4. ESTADÍSTICA OFICIAL

IMPUTATION IN THE SURVEY ON LIVING CONDITIONS

José María Méndez Martín*
Instituto Nacional de Estadística

Abstract

The Spanish "European Statistics on Income and Living Conditions" (EU-SILC) is one of the statistical operations that has been harmonised to EU standards. In this household survey there are two kinds of non-response: unit non-response (one or several household or individual questionnaires are missing) and item non-response (no questionaire is missing but some variables are). The main target variable is the total household income, which is defined as the aggregate of different income components. Components with missing values are imputed when they cannot be estimated with the help of other variables or other information in the questionnaire of the current or previous surveys. The procedure applied to the data preserves the variability of the variables and the correlations between them. The statistical software used for imputation is the IVEware. The IVEware implements a multivariate model involving a multiple regression sequence where imputation is carried out variable by variable generating draws from the predictive distribution specified by the regression model. An iterative imputation scheme is used, updating previous imputed values in order to better preserve the correlation among variables.

Keywords: EU-SILC, imputation, item non-response, IVEware.

1. Introducción

La Encuesta de Condiciones de Vida (ECV) es una encuesta anual del tipo panel rotatorio que recoge variables relacionadas con los ingresos y las condiciones de vida de los hogares. Se inició en 2004 y es una operación estadística armonizada de ámbito europeo. En INE (2007) se detalla la metodología de dicha encuesta.

Cada hogar que participa en la encuesta ha de cumplimentar un cuestionario de hogar y un cuestionario individual para cada persona adulta. Las áreas que abarca el cuestionario de hogar son: componentes de ingresos que son propios de la unidad hogar (ayudas sociales a la vivienda o al hogar, rentas de la propiedad, etc.), vivienda, equipamiento del hogar, etc. El cuestionario de personas adultas recoge información sobre componentes de ingresos que son propios de la unidad persona (salarios, renta de autónomos, prestaciones sociales, etc.), situación en la actividad, trabajo actual, estudios que realiza, máximo nivel de estudios alcanzado, salud, etc. En consonancia con los dos tipos de cuestionarios también hay dos factores de elevación, uno de hogares y otro de adultos, empleándose en las estimaciones uno u otro dependiendo de la unidad considerada.

En la ECV se pueden presentar tres tipos de falta de respuesta: total, individual y parcial. Cada tipo de falta de respuesta se trata de forma diferente. La falta de respuesta total consiste en que un hogar no colabora y, por tanto, no se recoge ningún cuestionario. El tratamiento de este tipo de falta de respuesta se realiza recalculando los factores de elevación de los hogares que sí han respondido. La falta de respuesta individual consiste en que en un hogar colaborador no se obtienen algunos de los cuestionarios de los adultos del hogar. La falta de respuesta individual se corrige ajustando los factores de elevación de los adultos de los que sí se obtiene cuestionario. Sin embargo este tipo de falta de respuesta también afecta a las variables del hogar que se obtienen por agregación de variables de adultos (por ejemplo los ingresos del trabajo por cuenta ajena del hogar, que es la suma de los ingresos por cuenta ajena de los miembros del hogar). Por ello, se calcula, para cada hogar, un factor multiplicador que intenta corregir la renta que falta de los cuestionarios individuales no cumplimentados. Fi-

*Corresponding Author. E-mail: jmmendez@ine.es

nalmente, la falta de respuesta parcial consiste en que no falta ningún cuestionario, pero algunas variables no están debidamente cumplimentadas. Para tratar este tipo de falta de respuesta se aplica el método de regresión secuencial multivariante, utilizando el software IVE, desarrollado por el *Institute for Social Research* de la Universidad de Michigan.

A partir de la información recogida en la encuesta se obtienen unos ficheros de trabajo que reorganizan la información, haciéndola más manejable para el investigador. En estos ficheros se trata la falta de respuesta en sus distintas modalidades. En particular, en el caso de la falta de respuesta parcial de las variables relacionadas con los ingresos, se lleva a cabo una imputación.

