de los verdaderos valores de los datos (Baraldi y Enders 2010).

Si el porcentaje de datos faltantes es pequeño, las estimaciones y los errores estándar serán muy cercanos entre las distintas imputaciones y la estimación global y el error estándar serán casi iguales. Se ha demostrado que incluso con un gran porcentaje de datos faltantes, un número relativamente pequeño de imputaciones provee estimaciones de los errores estándar que son eficientes (Horton y Lipsitz, 2001). Cabe señalar que aunque algunos autores afirman que el supuesto MAR, generalmente considerado para los métodos de imputación, es clave para la validez de la imputación múltiple, Rubin (2003) asegura que la imputación múltiple puede usarse con datos faltantes donde el patrón no puede ignorarse[5].

## 6. *Hot-deck*

En su forma más simple, el método *hot-deck*[6] es un procedimiento no paramétrico ampliamente usado para imputar valores faltantes. Esencialmente este método reemplaza un valor faltante con el valor de un caso similar en la base de datos actual (Kim y Fuller, 2004); es decir, se reemplazan los valores de una

5 Es decir, MNAR.

6 El término *hot* se debe a que actualmente es procesado, en oposición al *cold-deck,* en el cual se usan datos preprocesados para los donantes de una base de datos diferente.

persona que no respondió (receptor) por valores observados de una persona que respondió (donante), el cual es similar al receptor de acuerdo con ciertas características observadas para las dos unidades. Con el fin de identificar el donante, el investigador debe seleccionar las variables de clasificación. Todos los posibles donantes que coinciden en aquellas variables de clasificación se agrupan, de donde un caso se elige (Olinsky *et al.*, 2003).

En general se distinguen dos grandes tipos de métodos *hot-deck* (Andridge y Little, 2010), a saber: el determinístico, en el que un único donante se identifica y los valores se imputan de ese caso, usualmente el "vecino más cercano" basado en alguna métrica, y el aleatorio, en el cual el donante se selecciona aleatoriamente de un conjunto de posibles donantes. De acuerdo con Kim y Fuller (2004), la imputación *hot-deck* aleatoria preserva las características distribucionales del conjunto de datos imputado, y la selección aleatoria de donantes introduce variabilidad en el proceso, la cual se denomina *varianza de la imputación*.

Una de las mayores ventajas de este método es que imputa datos reales y, por tanto, realistas y además puede incorporar información procedente de otras covariables. Sin embargo, una gran debilidad de esta técnica es que requiere buenos emparejamientos[7] de donante y receptores que reflejen la información de las covariables disponibles. Cuando se eligen variables para crear el conjunto de donantes, la prioridad debe ser seleccionar variables que sean predictivas del ítem que se imputa. Sin embargo, los métodos de imputación *hot-deck* no pueden preservar todas las relaciones entre las variables del cuestionario a raíz de la necesidad de condicionar solo a un pequeño número de covariables. Finalmente, otra debilidad del método es que la escasez de donantes puede llevar al excesivo uso de un solo donante, por lo que muchas de las metodologías de *hot-deck* restringen el número de veces que un donante puede usarse para la imputación (Andridge y Little, 2010)[8].

Existen diversas aproximaciones alternativas al *hot-deck* tradicional: por ejemplo, la imputación *hot-deck* por regresión y los métodos bayesianos de imputación múltiple. El método *bootstrap* bayesiano aproximado presentado por

7 Encontrar buenos emparejamientos es más probable en muestras grandes que en pequeñas.

8 De acuerdo con estos autores, la elección óptima del número de donantes es un tema interesante de investigación.

Rubin y Schenker (1986) puede verse como un método de imputación *hot-deck* en un contexto de imputación múltiple (Kim *et al.*, 2004).

## II. Aplicación en la GEIH

Los datos de ingresos son de gran importancia en los análisis económicos; sin embargo, en las encuestas de hogares este tipo de información es propensa a problemas de no respuesta, como en el caso de la GEIH. Esta es realizada por el Departamento Administrativo Nacional de Estadística (DANE) y busca recolectar información sobre variables socioeconómicas de la población, especialmente sobre las condiciones de empleo y características generales de esta.

El tamaño de la muestra mensual es aproximadamente de 23.000 hogares y una muestra total anual de 271.620 hogares[9], con información de cobertura nacional, urbano-rural, regional, departamental y para las capitales de los departamentos, lo que la convierte en la fuente primaria de información para construir estadísticas oficiales y realizar todo tipo de estudios del mercado laboral.

