Abstract

This paper describes the methods used for imputation of the first wave of the Spanish Survey of Household Finances (EFF). It explains the motivation for using multiple imputation and describes its specific features, such as the use of the shadow values that flag each variable of the questionnaire, the different types of covariates used in the imputation models and the means of evaluating both the imputed values and the convergence of the imputation process.

Keywords: Household wealth survey, imputation methods.

JEL codes: D10, C81.

1 Introduction

In 2001 the Banco de España decided to launch a survey of Spanish household finances (*Encuesta Financiera de las Familias*, hereafter EFF) with similar features to those carried out in other countries, such as the *Survey of Consumer Finances* (SCF) in the US and the *Survey of Household Income and Wealth* (SHIW) in Italy.[1] This survey, whose first wave corresponds to 2002, collects information about households' holdings in real and financial assets, their debts, their different sources of income and their consumption. It provides microdata to study households' consumption, saving and investment decisions in Spain.[2]

By its own nature, non-response rates are typically high in this type of survey. Non-response takes place in two ways. First, there is a high percentage of households that do not want to participate in the survey or that cannot be located by the interviewers. Since household wealth distribution is heavily skewed and some types of assets, mainly financial assets, are only held by a low percentage of the population, the EFF oversamples wealthy households. Tables 3 and 4 in Bover (2004) show that non-response rates are not random; the higher the household wealth stratum, the higher these rates are. The *survey non-response* also depends on other characteristics, such as the total household income quartile, geographical factors, municipality size, household social status and the kind of neighbourhood and building in which the household lives. The means of taking this problem into account in the EFF is to use weights adjusted by the non-response in order not to bias the potential analysis of the data, as explained by Bover (2004).

A second type of non-response is *item non-response*: some households do not answer all questions asked by the interviewer, due to different reasons such as a lack of understanding of questions, a lack of knowledge of the answers, and reluctance and unwillingness to disclose some information. This entails the existence of missing data in some parts of the questionnaire completed by households. Like survey non-response,

[1] Both the description and the methodology of the EFF are explained by Bover (2004).

[2] The microdata and the corresponding documentation are available on the Banco de España website (http://www.bde.es/estadis/eff/eff.htm).

item non-response is not random, since it usually follows a pattern that depends on household characteristics. Item non-response occurs in euro questions more often than it does in questions involving a discrete number of alternatives (e.g. yes/no questions about the ownership of a particular asset). In wealth surveys like the EFF, item non-response affects mostly variables of income, wealth, debt and values invested in each type of asset in a non-random way; the problem becomes more serious in surveys with an oversampling of the wealthy, such as the SCF and the EFF, since usually the richer the households, the higher the item non-response rates are. Accordingly, the results of all potential analyses based on the EFF, ignoring the presence of missing data and not taking the item non-response into account, can be misleading.

Nowadays, many researchers, (see, for instance, Korinek *et al.,* 2005, and Vazquez Alvarez *et al.,* 1999), are concerned about how the non-response affects their estimates, and they use different parametric or non-parametric techniques to deal with this issue and control for non-response. Moreover, irrespective of whether the item missingness is random or not, it is useful to provide imputations, since the deletion of the missing data would discourage studies of how households decide to invest in different types of assets, due to the small sample sizes after deletion. For these reasons, wealth surveys like the SCF and the EFF provide imputations of missing data, so that correct inferences may be made by the users.

This paper explains the methods used for the imputation of the first wave of the EFF. It has two main purposes: first, to explain the motivation for imputation and to review the imputation techniques used in the EFF, which are very close to those of the SCF; and secondly, to explain in some detail how the imputation models are actually specified as well as more practical issues encountered when imputing the EFF data. For imputing the data, we have been very fortunate to use the programs written by Arthur Kennickell for the Survey of Consumer Finances (SCF) multiple imputation and to benefit from his invaluable help and advice.[3]

[3]See Kennickell (1991, 1998) for a detailed description of the imputation methods of the Survey of Consumer Finances.

The structure of this paper is as follows: Section 2 explains further why imputation is useful in wealth surveys like the EFF, and Section 3 explains the motivation behind multiple imputation and describes the imputation procedure implemented. The remaining sections describe the specific features of the EFF data imputation: Section 4 explains the use of the shadow values that flag all variables of the questionnaire, Section 5 describes the covariates used and the functional forms of the imputation models, and Section 6 explains the means of evaluating both the imputed data and the convergence of the imputation process. Finally, Section 7 summarises the main conclusions of this paper.

2 Why imputation is useful

Until recently, the most widespread ways of dealing with missing data were to fill in missing values with means of the observed data ("fill-in with means"), to delete cases or observations that have missing values in at least one variable in the empirical model proposed by the data users ("complete-case analysis") and to replace the missing values by other predicted values using non-stochastic imputation methods that best fit the observed data.

However, as many authors like Little and Rubin (1987), Rubin (1987, 1996) and Schafer (1997) emphasise, the goal of imputing is not to replace missing data by those predicted values that best fit the variables of interest, but to preserve the characteristics of their distribution and the relationships between different variables. In this way, all potential analyses carried out with different statistics, not only means but also medians, percentiles, variances and correlations, are unbiased. For this reason, all the imputation methods and the ways of dealing with missing data mentioned above, such as "fill-in with means", "complete-case analysis" and non-stochastic imputation, are not suitable, since they do not preserve the distribution of the complete-data (i.e. the joint distribution of both the observed and missing data). Non-stochastic imputation and the method of "fill-in with means" make the distribution more peaked around the mean of the observed data and the variance underestimated.

Only imputation methods based on stochastic imputation like those used in the EFF can help preserve the distribution of the complete-data, since the imputed values are the result of adding a random number to the value predicted by the imputation model using a distribution also specified by the imputation model. In this way, the imputed data preserve the distribution of the complete data, not only the mean of the variables but also other distribution characteristics such as percentiles, variances, covariances, etc. Finally, all results based on "complete-case analysis" will be biased, due to the fact that this method ignores the fact that item non-response is not random in wealth surveys like the EFF.