2. Imputación de los ingresos en la ECV

Los ingresos totales del hogar se calculan a partir de sus componentes. No sería adecuado dar por perdidos todos los ingresos cuando falta sólo algún componente. Por tanto resulta esencial realizar las imputaciones de componentes de los ingresos en aquellos casos donde razonablemente se puede llevar a cabo.

Un Reglamento de la Comisión Europea sobre aspectos de trabajo de campo y procedimientos de imputación da algunas recomendaciones. En concreto especifica que "el procedimiento aplicado a los datos debería preservar la variabilidad de las variables y la correlación entre ellas. Los métodos que incluyan un «componente de error» en los valores imputados serán preferibles a los que imputen simplemente un valor determinado. Los métodos que tengan en cuenta la estructura de las correlaciones (u otras características de la distribución conjunta de las variables) serán preferibles al enfoque marginal o univariante.". Como se verá más adelante estos principios están contemplados en el método de imputación de la ECV, que sigue un modelo similar al utilizado por Eurostat en el Panel de Hogares de la Unión Europea (EUROSTAT 2001).

La imputación en la ECV se lleva a cabo después de la depuración de los datos, que ha de ser realizada minuciosamente, tanto a nivel de microdatos como en los resultados agregados. La depuración se inicia en la entrevista personal durante la fase de recogida de la información, ya que la aplicación informática tiene incorporados una serie de controles que detectan posibles errores e inconsistencias. En los Servicios Centrales se aplican unos controles exhaustivos desarrollados por la unidad promotora del INE y también unos programas de chequeo de Eurostat. La depuración manual se realiza mediante una aplicación que visualiza el contenido de los datos recogidos en los cuestionarios de los diferentes años.

La depuración permite corregir una parte de la falta de respuesta parcial. También se detectan y eliminan los "outliers" antes de realizar las imputaciones. Con este fin se fijan unos límites mediante observación de la distribución de los importes extremos declarados y se ponen como valores perdidos los que no estén dentro de estos límites.

A partir del segundo año de producción de la ECV, las imputaciones de los ingresos se pueden realizar utilizando los datos disponibles del año anterior. Aprovechando la componente longitudinal de la encuesta, cuando un importe falta en el año t y no falta en el año t-1, se imputa el importe de t multiplicando por un valor el importe de t-1.

Cuando no es posible obtener el valor del ingreso en la depuración o con datos disponibles del año anterior, se realiza la imputación a partir de la información disponible. En algunos casos se dispone del tramo en el que está situado el importe que falta. En este caso se imputará el importe con la restricción del intervalo proporcionado. Cuando no se dispone ni siquiera del tramo entonces se imputa el importe con una restricción construida a partir de los percentiles 10 y 90 de la distribución de los respondientes. La imputación en esta fase se realiza aplicando el método de imputación de regresión secuencial multivariante, imputando solamente las variables de ingresos.

3. Método de imputación de regresión secuencial multivariante

En la imputación se utiliza una técnica regresión multivariante basada en unos modelos que implementa el software IVE. Es un procedimiento de imputación general multivariante que puede tratar datos con una estructura compleja y que permite añadir residuos aleatorios. La descripción completa del método se puede encontrar en Raghunathan, Lepkowski, Van Hoewyk y Solenberger (2001) y en las referencias que contiene esta publicación. Es posible descargar el software de imputación desde la

página web "www.isr.umich.edu/src/smp/ive".

El procedimiento se basa en crear imputaciones por medio de una secuencia de regresiones. Se pretende recoger la correlación de todas las variables, tanto las completas como de las que tienen valores perdidos. El programa permite distintos tipos de regresiones (lineal, logística, logística generalizada y de Poisson). Sin embargo, en el caso de la ECV solamente se utiliza regresión lineal para imputar los ingresos, previa aplicación a éstos de una transformación logarítmica. Las variables explicativas pueden ser discretas, continuas o binarias.