En este trabajo se usa la GEIH de 2010 con el fin de verificar el funcionamiento de algunos métodos de imputación[10]. En concreto, se realizará una aplicación para los ingresos laborales y las ganancias, de asalariados y cuenta propia, respectivamente, ya que de acuerdo con Guataquí, García y Rodríguez (2009), estas categorías son muy diferentes debido a las características laborales de cada grupo[11].

Como ya se reseñó, muchos autores consideran que la variable ingresos es MNAR ya que, a pesar de que en la práctica este supuesto no puede validarse, se supone que los individuos con ingresos altos son más renuentes a reportar su ingreso en una encuesta de hogares. Con el fin de explorar si en efecto los individuos con ingresos más altos no reportan el dato, en este trabajo se realiza

9 En este caso la unidad de análisis es el individuo y no el hogar.

10 La imputación de ingresos es muy útil cuando se van a realizar estudios sobre pobreza, distribución del ingreso y política laboral general.

11 Como en la mayor parte de la literatura en el mercado laboral y en métodos de imputación, en este trabajo se usará el logaritmo de las variables de ingreso y ganancias.

un análisis de la tasa de no respuesta por el estrato de la vivienda, de manera que se da por sentado que a un estrato más alto el ingreso es mayor[12].

Como puede observarse en el cuadro 1, en efecto las personas de ingresos altos son las que menos reportan este dato en la encuesta. De manera que se podría suponer, sin ser de ninguna manera algo concluyente, que el patrón de datos faltantes es MNAR. Esto tiene importantes implicaciones porque en la mayoría de los métodos de imputación se presume que el patrón de datos faltantes es MAR (o MCAR). Sin embargo, en este trabajo se van a explorar dos visiones alternativas. Por un lado, Rubin (2003) asegura que la imputación múltiple puede usarse con datos faltantes donde el patrón no puede ignorarse (es decir, MNAR) y, por otro, Barceló (2008) señala que se puede considerar que la variable ingreso sigue un patrón MAR si, en promedio, las personas que no respondieron tienen un ingreso similar a las que sí lo hicieron. En ese sentido, se van a realizar las imputaciones para todas las observaciones e imputaciones por estratos con un ingreso promedio "similar" (1 y 2, 3 y 4, 5 y 6).

**Cuadro 1.** Tasa de no respuesta por estrato de la vivienda

| Estrato | Ganancia (cuenta propia) | Ingreso (asalariados) |
|---|---|---|
| 1 | 6,84 | 2,96 |
| 2 | 5,99 | 3,09 |
| 3 | 6,00 | 3,65 |
| 4 | 9,35 | 6,78 |
| 5 | 13,18 | 11,25 |
| 6 | 16,33 | 10,71 |
| Todos | 6,65 | 3,82 |

*Fuente*: cálculos propios.

Puesto que algunos métodos de imputación requieren información complementaria recogida por otras covariables, bien sea para encontrar individuos con características similares o para aplicar una regresión, aquí se va a considerar una especificación de una ecuación minceriana[13] (Mincer, 1974) cuyas estimacio-

12 Cabe señalar que se consideran solo los estratos 1 a 6, dejando por fuera las zonas rurales.

13 Aunque este tipo de especificación ha sido objeto de algunas críticas, aquí se estima no con el fin de realizar inferencia, sino de predecir.

nes[14] se presentan en el cuadro 2. De manera que se tienen en cuenta la edad, la experiencia[15], el género, el nivel de escolaridad y las horas trabajadas.