However, as Rubin (1987, 1996) states, one single stochastic imputation does not take into account the uncertainty about the imputation model due to the fact that it treats the imputed value as if it was an actual one; we need to draw several imputed values to assess the uncertainty about the imputation. This is the reason why the EFF provides multiple imputations instead of one single stochastic imputation of the missing data; we do not want to provide one single imputation of the missing data that can be used as if they were the true data using complete-data econometric tools and forgetting that they are not really observed. As single imputation only provides one value, it does not reflect the uncertainty about both the imputation and the non-response models, and it underestimates the standard errors of all statistics used. The EFF imputes five values for each missing item of each household observation, whereby these five values may differ depending on the degree of uncertainty about the imputation model.

Finally, Rubin (1996) gives the two most important reasons why the database constructors should provide imputations of the missing data, instead of letting the potential users impute their own data. First, the potential users of the data may neither know the modelling and the tools required to impute the missing data nor devote enough time, effort and computation requirements to obtain acceptable imputations. Second, to preserve confidentiality, the potential users of the data will not receive information about some relevant variables that are both major determinants of the non-response and very good predictors of the imputed income and wealth variables. In the case of the EFF, random wealth strata indicators and location variables will not be available for the potential users; these variables are not only very good predictors of many variables, but they are also important factors of item non-response. Moreover, due to confidentiality reasons, the potential users of the imputation models will not have some key covariates available for satisfying the main assumption made by many imputation methods like that carried out here, which is called *missing at random*. As explained in Section 3, this assumption asserts that the distribution of the non-response conditional on some

covariates and on the complete-data (the observed and the missing data) is independent of the missing data. Thus, the lack of some key covariates to satisfy this assumption will make the final data users' own imputations not acceptable. However, if the potential users of the data want to carry out more complex imputation methods or to deal with missing data using maximum likelihood models or other approaches, they can do so as all survey variables are suitably flagged in such a way that the users know both the nature of the data (whether they are observed or missing) and the reason for item missingness.

3 General features of multiple imputation in the EFF

3.1 Assumptions and theoretical framework

Missing at random The imputation of the EFF data is done assuming *missing at random* (MAR) as explained by Rubin (1976). This assumption implies that the conditional distribution of the household responses, R, only depends on the observed data, Y_{obs}, but not on the missing data, Y_{mis}. Let Y be the $N \times K$ matrix formed by the K variables available for each of the N participants in the EFF survey; this matrix can be decomposed into two matrices, Y_{obs} and Y_{mis}, containing the observed and the missing data separately, so we have $Y = (Y_{obs}\, Y_{mis})$. The non-response model depends on the parameter vector, ϕ. The MAR assumes:

$$P\left(R \mid Y, \phi\right) = P\left(R \mid Y_{obs}, \phi\right); \quad Y = (Y_{obs}\, Y_{mis}) \tag{1}$$

Ignorable missing data mechanism As Rubin (1976) and Cameron and Trivedi (2005) explained, another assumption made by the imputation methods like that of the SCF and the EFF is that the missing data mechanism is *ignorable*. This occurs when the household response is missing at random (MAR) and the parameters of the missingness mechanism, ϕ, are distinct from θ, the parameters of our imputation model of the missing data, $P\left(Y_{mis} \mid Y_{obs}, \theta\right)$ (i.e. ϕ and θ are independent). If so, we do not need to specify the non-response model, $P\left(R \mid Y_{obs}, \phi\right)$, for imputing missing data.

Stochastic imputation In large surveys like the EFF (containing around 3,000 variables), the pattern of item missingness may be very different across household observations, so the number of variables to be imputed and the variables included in the two vectors defined for each household i, $Y_{obs,i}$ and $Y_{mis,i}$, are specific to each household.[4]

[4] That is to say, the variables included in the vectors, $Y_{obs,i}$ and $Y_{mis,i}$, and their dimension are different across households, and they depend on the pattern of item missingness across households. If K is the number of variables included in the survey (i.e. the number of columns of matrix Y), we generally observe for two different households, i and j, the following: no. of variables in $Y_{obs,i} \neq$ no. in $Y_{obs,j}$, no. of variables in $Y_{mis,i} \neq$ no. in $Y_{mis,j}$, no. of variables in $Y_i =$ no. in $Y_j = K$, and the

At the beginning of the imputation process, the original sample of N households, $Y = (Y_{obs}\ Y_{mis})$ has the following structure:[5]

$$\begin{matrix} Y_{obs,1} & Y_{mis,1} \\ Y_{obs,2} & Y_{mis,2} \\ \vdots & \vdots \\ Y_{obs,N} & Y_{mis,N} \end{matrix} \tag{2}$$

We impute missing data stochastically for preserving the characteristics of the data distribution. Suppose that the imputation model we propose for the variable of interest, say y, is as follows:

$$y = X\beta + u, \quad u \mid X \sim N\left(0, \sigma^2 I\right) \tag{3}$$

Stochastic imputation replaces the missing value, y_{mis}, by its best linear predicted value, $X\widehat{\beta}$, plus a random draw, $\widehat{u}$, coming from a normal distribution function with the following variance-covariance matrix:

$$\begin{gathered} \widehat{y}_{mis} = X\widehat{\beta} + \widehat{u}, \quad \widehat{u} \mid X \sim N\left(0, \widehat{\sigma}^2 I\right) \\ \widehat{\beta} = \left(X'X\right)^{-1}\left(X'y\right); \quad \widehat{\sigma}^2 = \frac{1}{n}\left(y'y - y'X\left(X'X\right)^{-1}X'y\right) \end{gathered} \tag{4}$$

The matrix X has $n \times k$ dimension and contains k covariates that the model includes for imputing the variable of interest, y; n denotes the subsample size of respondents over which the imputation model is estimated. If X is properly constructed, stochastic im-

number of variables in both $Y_{obs,l}$ and $Y_{mis,l}$ is equal to K, for $l = 1, \ldots, N$.