En el modelo de regresión la aleatoriedad en la estimación se introduce por dos vías: por una parte se considera el término correspondiente al residuo aleatorio y por otra se incorpora una perturbación en los coeficientes de regresión estimados. La distribución que se obtiene con este enfoque se puede consultar en Gelman, Carlin, Stern y Rubin (1995).

Con este software es posible considerar el intervalo en el que está el valor imputado, es decir, se puede imponer un valor mínimo y máximo al valor imputado para cada registro. Este aspecto es importante en la imputación de ingresos ya que en el cuestionario, para muchos componentes de la renta, se solicita en primer lugar el importe exacto y, cuando éste se desconoce, se solicita el intervalo.

El procedimiento general de imputación sigue la siguiente estrategia. Supongamos que X es la matriz de datos construida con todas las variables completas (es decir, que no tienen ningún valor perdido). X se compone de variables explicativas como sexo, edad, región, nivel de estudios y otras que pueden ser continuas (ya transformadas si es necesario), binarias o categóricas.

Por otra parte sean Y1 , Y2 , Y3 , Y4 , ... , Yk las variables que tienen valores perdidos. Suponemos que las variables Y están ordenadas de menor a mayor falta de respuesta. En total se tienen las variables:

X1 , X2 , X3 , ... , Y1 , Y2 , Y3 , Y4 , ... , Yk

En la iteración inicial se imputa según las siguientes distribuciones condicionales:

[Y1 /X]
[Y2 /X,Y1]
...
[Yk /X,Y1 ... Yk-1]

donde [A /B] denota la distribución de A condicionada por B.

Por tanto, en esta iteración inicial se empieza haciendo la regresión de la variable con menos falta de respuesta Y1 sobre las variables explicativas X. Una vez obtenida una "predicción" de Y1 se incorpora esta variable a la matriz X de las variables completas y se obtiene la matriz [X,Y1]. A continuación se repite el proceso con la siguiente variable (Y2) tomando como variables explicativas [X,Y1] Se repite el proceso con Y3 , Y4 , ... , Yk hasta que todas las variables han sido imputadas.

Una vez que se ha realizado esta iteración de regresiones se tiene una primera imputación de todos los valores perdidos. Esta imputación mantiene la estructura multivariante de las variables imputadas con X y con algunas de las Y. Así en la imputación de Ym se tienen en cuenta las X y las variables Y1 ... Ym-1. Sin embargo las variables Ym+1 ... Yk no se han tenido en cuenta en la imputación de Ym.

En las iteraciones siguientes, utilizando la estrategia de esta iteración inicial, lo que se hace es repetir esta iteración pero incluyendo como variables explicativas todas las variables, ya que ahora no hay valores perdidos en ninguna de ellas.

Iteración 2:
[Y1 /X,Y2 ... Yk]
[Y2 /X,Y1,Y3,Y4 ... Yk]
...
[Yk /X,Y1 ... Yk-1]
...
...
Iteración "n":
[Y1 /X,Y2 ... Yk]
[Y2 /X,Y1,Y3,Y4 ... Yk]
...
[Yk /X,Y1 ... Yk-1]

En cada iteración se actualizan las imputaciones hechas en la iteración anterior. Así se obtienen actualizaciones que van recogiendo de una manera más completa la estructura de correlaciones del conjunto de las variables. Este proceso se detiene cuando se alcanza el número de iteraciones especificado por el usuario.

Referencias

[1] EUROSTAT (2001). Imputation of Income in the ECHP. Doc. Pan 164/2001.

[2] Gelman, Carlin, Stern y Rubin (1995). *Bayesian Data Analysis*, Chapman and Hall, London.

[3] INE (2007). Encuesta de Condiciones de Vida. Metodología.

[4] Raghunathan, Lepkowski, Van Hoewyk y Solenberger (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models, Survey Methodology. *Statistics Canada*, **27(1)**.