**Cuadro 2 (a).** Ecuación minceriana para el logaritmo de las ganancias

| Variables | LogGanancia | | | |
|---|---|---|---|---|
| | Total | 1 y 2 | 3 y 4 | 5 y 6 |
| Edad | 0,065276*** (0,001403) | 0,069716*** (0,001599) | 0,047875*** (0,002683) | 0,055209*** (0,00943) |
| Exper. | -0,00065*** (0,000015) | -0,000764*** (0,000018) | -0,000487*** (0,000029) | -0,000484*** (0,000101) |
| Género | -0,562858*** (0,007519) | -0,621465*** (0,008703) | -0,446424*** (0,01378) | -0,25838*** (0,047129) |
| Esc. | 0,096432*** (0,000835) | 0,057910*** (0,00111) | 0,103919*** (0,001583) | 0,134988*** (0,005983) |
| Horas | 0,016781*** (0,000172) | 0,016397*** (0,000198) | 0,018511*** (0,000319) | 0,02129*** (0,001314) |
| Constante | 10,51363*** (0,03283) | 10,8273*** (0,03745) | 10,7244*** (0,064510) | 10,1447*** (0,223868) |
| Observaciones | 58.285 | 38.703 | 18.036 | 1.546 |
| $R^2$ | 0,3845 | 0,3676 | 0,3773 | 0,3885 |
| Test F | 7.280,21*** | 4.498,7*** | 2.184,68*** | 195,68*** |

*Fuente*: cálculos propios. ***$p < 0,01$, **$p < 0,05$. Errores estándar en paréntesis.

Como se observa en el cuadro 2, todas las variables son significativas al 1%. Tanto la experiencia —calculada como el cuadrado de la edad— como el género tienen signo negativo. En el caso del primero, de acuerdo con Guataquí *et al.* (2009) se puede decir que muestra el comportamiento esperado de una función cóncava, indicando que el ingreso crece con la edad pero a una tasa decreciente. Además se observa que la escolaridad y las horas trabajadas tienen en todos los casos signo positivo, donde la escolaridad tiene mayor efecto en el ingreso de los asalariados y las horas trabajadas en las ganancias de los cuenta propia.

14 Para los casos con información completa.

15 Se consideran *proxys* de la experiencia la edad y la edad al cuadrado.

**Cuadro 2 (b).** Ecuación minceriana para el logaritmo de los ingresos

| Variables | LogIngreso | | | |
|---|---|---|---|---|
| | Total | 1 y 2 | 3 y 4 | 5 y 6 |
| Edad | 0,061609*** (0,001108) | 0,065875*** (0,001346) | 0,056357*** (0,001820) | 0,079776*** (0,00725) |
| Exper. | -0,000547*** (0,000014) | -0,000670*** (0,000017) | -0,00049*** (0,000023) | -0,000726*** (0,000085) |
| Género | -0,144719*** (0,004547) | -0,131305*** (0,005649) | -0,143971*** (0,007032) | -0,23886*** (0,027854) |
| Esc. | 0,107957*** (0,000562) | 0,070739*** (0,000767) | 0,120694*** (0,000965) | 0,146435*** (0,004606) |
| Horas | 0,013392*** (0,000157) | 0,014073*** (0,000185) | 0,013039*** (0,000259) | 0,013886*** (0,001247) |
| Constante | 10,3495*** (0,021839) | 10,6081*** (0,026327) | 10,3672*** (0,036115) | 9,9687*** (0,151975) |
| Observaciones | 69.935 | 40.197 | 27.151 | 2.587 |
| $R^2$ | 0,4349 | 0,3215 | 0,4521 | 0,4288 |
| Test F | 10.763,84*** | 38.080,03*** | 4.479,97*** | 387,50*** |

*Fuente*: cálculos propios. ***p < 0,01, ** p < 0,05. Errores estándar en paréntesis.

En ese sentido, se corrobora la necesidad de considerar ambas categorías de ocupados separadamente, como lo sugieren Guataquí *et al.* (2009) en un estudio para estimar los determinantes de los ingresos laborales en Colombia, y en este caso concreto para realizar la imputación de ingresos en la GEIH. Finalmente, el $R^2$ obtenido, pese a no ser muy alto, corresponde a lo esperado, de acuerdo con la evidencia empírica sobre ecuaciones mincerianas, puesto que hay factores que no pueden capturarse en el modelo, como las habilidades específicas y si la persona tiene o no redes de contactos laborales.

A continuación se estiman los modelos de imputación para el total y por grupos de estratos[16], tres de ellos por imputación simple —eliminación del caso, media no condicionada y regresión simple— y cuatro por imputación múltiple —*hot-deck, hot-deck* con regresión, normal multivariado y de ecuaciones

16 Los resultados por estratos aparecen en el anexo 1.

encadenadas—. En el método de eliminación del caso simplemente se consideran las unidades con información completa. En el método de imputación de la media no condicionada se calcula la media de ingresos y ganancias con los casos que reportaron y este dato se usa para rellenar las celdas con valores faltantes. Por su parte, el método de regresión simple estocástico relaciona $Y$ con las covariables de la ecuación minceriana más un término de error aleatorio, y los valores predichos de $\hat{Y}$ se usan para la imputación.