[5] If the number of households were 2 and the number of variables in the survey 3, the sample structure would be:

$$(Y_{obs}\, Y_{mis}) = \begin{pmatrix} y_{obs,11} & y_{obs,12} & y_{obs,13} & y_{mis,11} & y_{mis,12} & y_{mis,13} \\ y_{obs,21} & y_{obs,22} & y_{obs,23} & y_{mis,21} & y_{mis,22} & y_{mis,23} \end{pmatrix}$$

One example of this structure is the following:

$$(Y_{obs}\, Y_{mis}) = \begin{pmatrix} 1 & \cdot & 4 & \cdot & 3 & \cdot \\ 5 & 9 & \cdot & \cdot & \cdot & 7 \end{pmatrix}$$

The matrix Y_{mis} contains the missing information in the survey, which is unobserved and must be imputed. Another equivalent notation of the survey structure is written in terms of the number of survey variables also separating observed and missing data, as follows:

$$(Y_{obs}\, Y_{mis}) = (y_{obs,1}\, y_{obs,2}\, y_{obs,3}\, y_{mis,1}\, y_{mis,2}\, y_{mis,3})$$

This notation will be used later, when describing the first iteration of the imputation process.

putation preserves the characteristics of the distribution among the variable of interest, y, and other variables of the survey. This is due to the fact that the randomisation does not make the distribution of the complete data more peaked around the mean of the observed data nor underestimate the variance, unlike other methods, such as "fill-in with means" and non-stochastic imputation.

Multiple imputation The EFF provides multiply imputed values of the missing data instead of one single value, since we want to reflect the uncertainty about the imputation and non-response models. Single stochastic imputation only takes into account the within-imputation variance of the statistics constructed using a single imputed data set, but ignores the between-imputation variance due to the uncertainty about the true imputation and non-response models. The EFF provides m plausible values of the missing data; m is equal to 5 in the final sample provided for the potential users, but it takes a different value in each iteration of the imputation process, as explained in Section 3. Thus, for each missing value of each variable k, $y_{mis,ik}$, we have m imputed values, $\widehat{y}^{(1)}_{mis,ik}, \cdots, \widehat{y}^{(m)}_{mis,ik}$, which will differ between themselves, depending on the degree of uncertainty we have about the imputation model. After imputing all variables of the survey, we have m complete-data sets, where the observed data of household i, $Y_{obs,i}$, are repeated in each data set and its missing data, $Y_{mis,i}$, are replaced by each one of the m imputed values, $\widehat{Y}^{(s)}_{mis,i}$, $s = 1, 2, \ldots, m$. As a result, the final data sample has the following structure:

$$\begin{array}{lcc}
\text{Data set 1:} & \left.\begin{array}{cc} Y_{obs,1} & \widehat{Y}^{(1)}_{mis,1} \\ Y_{obs,2} & \widehat{Y}^{(1)}_{mis,2} \\ \vdots & \vdots \\ Y_{obs,N} & \widehat{Y}^{(1)}_{mis,N} \end{array}\right\} \rightarrow \widehat{Q}^{(1)}, U^{(1)} \\
\vdots & & \\
\text{Data set } m\text{:} & \left.\begin{array}{cc} Y_{obs,1} & \widehat{Y}^{(m)}_{mis,1} \\ Y_{obs,2} & \widehat{Y}^{(m)}_{mis,2} \\ \vdots & \vdots \\ Y_{obs,N} & \widehat{Y}^{(m)}_{mis,N} \end{array}\right\} \rightarrow \widehat{Q}^{(m)}, U^{(m)}
\end{array} \tag{5}$$

Let $\widehat{Q}^{(s)}$ and $U^{(s)}$ be the statistic vector of interest and its estimated variance-covariance matrix from the complete data set s. Following Rubin (1976, 1987) and Cameron and Trivedi (2005), one possible way of treating multiply imputed data sets is to carry out the empirical analysis separately in each complete-data set, and then to combine these estimands by averaging over the m multiply imputed data sets, as follows:

$$\begin{aligned} \overline{Q} &= \frac{1}{m}\sum_{s=1}^{m}\widehat{Q}^{(s)}; \quad \overline{U} = \frac{1}{m}\sum_{s=1}^{m}U^{(s)} \\ B &= \frac{1}{m-1}\sum_{s=1}^{m}\left(\widehat{Q}^{(s)}-\overline{Q}\right)\left(\widehat{Q}^{(s)}-\overline{Q}\right)' \\ T &= \overline{U}+\left(1+\frac{1}{m}\right)B \end{aligned} \tag{6}$$

The estimated variance-covariance matrix, T, of the combined statistic vector, $\overline{Q}$, takes into account the within-imputation variability, $\overline{U}$, and the between-imputation variability, B. The latter is due to the uncertainty about the imputation and is ignored by single imputation methods; this is the reason why single imputation underestimates the variance of the statistics. Equation (6) shows that, the higher the value of m, the lower the loss of efficiency due to imputation is in T; Rubin (1976) shows how the loss of efficiency varies depending on both the number of multiply imputed values, m, and the fraction of missing data. For the most common values of the fraction of missing information (normally less than 30%), as the number of multiple imputations increases from 5, the efficiency gain is very low and it does not offset the effort in terms of time, storage and computational requirements. Therefore, in the final data sample, the number of multiple imputations is 5, as in other surveys like the SCF.

3.2 Description of the imputation procedure

Iterative and sequential imputation process The imputation procedure is based on the data augmentation algorithm (see Tanner and Wong, 1987) and Markov chain Monte Carlo method, and has a sequential and iterative structure (see Schafer 1997), as follows:

$$\begin{gathered}\text{I-step (Imputation step):} \quad \widehat{Y}_{mis}^{(t)} \sim P\left(Y_{mis} \mid Y_{obs}, \widehat{\theta}^{(t-1)}\right) \\ \text{P-step (Posterior step):} \quad \widehat{\theta}^{(t)} \sim P\left(\theta \mid Y_{obs}, \widehat{Y}_{mis}^{(t)}\right) \\ \left(\widehat{Y}_{mis}^{(1)}, \widehat{\theta}^{(1)}\right), \left(\widehat{Y}_{mis}^{(2)}, \widehat{\theta}^{(2)}\right), \cdots \xrightarrow{d} P\left(Y_{mis}, \theta \mid Y_{obs}\right) \end{gathered} \tag{7}$$

Each iteration t consists of two steps, the first step is called *imputation step*, here the missing data are imputed, $\widehat{Y}_{mis}^{(t)}$, using the previous-iteration estimates, $\widehat{\theta}^{(t-1)}$, of the parameters that come from the missing data distribution conditional on observed data. The second step is called *posterior step*, and it estimates the parameters of the complete data distribution, $\widehat{\theta}^{(t)}$, coming from the imputation model and using the imputations of the first step, $\widehat{Y}_{mis}^{(t)}$, as if the imputed values were really known or observed. Then, we start another iteration, $t+1$, repeating both steps until the convergence of the process.

Following the SCF imputation programs (see Kennickell, 1991), this iterative process is carried out from the second iteration of the imputation process; the first iteration is slightly different from that explained above, and it would be identical to the rest if we only had to impute one variable. However, in large surveys like the SCF and the EFF, a high percentage of the survey variables must be imputed; so, within one iteration, these two steps (I and P-steps) are repeated sequentially for each one of the survey variables having missing information.

First iteration of the imputation process The order in which the variables are imputed sequentially matters, since the imputed values of one variable are used to impute the remaining variables in the first iteration of the imputation process, and these imputed

data are treated as if they were really observed for imputing the remaining variables in the iteration. That is, if the survey has K variables, the way of imputing in the first iteration is as follows:

$$\text{I-step} \quad : \quad \begin{cases} \widehat{y}^{(1)}_{mis,1} \sim P\left(y_{mis,1} \mid Y_{obs}, \widehat{\theta}^{(0)}_1\right) \\ \widehat{y}^{(1)}_{mis,2} \sim P\left(y_{mis,2} \mid Y_{obs}, \widehat{y}^{(1)}_{mis,1}, \widehat{\theta}^{(0)}_2\right) \\ \widehat{y}^{(1)}_{mis,3} \sim P\left(y_{mis,3} \mid Y_{obs}, \widehat{y}^{(1)}_{mis,1}, \widehat{y}^{(1)}_{mis,2}, \widehat{\theta}^{(0)}_3\right) \\ \quad \vdots \\ \widehat{y}^{(1)}_{mis,K} \sim P\left(y_{mis,K} \mid Y_{obs}, \widehat{y}^{(1)}_{mis,1}, \widehat{y}^{(1)}_{mis,2}, \ldots, \widehat{y}^{(1)}_{mis,K-1}, \widehat{\theta}^{(0)}_K\right) \end{cases}$$

$$\text{P-step} \quad : \quad \widehat{\theta}^{(1)} \sim P\left(\theta \mid Y_{obs}, \widehat{Y}^{(1)}_{mis}\right)$$

$$\theta' \quad = \quad (\theta'_1 \, \theta'_2 \ldots \theta'_K) \tag{8}$$

The parameter vector, θ, collects all the parameter subvectors, $\theta_i \; i = 1, \ldots K$, implied by the imputation model of each variable. The imputed values of the first variable, $\widehat{y}^{(1)}_{mis,1}$, are used to impute the second and the remaining variables; the imputed values of the second variable, $\widehat{y}^{(1)}_{mis,2}$, are used to impute the third and the remaining variables, and so on. Thus, the choice of the order in which the variables are imputed sequentially within the same iteration is not trivial; once we impute one variable, we have to update the missing values of all covariates that are derived from the imputed variable and that take part in the imputation models of the remaining variables.

In the EFF data, we start imputing those variables not having a high percentage of missing information and those variables that are considered to be very good predictors of the remaining variables to be imputed. Once all variables are already imputed, the parameter vector of the imputation model, θ, is estimated for imputing missing information in the next iteration. This sequential and iterative process continues until the sixth iteration, when the missing data and the parameter values of the imputation models are expected to converge in distribution.

The starting values of the parameters in the first iteration, $\widehat{\theta}^{(0)}$, correspond to the estimates of the imputation model of each variable, but using the subsample of both the

observed data and the values of the missing data previously imputed within the first iteration:

$$\begin{aligned}
\widehat{\theta}_1^{(0)} &\sim P\left(\theta_1 \mid Y_{obs}\right) \\
\widehat{\theta}_2^{(0)} &\sim P\left(\theta_2 \mid Y_{obs}, \widehat{y}_{mis,1}^{(1)}\right) \\
&\vdots \\
\widehat{\theta}_K^{(0)} &\sim P\left(\theta_K \mid Y_{obs}, \widehat{y}_{mis,1}^{(1)}, \widehat{y}_{mis,2}^{(1)}, \ldots, \widehat{y}_{mis,K-1}^{(1)}\right)
\end{aligned} \tag{9}$$

Therefore, in the I-step of the first iteration, we first estimate the initial value of θ_1, $\widehat{\theta}_1^{(0)}$, from an empirical model with the same covariates as those included in the imputation model of the first variable to be imputed, $y_{mis,1}$, but only using the subsample of the observed data, $P\left(\theta_1 \mid Y_{obs}\right)$. When we impute this first variable stochastically, $\widehat{y}_{mis,1}^{(1)}$, from the distribution $P\left(y_{mis,1} \mid Y_{obs}, \widehat{\theta}_1^{(0)}\right)$, we estimate the initial values of the parameters of the imputation model of the second variable to be imputed, $\widehat{\theta}_2^{(0)}$, using the empirical model that includes the same covariates as its imputation model, $P\left(\theta_2 \mid Y_{obs}, \widehat{y}_{mis,1}^{(1)}\right)$, but restricting the sample to both the observed data and the imputed data of the first variable, denoted as Y_{obs} and $\widehat{y}_{mis,1}^{(1)}$, respectively. This sequential process continues until imputing the last variable of the survey, $\widehat{y}_{mis,K}^{(1)}$, then we implement the P-step as in equation (8), and we start the second and the remaining iterations following the two steps described in equation (7).[6]

Following the SCF imputation process, the number of multiply imputed values for the missing data increases with the iteration number; in this way, we only impute one value ($m = 1$) in the first iteration, three multiple values ($m = 3$) in the second iteration, and five values ($m = 5$) from the third to the last iteration.

[6] At the end of the imputation process (in the sixth iteration), the P-step is not carried out; we do not need to obtain estimates of the parameters for the next iteration. In the last iteration, we are only interested in imputing the missing data, $\widehat{Y}_{mis}^{(6)}$, i.e. the multiply imputed data of the final sample provided for the potential users.

Functional form of the imputation models Concerning the imputation model, we distinguish three different types of variables using the SCF multiple imputation macro programs: continuous, binary and categorical variables.