Todos los métodos de imputación múltiple se estimaron fijando un $m = 5$, pues como ya se mencionó, es una cantidad razonable cuando el porcentaje de datos faltantes no es muy alto y a causa del volumen del conjunto de datos, dados los requerimientos computacionales de algunas técnicas. Para combinar los parámetros estimados y los errores estándar de los $m$ conjuntos de datos imputados se usan las fórmulas presentadas en Rubin (1987, citado por Baraldi y Enders, 2010).

La aproximación básica de imputación múltiple extrae valores plausibles de la distribución predictiva *a posteriori* de los valores faltantes de ingresos y ganancias, dada la información observada de estas variables, junto con el efecto de covariables relacionadas, bajo un modelo de regresión bayesiano (Rubin y Schenker, 1986). Habitualmente se usan métodos de simulación basados en cadenas de Markov (MCMC) que generan muestras de la distribución predictiva *a posteriori* de $f(Y_{mis} \mid Y_{obs})$.

En el procedimiento *hot-deck* se obtienen las estimaciones siguiendo el método *bootstrap* bayesiano aproximado de Rubin y Schenker (1986). En el procedimiento simple, los datos de ingresos y ganancias de los receptores serán imputados con valores de donantes que sean similares, según las características modeladas en la ecuación minceriana. Por su parte, el procedimiento *hot-deck* con regresión calcula los valores predichos $\hat{Y}$ de la regresión de los ingresos y las ganancias con las variables de la misma ecuación, para cada uno de los casos, y a partir del $\hat{Y}$, se calculan las distancias que permitan encontrar casos similares. Para imputar se considera un donante y receptor cuya distancia en $\hat{Y}$ sea mínima y se asigna el dato de ingreso o ganancia del donante al receptor.

Los métodos de imputación normal multivariante y de ecuaciones encadenadas extraen aleatoriamente los valores de la distribución de $f(Y_{mis} \mid Y_{obs})$

por medio de un proceso iterativo. Aunque son muy similares, los métodos de imputación por ecuaciones encadenadas y normal multivariante tienen algunas diferencias. En el modelo normal multivariante las imputaciones se realizan a partir de una única distribución normal multivariante; es decir, se usa la información de todas las variables para imputar todas las otras variables con base en un único modelo. En el método de ecuaciones encadenadas no se supone una distribución multivariante, y los valores imputados se generan a partir de un conjunto de modelos univariantes en el que una única variable se imputa a partir de un grupo de covariables.

En el caso de la imputación por ecuaciones encadenadas se estima primero el modelo de regresión para la variable que tenga menos datos faltantes y se imputa con los valores predichos obtenidos. Luego con esta información adicional se estima la regresión para la variable que a continuación tenga menos datos faltantes y así sucesivamente. Aunque este método es mucho más flexible que el normal multivariante al permitir variables de distinto tipo, que no siempre se ajustan a la normalidad, carece de fundamentos teóricos sólidos e implica un gran esfuerzo computacional en conjuntos de datos grandes.

Para validar la robustez de los resultados en este ejercicio se realiza la comparación por grupos de estratos con el fin de verificar la aleatoriedad de los datos faltantes y adicionalmente se realizan las estimaciones con un tamaño de la muestra reducido. Dado que se trabaja con datos reales, de manera que no se conocen los verdaderos ingresos y ganancias medias, se han reestimado todos los modelos de imputación eliminando aleatoriamente un 10% de los datos (conocidos).

En el cuadro 3 se presentan los resultados de los ingresos y ganancias medias imputados para todos los estratos y en el anexo 1 se presentan los resultados por grupos de estratos.

Como se observa en el cuadro 3, los resultados son relativamente similares en la mayor parte de los casos; sin embargo, al aumentar el porcentaje de datos faltantes, los resultados se distorsionan mucho con la imputación por regresión estocástica. Además, para ese 10% de datos que se extrajeron y que se conocen se calcularon tres medidas para evaluar la predicción con el valor imputado y el dato real: el error cuadrático medio, el error absoluto medio y la U de Theil.