Continuous variables Continuous variables are imputed stochastically using linear regression models; if y is the vector with dimension $n \times 1$ containing the household observations of the variable of interest to be imputed and if X is the matrix with dimension $n \times k$ that includes the values of the k covariates involved by the imputation model, missing information on continuous variables is imputed as follows:

$$\begin{gathered} y = X\beta + u, \quad u \mid X \sim N\left(0, \sigma^2 I\right) \\ \widehat{y}_{mis} = X\widehat{\beta} + \widehat{u},\ \widehat{u} \mid X \sim N\left(0, \widehat{\sigma}^2 I\right) \\ \widehat{\beta} = \left(X'X\right)^{-1} X'y, \quad \widehat{\sigma}^2 = \frac{1}{n}\left(y'y - y'X(X'X)^{-1}X'y\right) \end{gathered} \tag{10}$$

We cannot estimate imputation models by maximum likelihood, non-parametrically or non-linearly due to the enormous costs in terms of effort and time; we have very different patterns of item missingness among the household covariates in large surveys like the EFF and the SCF. We restrict the imputation of continuous variables to linear regression models such as that in equation (10), since we can accommodate very easily a huge number of different patterns of item missingness across households in the first iteration, as if we implement different linear imputation models for each observation i depending on the non-missing covariates in X_i. For example, if the imputation model of the variable of interest, y, is specified to have three covariates in the matrix, $X = (x_1\ x_2\ x_3)$, the missing values of households having observed data in the three covariates are imputed using the following estimated parameters of the imputation model:

$$\widehat{\beta} = \begin{pmatrix} x_1'x_1 & x_1'x_2 & x_1'x_3 \\ x_2'x_1 & x_2'x_2 & x_2'x_3 \\ x_3'x_1 & x_3'x_2 & x_3'x_3 \end{pmatrix}^{-1} \begin{pmatrix} x_1'y \\ x_2'y \\ x_3'y \end{pmatrix} \tag{11}$$

However, if the second covariate, x_{i2}, is missing for household i, the imputation model of the variable, $y_{mis,i}$, should be only based on the other non-missing covariates, x_{i1} and x_{i3}, and the parameter estimates of this new imputation model can be obtained easily by removing the rows and columns of matrices, $(X'X)^{-1}$ and $X'y$, referring to the missing second covariate in equation (11), as follows:

$$\begin{aligned}
\widehat{\beta}_i &= \begin{pmatrix} x_1'x_1 & x_1'x_3 \\ x_3'x_1 & x_3'x_3 \end{pmatrix}^{-1} \begin{pmatrix} x_1'y \\ x_3'y \end{pmatrix} \\
\widehat{y}_{mis,i} &= x_{i1}\widehat{\beta}_1 + x_{i3}\widehat{\beta}_3 + \widehat{v}_i,\ \widehat{v}_i \mid x_1, x_3 \sim N\left(0, \widehat{\omega}^2\right) \\
\widehat{\omega}^2 &= \frac{1}{n}\left[y'y - y' \begin{pmatrix} x_1 & x_3 \end{pmatrix} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_3 \end{pmatrix} \right]
\end{aligned}$$

Consequently, using linear regression models, we take advantage of reshaping easily the matrices, $(X'X)^{-1}$ and $X'y$, involved in the estimation of the imputation model parameters in equation (10). For imputing the missing value of household i, $\widehat{y}_{mis,i}$, we reshape these matrices rapidly depending on the particular pattern of item missingness in the covariates, X_i. This property of the linear regression models is very useful for accommodating a huge number of different patterns of item missingness in the first iteration of the imputation process (see equations (8) and (9)), when we do not have any previously imputed data for missing information and when the number of covariates is very high (from 100 to 200 in most imputation models). In this way, we can save much effort and time and many computational resources in the imputation process. One very useful feature of the SCF multiple imputation programs is that they exploit this property of the linear regression models in a very simple way. The SCF imputation programs also deal with non-monotone patterns of the item non-response across households, since these programs allow us to select one set of covariates specific to each household, depending on the missing information.

Binary and categorical variables We estimate linear probability models for imputing binary variables and use hot deck procedures to impute categorical variables. Once again, the reason why we do not estimate discrete choice models by maximum likelihood or non-parametric models for imputing both binary and multinomial variables is the large number of different patterns of item missingness across observations in large surveys like the EFF and the SCF.

Concerning the imputation of multinomial variables by hot deck procedures, the SCF macro programs only allow us to use two covariates (either two discrete variables or one discrete and one continuous variable). However, we can construct interactions between two or more variables and use these interactions as the two covariates of the imputation model. The hot deck method imputes the most likely value of the variable to be imputed in the cell formed by the household observations having identical covariate values. Moreover, when the cell size resulting from the tabulation of the two covariates is very small or when there are no household observations having identical covariate values (mainly when we use one continuous covariate, such as total household income, or when the covariates consist of interactions between variables), the SCF hot deck procedure makes the cell size larger by merging adjacent cells having the nearest values of the covariates and imputes the most likely value of the variable of interest in the enlarged-size cell.

Bounds Another very useful feature of the SCF imputation programs is the possibility of restricting the imputed values of missing data to one upper and one lower bound specific to each observation. The upper and lower bounds are constructed using the information provided either by the EFF survey or previously imputed, whereby the way of constructing these constraints depends greatly on the information available for each household. The use of these bounds allows us to maintain consistency between the observed data and the imputed values of missing information in the EFF survey.

Some examples of these bounds are the following: when we impute the household average monthly food expenditures (obtained using questions 9.2 and 9.2b), we know logically that spending on food cannot be negative, so the lower bound we can impute is at least zero or one euro; furthermore, the imputed value of the average monthly food expenditures also has to satisfy an upper bound given by question 9.1, the average monthly spending on non-durable goods, which also includes food expenditures. When we impute food consumption, the lower bound is always satisfied trivially; however, the imputation model may need many trials to draw a sufficiently small random number that satisfies the upper bound, if the average monthly household non-durable expenditures (question 9.1) declared by the household is relatively small. When the imputation model is not able to impute stochastically one value inside the range defined by these two bounds, the imputed value is set equal to the nearest bound. Normally, the imputation model needs several trials to draw one either sufficiently large or small random number making the imputed value satisfy these upper and lower bounds.

4 Logical trees and shadow values of the EFF data

This section describes both the use and the construction of the shadow values of the EFF data. All survey variables are appropriately flagged by shadow values that indicate the nature of the data. Hence, if potential users of the EFF data wish to impute their data using non-parametric methods or more complex response and imputation models, they can do so using the shadow values.

4.1 Why shadow values are extremely useful in the imputation process

In order to know the origin of the EFF data for one particular observation and one variable, we have created as many flags as variables in the EFF questionnaire. These flags give information about whether the values provided have been answered by the households (i.e. they are actually observed values) or whether they have been imputed; these flags also show why these values are missing.

Moreover, the flags indicate whether the existing missing values in variables after imputing are really *true missing values* (i.e. given the household responses to previous questions of the interviewers, the households did not have to respond this question in particular) or whether they have been imputed as "true missing" during the imputation process. For instance, in the variable of the current value of the main residence (question 2.5 of the EFF questionnaire), the tenants' responses are logically missing, since the respondents of this question are only the homeowners. Thus, the missing values of all households renting are called *true missing values*. Moreover, if one household does not know or answer which is the housing tenure regime of the main residence (question 2.1 of the EFF questionnaire) and if we impute that this household is a renter, we also have to impute the response to question 2.5 as "true missing". The different values that these flags take for indicating the origin of the data and the reason for item missingness are called *shadow values*.

Consequently, before starting to impute, we first have to create these flags for all variables of the original data set. This task is done in two stages: in the first stage, we convert the codes of "don't know" and "no answer" responses (DK and NA responses) into missing values and assign their corresponding shadow values to all survey variables; both the DK and NA codes vary across variables.

In the second stage, we specify and program all the potential and logical relationships among the variables of the questionnaire, so that we can assign the shadow values correctly in all observations and in all variables, mainly in those variables having either true missing values or item missing values derived from the household non-response to a previous related question (i.e. missing values not derived from neither DK nor NA responses). The logical relationships that exist among the EFF variables are grouped in *logical trees* of variables; thus, we need to identify the total number of logical trees existing in the survey. In each tree, one variable is called *head-variable* and the remaining variables *branch-variables*. The household response (or non-response) to one head-variable affects both the values and the shadow values of the branch-variables, since some observed values of one head-variable may involve true missing values and may restrict the values in some branch-variables of its logical tree. The non-response to head-variables may imply that the branch-variables of their logical tree will be missing according to the EFF questionnaire.

Consequently, we have to program all the potential logical trees for assigning the correct shadow value to the branch-variables according to the values of the head-variable in all household observations and in all variables of the EFF survey. This process is quite time-consuming and depends greatly on how large and complex the survey questionnaire is. Compliance by the actual answers with these theoretical logical trees is facilitated by the use of Computer Assisted Personal Interviews (CAPI). The meaning and the total number of different shadow values are specific to the EFF survey, as they depend greatly on both the survey characteristics and its implementation. In the 2002 wave of the EFF, the flag variables can take the following values: 0, 1, and from 2047 to 2055.

The reason why the flags are constructed before the imputation stage is that the shadow values are continuously used to impute the missing data, since we only impute those variables having a shadow value 2050 or higher, as explained later. Moreover, the imputation stage relies greatly on all the logical trees established among the variables of the survey. The order in which the variables are imputed and the way in which the imputed value of one head-variable determines the value imputed subsequently to its branch-variables are based on all logical trees involving these variables.

4.2 Meaning of the EFF shadow values

Before defining the meaning of each shadow value, here are some examples of potential and logical relationships between survey variables. If households are renters (i.e. the response to question 2.1 is 1), questions in Section 2 of the questionnaire related to the characteristics of owner-occupied housing, such as the means of acquiring home ownership (question 2.2), questions from 2.3 to 2.8.a, and those related to the loans taken out for the purchase of the main residence (questions from 2.9 to 2.18 for the four most important outstanding loans), are not asked, and the renters' "responses" are "true missing". Consequently, these variables will have a shadow value of 0, which indicates that their values are "true missing". In this example, the logical tree involves the variables given by questions from 2.1 to 2.8.a and from 2.9 to 2.18; the head-variable is the indicator of housing tenure regime, question 2.1, and the group of branch-variables is formed by the rest.

If the household is a homeowner (the response to question 2.1 is 2), responds correctly to all questions asked and has two bank loans outstanding on the purchase of the main residence (the answer to question 2.8.a is 2), the shadow values of questions from 2.1 to 2.8.a and from 2.9 to 2.18 for the first two outstanding loans will be 1, indicating that the household has responded to these questions. However, the shadow values of questions from 2.9 to 2.18 referring to the third and fourth outstanding loans will be 0; they are "true missing" due to the fact that the household only has two outstanding loans; questions related to the third and fourth loans are not asked by the interviewer.

Moreover, another logical relationship among the same variables occurs when the household does not know which is the housing tenure regime. In this case, the shadow value of variable 2.1 will be 2050; the values of the variables related to the characteristics of the owner-occupied housing, such as questions from 2.2 to 2.8.a and from 2.9 to 2.18 among others, will be missing. As a result, these questions will have to be imputed as "true missing" or not depending on the previously imputed value of the housing tenure status, i.e. depending on whether we impute that the household is a renter or a home-owner, respectively. Finally, all these variables having missing values as a consequence of the item non-response to a previous question in the survey (in the example, to question 2.1) will have 2052 as the shadow value.

The most common shadow values: 0, 1, 2050, 2051, and 2052

0: The value of 0 is set to flags of all variables in which the sample observations have true missing values (as the flag of question 2.5 for those respondents being renters of their main residence).

1: The shadow value of 1 indicates that the value of the flagged variable is really observed, since the household has answered the question correctly.

2050 and 2051: The shadow values of 2050 and 2051 are assigned to the flags of those questions to which the household has responded "Don't know" and "No answer", respectively.

2052: As mentioned above, the shadow value of 2052 indicates that the flagged variable has a missing value due to the fact that one question previously asked has not been responded correctly. That is due to one of the following three cases: the household did not answer one previous related question, the household claimed not to know the answer or this previous question was never asked as a consequence of the non-response to another previous related question.

Rare shadow values The remaining values of the flag variables are not very common, and they are the result of some inconsistencies and errors detected in the data.