**Cuadro 3.** Ganancias e ingresos medios de todos los estratos

| Método | Ganancia media | | Ingreso medio | |
|---|---|---|---|---|
| | Datos disponibles | 10% menos | Datos disponibles | 10% menos |
| Eliminación del caso | 579.101 (3.805,55) | 580.872 (4.041,79) | 934.561 (4.911,22) | 935.429 (5.203,53) |
| Media no condicionada | 558.855 (3.495,77) | 557.588 (3.697,54) | 923.848 (4.713,26) | 923.173 (5.017,57) |
| Imputación por regresión estocástica | 577.164 (3.604,41) | 715.214 (5.113,37) | 950.557 (5.033,04) | 828.227 (7.986,27) |
| *Hot-deck* | 579.638 (3.786,98) | 579.615 (4.018,00) | 934.586 (5.035,87) | 935.963 (5.238,53) |
| *Hot-deck* con regresión | 578.321 (3.705,54) | 580.771 (3.901,31) | 935.769 (5.036,16) | 937.006 (5.211,37) |
| Imputación múltiple nmv | 584.154 (1.106,14) | 586.525 (1.171,90) | 941.864 (1.442,27) | 943.530 (1.523,22) |
| Imputación múltiple ec. encadenadas | 580.583 (1.495,53) | 581.906 (1.584,07) | 943.240 (2.001,46) | 943.604 (2.108,22) |

*Fuente*: cálculos propios. Errores estándar en paréntesis.

$$ECM = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{4}$$

$$EAM = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \;=\; \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{5}$$

$$U = \frac{\sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{n}}}{\sqrt{\frac{\sum_{i=1}^{n}\hat{y}_i^{\,2}}{n}} + \sqrt{\frac{\sum_{i=1}^{n}y_i^{\,2}}{n}}} \tag{6}$$

Donde $\hat{y}_i$ es el valor imputado e $y_i$ es el valor real. Por otro lado, el estadístico $U$ toma como valores $0 < U < 1$, donde 0 significa predicción perfecta. Para calcular estas medidas se repitió el proceso 50 veces y se calculó el promedio de cada una, a causa del componente de incertidumbre que existe en el proceso de imputación (*véase* cuadro 4).

**Cuadro 4.** Medidas de evaluación del error de las imputaciones

| Método | Ganancia media | | | Ingreso medio | | |
|---|---|---|---|---|---|---|
| | ECM | EAM | UT | ECM | EAM | UT |
| Eliminación del caso | 0,72430 | 0,64487 | 0,00031 | 0,33175 | 0,41677 | 0,00019 |
| Media no condicionada | 0,70133 | 0,63385 | 0,00038 | 0,33272 | 0,41738 | 0,00020 |
| Imputación por regresión estocástica | 0,72541 | 0,64537 | 0,00037 | 0,33485 | 0,41722 | 0,00022 |
| *Hot-deck* | 0,18375 | 0,14298 | 0,00290 | 0,37659 | 0,21221 | 0,00001 |
| *Hot-deck* con regresión | 0,14803 | 0,14273 | 0,00004 | 0,26780 | 0,21036 | 0,00002 |
| Imputación múltiple nmv | 0,71881 | 0,64211 | 0,00028 | 0,33285 | 0,41727 | 0,00014 |
| Imputación múltiple ec. encadenadas | 0,72025 | 0,64395 | 0,00037 | 0,33365 | 0,41681 | 0,00019 |

*Fuente*: cálculos propios.

De acuerdo con las medidas de evaluación del error de predicción del 10% de los datos conocidos, que fueron eliminados, el método *hot-deck* por regresión muestra un comportamiento claramente superior a las otras medidas. Esto es importante, en la medida en que pese a que los ingresos y ganancias medios obtenidos son similares a los otros métodos, particularmente al de eliminación del caso (que es el que se usa comúnmente), no se estaría desperdiciando información adicional del conjunto de datos y las predicciones se pueden mejorar.

Finalmente, con el fin de verificar si existen diferencias entre las ganancias e ingresos medios, obtenidos a partir de la imputación de todos los estratos y de los grupos de estratos, se presenta el cuadro 5, comparativo. Para los grupos de estratos se ponderó por el número de observaciones en cada estrato con respecto al total.