2053: The value of 2053 means that the household response to the flagged question is incorrect, overridden and has to be imputed.

2054: The value of 2054 indicates that the flagged variable is answered incorrectly by the household due to a mistake made by the interviewers. This happens to question 5.18, since the interviewers allowed the households to answer that the person that took out the insurance policy was the family or the employer instead of indicating which of the household members was the person.

2055: The shadow value of 2055 is set to some responses to question 5.23, the annual payment made on average for other forms of insurance. Due to an error in programming the CAPI, question 5.23 was not asked to households whose response to question 5.22 (which other forms of insurance the household has taken out) was to say "Other".

2047 and 2048: The shadow values of 2047 and 2048 are assigned to some of the questions that were not asked of the household members aged over 16, due to the missing information on their year of birth. The value of 2048 flags question 6.1, the labour market situation of the household members with missing age; this shadow value indicates that, in the 2002 wave of the EFF, the labour status of the two household members with missing age is fixed after checking that the preliminary imputations of the EFF data always imputed the same labour status to these household members.

The reason why we fix their labour market situation is to avoid constructing specifically imputation models for all the variables concerning characteristics of the employees, self-employed workers, and unemployed, retired or disabled household members. We save much time and effort and many computational requirements if we fix the current

employment situation of these household members, since we nearly always imputed them the same labour status (housewife/house-husband and retiree/early retire). In this way, we do not need to construct models for imputing the characteristics of the labour market situation for only two household members with missing age.

The shadow value of 2047 is only assigned to those variables for which we impute true missing values as a consequence of taking the current employment situation of the household members with missing age fixed.

2049: The shadow value of 2049 indicates that the value of the flagged variable has been edited using other information provided by the household.

The SAS programming for assigning shadow values to all survey variables was made easier due to the fact that the households were interviewed by CAPI and due to the fact that the original data was previously inspected; in this way, the number of inconsistencies and errors detected in the household answers at this stage was very small and not very serious. The logical relationships being the most difficult to specify and to program were those related to Section 6 of the EFF questionnaire, concerning the current labour market situation and labour earnings of all household members, especially the questions asked to the self-employed workers.

As said before, only variables with shadow values equal or higher than 2050 are multiply imputed. A small list of variables has not been imputed, due to the fact that either the fraction of missing information is higher than 60%, the number of respondents is very small to impute the variables suitably or due to the fact that the households have not generally understood the questions very well. The list of questions not imputed is the following: 4.8.1, 4.8.3, 4.40, 6.28.d, 6.28.f, 6.51.b, 6.57.b, 6.59.b, 6.60.b, 7.4.b, 7.8.b, and 7.10.[7] Missing values not imputed in this list of variables are marked with a value of -9999.

[7]Information about questions 1.6, 1.6.a, 1.6.b, 6.84, and from 8.8 to 8.10, are not provided in the public data set.

5 Imputation models in the EFF: covariates and specifications

5.1 Description of the imputation model covariates

As mentioned in Section 2, the goal of imputation is not to replace the missing data by the most accurate predicted values, but to preserve the characteristics of the distribution and the relationships between the different variables of the survey, so that the potential analyses based on statistics, such as means, percentiles and correlations among different variables, are unbiased. For this purpose, we need to include a high number of covariates in the imputation models in order not to bias the tests of different hypotheses about economic theories (for example, the permanent income hypothesis versus precautionary saving motive in consumption topics). We classify the covariates included in the EFF data imputation models into four groups, although some covariates may lie on several groups at the same time as their use may be motivated by the goal of different groups of covariates.

First group of covariates The first group of variables is formed by the determinants of non-response. The EFF data imputation models rely on the assumptions of "missing at random" and "ignorable missing data mechanism"; consequently, we should condition on a set of variables explaining or being related to the non-response to satisfy both assumptions. According to the MAR assumption, the distribution of the complete data only depends on the observed data, conditional on the determinants of the item non-response and other covariates.

Concerning the EFF data, variables that may explain the non-response and that should be included in the first group are the following: total household income (constructed by us using the information of the EFF); random wealth strata indicators; regional indicators; age and education of both the household head and the partner; and information provided by the interviewers, such as indicators of the type of both building

and neighbourhood, social status and house quality indicators, the respondent's degree of understanding and sense of responsibility in answering the questionnaire, indicators of where the interview took place (either inside the house or at the front door), and the number of other household members attending the interview.

Second group of covariates The second group is formed by covariates that are very good at predicting and explaining the variable of interest we want to impute. For example, among the variables included in this group to impute household income variables and amounts of wealth held in each kind of asset, we usually include non-durable consumption, since most regression estimates reveal that consumption is a good predictor.

When we impute the amount of wealth invested in each asset individually, we also include total household income, indicators of the different types of assets owned by the household (the yes/no questions about asset holdings have very small fractions of missing information), the current value of the owner-occupied house, the type and number of real estate properties owned, and the total value of these properties. Both the main residence and the other real estate properties are the most important assets in which Spanish households usually invest a great percentage of their wealth.[8]

At the beginning of each iteration of the imputation process, depending on both the sample size and the fraction of missing information on the values held in each asset, we always try to use the wealth values held in the most common assets as covariates of the imputation model. The most common assets are the main residence, other real estate properties, stocks, mutual funds and pension schemes.

Moreover, we first start imputing variables defined at the household level and the total values held in one particular type of asset (for instance, the total value of the portfolio invested in mutual funds). In this way, we first impute total household income,

[8]See the following articles of the Economic Bulletin publications: "Survey of Household Finances (EFF): Description, Methods, and Preliminary Results" (2005a) and Box 5 of "Quarterly Report on the Spanish Economy" (2005b) on pages 62-63.

instead of imputing either the income of each household member or the different income sources separately. This is done due to the fact that we have richer information at the level of the household or at the level of one given type of asset than the information we have separately for each household member or for the participation in each mutual fund. Furthermore, we also impute first not only those variables having a low percentage of missing values, but also those variables that are key covariates and very good predictors of a large number of the remaining variables to be imputed in the survey.

Third group of covariates The third group of covariates used in the imputation models are formed by those variables that are expected to affect or explain the variable to be imputed according to different economic theories, in order to preserve the existing relationships between these variables. The inclusion of this group of variables in the imputation model is very important, in order not to condition or bias the estimates made by the potential users of the data when they test the hypothesis of one particular economic model.