Como se observa en el cuadro 5, no existen diferencias muy significativas entre la imputación para el total del conjunto de datos y por grupos de estratos. De manera que si se quiere ser riguroso con el supuesto del patrón de datos faltantes, se puede realizar la imputación estratificada. Sin embargo, para la GEIH es posible considerar que los datos de ingresos y ganancias pueden imputarse para el conjunto total de información suponiendo que son MAR.

**Cuadro 5.** Ganancias e ingresos medios mediante la imputación de todos los estratos y por grupos de estratos

| Método | Ganancia media | | Ingreso medio | |
|---|---|---|---|---|
| | Todos | Por estratos | Todos | Por estratos |
| Eliminación del caso | 579.101 | 579.104 | 934.561 | 934.538 |
| Media no condicionada | 558.855 | 555.172 | 923.848 | 918.194 |
| Imputación por regresión estocástica | 577.164 | 587.164 | 950.557 | 954.036 |
| *Hot-deck* | 579.638 | 578.665 | 934.586 | 933.979 |
| *Hot-deck* con regresión | 578.321 | 580.616 | 935.769 | 937.913 |
| Imputación múltiple nmv | 584.154 | 585.432 | 941.864 | 942.161 |
| Imputación múltiple ec. encadenadas | 580.583 | 579.821 | 943.240 | 939.606 |

*Fuente*: cálculos propios.

## III. Conclusiones

Gran cantidad de estudios que hacen uso de bases de datos con información faltante recurren a la técnica de eliminación de casos o unidades que registren información incompleta, porque es la técnica más sencilla para enfrentar este tipo de problema. Sin embargo, esta opción no está exenta de inconvenientes y por eso debe considerarse con cautela. En efecto, en muchos casos puede inducir a errores graves de estimación cuando los individuos que no respondieron son sistemáticamente diferentes de los que sí lo hicieron. En ese sentido, a pesar de que en este estudio la eliminación de unidades con datos incompletos parece funcionar bien, en muchos otros estudios puede ser la peor alternativa.

En ese sentido, la imputación de datos faltantes surge como una opción que más allá de buscar valores imputados plausibles, lo que busca es evitar perder información cuando se realizan análisis solo con casos completos, preservando las características de la distribución de los datos y las relaciones entre variables. En consecuencia, es importante que el investigador trate de inferir el mecanismo que generó la ausencia del dato, es decir, si el patrón es MCAR, MAR o MNAR. Tal inferencia resulta de gran dificultad en la práctica, pues, por ejemplo, no se puede verificar el último mecanismo, al depender de la misma información que no se observa.

En este estudio se partía de la premisa inicial, ampliamente reseñada en la literatura, de que los individuos con mayores ingresos son renuentes a revelar esta información y por tanto el patrón de datos era posiblemente MNAR. Con el fin de evaluar el porcentaje de no respuesta por nivel de ingreso, en este trabajo se usó la información del estrato de la vivienda, dando por cierto que a un estrato mayor corresponde mayor ingreso. Al inducir que posiblemente el patrón era MNAR, se procedió a realizar la imputación por grupos de estratos y para el total.

En este trabajo se evaluaron siete métodos, a saber: eliminación del caso, imputación por media no condicionada, imputación simple, *hot-deck*, *hot-deck* con regresión, imputación múltiple normal multivariada e imputación múltiple con ecuaciones encadenadas. En general, las ganancias e ingresos medios obtenidos con los métodos evaluados arrojaron resultados similares, quizá porque no se cuenta con porcentajes altos de no respuesta y porque es posible que los datos no sean MNAR.

Con el fin de verificar qué mecanismos se desempeñaban mejor, se eliminó un 10% adicional de datos. Se reestimaron los modelos y se calcularon tres medidas de evaluación de las predicciones de los datos imputados con los conocidos: el error cuadrático medio, el error absoluto medio y la U de Theil. Como la mayoría de los métodos se basan en simulaciones o incluyen errores aleatorios, se replicó este proceso cincuenta veces y se calcularon las medias de las tres medidas mencionadas. De acuerdo con los resultados, el mejor método de imputación en este caso fue el *hot-deck* con regresión.

## Referencias

1. ACOCK, A. (2005). "Working with missing values", *Journal of Marriage and Family,* 67:1012-1028.

2. ANDRIDGE, R. y LITTLE, R. (2010). "A review of hot deck imputation for survey non-response", *International Statistical Review,* 78:40-64.

3. BARALDI, A. N. y ENDERS, C. K. (2010). "An introduction to modern missing data analyses", *Journal of School Psychology,* 48:5-37.

4. BARCELÓ, C. (2008). "The impact of alternative imputation methods on the measurement of income and wealth: Evidence from the Spanish survey of household finances", *Documentos de Trabajo* (0829):9-64.

5. GREENLEES, J. S., REECE, W. S. y ZIESCHANG, K. D. (1982). "Imputation of missing values when the probability of response depends on the variable being imputed", *Journal of the American Statistical Association*, 77(378): 251-261.

6. GUATAQUÍ, J. C., GARCÍA, A. F. y RODRÍGUEZ, M. (2009). "Estimaciones de los determinantes de los ingresos laborales en Colombia con consideraciones diferenciales para asalariados y cuenta propia", *Documentos de Trabajo*, 70:1-22.

7. HAZIZA, D. (2009). "Imputation and inference in the presence of missing data", *Sample Surveys: Design, Methods and Applications*, 29A: 215-246.

8. HLALELE, N., NELWAMONDO, F. y MARWALA, T. (2009). "Imputation of missing data using PCA, neuro-fuzzy and genetic algorithms", en *Advances in neuro-information processing, lecture notes in computer science* (pp. 485-492). Berlin-Heidelberg, Springer.

9. HORTON, N. J. y LIPSITZ, S. R. (2001). "Multiple imputation in practice: Comparison of software packages for regression models with missing variables", *The American Statistician*, 55(3):244-254.

10. HRON, K., TEMPL, M. y FILZMOSER, P. (2010). "Imputation of missing values for compositional data using classical and robust methods", *Computational Statistics and Data Analysis*, 54:3095-3107.

11. KIM, J. K. y FULLER, W. (2004). "Fractional hot deck imputation", *Biometrika*, 91(3):559-578.

12. LITTLE, R. J. (1992). "Regression with missing X's: A review", *Journal of the American Statistical Association*, 87(420):1227-1237.

13. MINCER, J. (1974). *Schooling, experience and earnings.* Nueva York, Columbia University Press.

14. MUÑOZ, J. y RUEDA, M. (2009). "New imputation methods for missing data using quantiles", *Journal of Computational and Applied Mathematics*, 232:305-317.

15. NICOLETTI, C. y PERACCHI, F. (2006). "The effects of income imputation on microanalyses: Evidence from the European Community Household Panel", *Journal of the Royal Statistical Society*, 169(3):625-646.

16. OLINSKY, A., CHEN, S. y HARLOW, L. (2003). "The comparative efficacy of imputation methods for missing data in structural equation modeling", *European Journal of Operational Research*, 15:53-79.

17. ROBINS, J. M. y WANG, N. (2000). "Inference for imputation estimators", *Biometrika*, 87(1):113-124.

18. RUBIN, D. B. (1996). "Multiple imputation after 18 + years", *Journal of the American Statistical Association*, 91(434):473-489.

19. RUBIN, D. B. (2003). "Discussion on multiple imputation", *International Statistical Review*, 71(3):619-625.

20. RUBIN, D. B. y SCHENKER, N. (1986). "Multiple imputation for interval estimation from simple random samples with ignorable nonresponse", *Journal of the American Statistical Association*, 81(394):366-374.

21. SANDE, I. G. (1982). "Imputation in surveys: Coping with reality", *The American Statistician*, 36(3):145-152.

22. SCHENKER, N. y TAYLOR, J. (1996). "Partially parametric techniques for multiple imputation", *Computational Statistics and Data Analysis*, 22:425-446.

23. WALCZAK, B. y MASSART, D. (2001). "Dealing with missing data: part II", *Chemometrics and Intelligent Laboratory Systems*, 58:29-42.

24. ZHANG, P. (2003). "Multiple imputation: Theory and method", *International Statistical Review*, 71(3):581-592.

# Anexo 1

**Cuadro A1.1.** Ganancias e ingresos medios para estratos 1 y 2

| Método | Ganancia media | | Ingreso medio | |
|---|---|---|---|---|
| | Datos disponibles | 10% menos | Datos disponibles | 10% menos |
| Eliminación del caso | 422.608 (2.513,77) | 422.601 (2.539,40) | 655.576 (2.608,76) | 656.083 (2.786,91) |
| Media no condicionada | 415.871 (2.315,80) | 416.712 (2.475,86) | 656.228 (2.515,94) | 655.969 (2.640,32) |
| Imputación por regresión estocástica | 417.696 (2.366,83) | 706.751 (5.170,39) | 656.044 (2.549,85) | 820.270 (8.063,29) |
| *Hot-deck* | 422.946 (2.584,26) | 422.273 (2.519,10) | 655.434 (2.646,08) | 656.538 (2.743,73) |
| *Hot-deck* con regresión | 422.840 (2.461,81) | 422.673 (2.699,34) | 656.092 (2.630,92) | 656.305 (2.745,67) |
| Imputación múltiple nmv | 425.414 (740,52) | 424.929 (746,08) | 660.646 (565,49) | 659.262 (826,70) |
| Imputación múltiple ec. encadenadas | 422.223 (982,10) | 423.908 (1.055,85) | 657.374 (1.055,14) | 656.564 (1.106,01) |

*Fuente*: cálculos propios. Errores estándar en paréntesis.

**Cuadro A1.2.** Ganancias e ingresos medios para estratos 3 y 4

| Método | Ganancia media | | Ingreso medio | |
|---|---|---|---|---|
| | Datos disponibles | 10% menos | Datos disponibles | 10% menos |
| Eliminación del caso | 787.607<br>(7.937,21) | 791.032<br>(8.469,15) | 1.147.913<br>(8.438,47) | 1.144.520<br>(8.760,22) |
| Media no condicionada | 749.238<br>(7.314,39) | 744.154<br>(7.684,19) | 1.127.469<br>(8.094,70) | 1.127.403<br>(8.688,97) |
| Imputación por regresión estocástica | 796.717<br>(7.953,93) | 725.328<br>(5.337,05) | 1.169.654<br>(8.338,69) | 825.654<br>(8.082,14) |
| *Hot-deck* | 786.077<br>(8.440,02) | 790.089<br>(8.944,03) | 1.148.430<br>(8.346,16) | 1.144.313<br>(9.159,15) |
| *Hot-deck* con regresión | 787.791<br>(7.659,49) | 789.946<br>(9.037,17) | 1.152.564<br>(8.718,12) | 1.147.192<br>(8.644,93) |
| Imputación múltiple nmv | 796.195<br>(2.341) | 800.125<br>(2.507,27) | 1.158.798<br>(2.483,86) | 1.157.830<br>(2.482,30) |
| Imputación múltiple ec. encadenadas | 788.945<br>(3.105,19) | 786.828<br>(3.267,62) | 1.156.585<br>(3.394,18) | 1.150.120<br>(3.715,08) |

*Fuente*: cálculos propios. Errores estándar en paréntesis.

**Cuadro A1.3.** Ganancias e ingresos medios para estratos 5 y 6

| Método | Ganancia media | | Ingreso medio | |
|---|---|---|---|---|
| | Datos disponibles | 10% menos | Datos disponibles | 10% menos |
| Eliminación del caso | 2.064.456<br>(76.839,08) | 2.092.839<br>(83.522,07) | 3.029.668<br>(76.095,74) | 3.077.876<br>(82.345,3) |
| Media no condicionada | 1.778.456<br>(65.817,05) | 1.755.391<br>(69.504,33) | 2.792.291<br>(69.692,13) | 2.762.313<br>(73.566,2) |
| Imputación por regresión estocástica | 2.385.003<br>(78.075,32) | 729.115<br>(5.512,54) | 3.321.312<br>(75.510,22) | 826.102<br>(8.147,69) |
| *Hot-deck* | 2.057.286<br>(76.271,83) | 2.077.706<br>(74.835,04) | 3.011.352<br>(72.928,51) | 3.064.209<br>(77.693,73) |
| *Hot-deck* con regresión | 2.113.499<br>(79.135,50) | 2.141.090<br>(87.592,45) | 3.064.090<br>(72.072,55) | 3.115.182<br>(79.859,48) |
| Imputación múltiple nmv | 2.132.583<br>(22.300,55) | 2.179.114<br>(26.335,61) | 3.042.730<br>(21.563,82) | 3.103.126<br>(23.102,02) |
| Imputación múltiple ec. encadenadas | 2.085.513<br>(28.505,05) | 2.104.259<br>(32.173,65) | 3.047.740<br>(29.328,4) | 3.021.650<br>(33.108,19) |

*Fuente*: cálculos propios. Errores estándar en paréntesis.