For example, irrespective of whether the current income may lie in the other groups of covariates, when we impute non-durable consumption, we need to include current income as a covariate, in order not to lead to misleading results and not to bias the estimates of the potential users in favour of economic theories based on the permanent income hypothesis.[9] Moreover, in the imputation of non-durable consumption we also need to include variables explaining household income uncertainty, so that we do not bias the empirical evidence against precautionary saving motive models [see Dynan (1993), Carroll (1994), and Albarran (2000), among others].

Fourth group of covariates The fourth group of covariates is formed by those variables that are determinants or very good predictors of the covariates included in the rest

[9] See Browning and Lusardi (1996) and Attanasio (1999) for recent surveys about household saving and consumption.

of the groups of variables. Its role is very important at least in the first iteration, since variables are imputed sequentially based on both the observed data and the values of the previously imputed variables. This is due to the fact that there is a very large number of different patterns of item missingness across observations in the EFF. For this reason, we need to include variables that explain the covariates included in the rest of the groups; if we have missing information on some key covariates, we need other covariates that explain or predict the missing covariates very well. In this way, our imputation model will not be very poor, as we can reshape the matrices, $(X'X)^{-1}$ and $(X'y)$, in equation (10) restricting the set of covariates of household i, X_i, to those not having missing information when imputing the missing value of the variable of interest, y_i.[10]

Therefore, the last group of covariates tries to predict and capture the explanatory power of other missing predictors of the imputation model for some household observations; we usually try to include a set of key variables as large as that allowed by the sample size available to impute the variable of interest. Some of these essential household characteristics are the following: both household composition and structure (number of children, children's age, household head's civil status, number of adults in the household, number of household member adults broken down by their labour market situation, among others) as well as personal characteristics of both the household head and the partner, such as age, education, labour history, current labour status, type of work done, economic activity and other characteristics of the main job.

As a result, many variables of the survey, such as income, age and education, take part as covariates in the imputation model due to the fact that they help achieve the different aims of more than one group of covariates at the same time.

[10] That is, the number of non-missing covariates we have for household i may be smaller than that of X, the total number of covariates involved by the imputation model of variable, y.

5.2 Specifications of the imputation models

Continuous variables Concerning the functional form of the imputation models, in order to take non-linearities into account, regressors may either be formed by interactions between variables or introduced in logarithms or as polynomials. To impute euro questions that may have a zero value, we first impute a binary variable indicating whether or not the variable of interest has a positive value, and then impute the logarithm of the positive values.

Concerning some questions about percentages, such as question 6.38.2.4 (the percentage of the business value owned by the household when one household member is a self-employed partner in a non-family partnership), we first impute a categorical variable indicating whether the percentage is lower than 50%, equal to 50%, between 50% and 100% or equal to 100%. This is done when a histogram of the percentage shows some probability mass points like 50% and 100%. Next, we impute the logarithm of the percentage as a continuous variable restricting its imputed value to lying in the range previously imputed by the categorical variable.

Continuous variables are usually imputed using models based on their logarithm. Exceptions are the questions about the amount of money that the household has to repay in outstanding loans (loans taken out for the purchase of either the main residence or the other real estate properties, and other debts), such as questions 2.12, 2.55 and 3.6 of the EFF questionnaire. For these variables, we set up an imputation model for the logarithm of the ratio of the amount of money not repaid to the total value paid back; next, we recover the imputed value of the amount of money not repaid in the outstanding loan using either observed or previously imputed data of the initial amount of the loan. These model specifications work much better than the models that impute directly the logarithm of the total amount pending repayment.

Sometimes, to impute continuous variables, we specify an imputation model for another variable highly related to the variable of interest. This is done when the latter model makes more economic sense and has a greater explanatory power. For example,

to impute question 6.23.a, in what year the household member became unemployed, we specify the imputation model for another variable, in particular for the logarithm of the number of years that the unemployment spell has lasted.

Multinomial variables As mentioned in subsection 3.2, the imputation of categorical variables is done by hot deck procedures. One drawback of using hot deck is that we cannot take into account a high number of covariates in the imputation models. The SAS macro programs written for imputing the SCF data do not allow us to include more than two covariates. However, these macro programs allow us to include one continuous variable, such as income or age, or two covariates being the result of interactions between other variables. In this way, the problem of not controlling for a sufficient number of covariates in the hot deck imputation is partially solved.

For imputing the EFF data, we usually include one discrete variable formed by the interaction of the random wealth strata indicators with the total household income quartile. Depending on the sample size over which the variable of interest is imputed, we also interact this interaction with other variables, such as the family head's age bands, education indicators, or some other characteristic specific to the variable of interest. Both covariates used for imputing categorical variables by hot deck procedure should not have missing values, otherwise they need to be imputed previously.

Questions asked separately to each household member over 16, each particular asset within an asset type, each job, etc. Next, I will describe how we impute missing information about questions posed individually to each household member and related to characteristics of the self-employed workers' jobs, employees' jobs, and pensions; or questions posed to households concerning their holdings in different mutual funds, pensions schemes, real estate properties, loans, etc. For example, question 6.14 (the regular monthly gross earnings) is collected for up to nine household members in their three most important jobs working as employees. The way of imputing variables like 6.14 is to construct a pooling of subsamples defined for each household member and

for each job: first, we generate the covariates of the imputation model separately for each household member and job; next, we pool all these subsamples and estimate the parameters of the imputation model over the pooled sample; and finally, the imputed values of the variable of interest (like 6.14) are updated in the original data set.

The imputation of these variables by pooling subsamples is laborious and computer-intensive; moreover, running the imputation programs becomes very slow. The main difficulty in programming the imputation models by pooling samples appears in the first iteration of the imputation process. In the remaining iterations, the parameters of the imputation models are estimated over the sample formed by observed data and missing data imputed in the previous iteration. However, in the first iteration, the estimates of the imputation models are based on both the observed data and the missing data previously and sequentially imputed within the iteration. Therefore, once we finish imputing one variable in the first iteration, we have to update the missing values of all predictors and covariates derived from the imputed variable, since they will be used as covariates in the imputation models of the remaining survey variables with missing information. We have to update the implied values of all these missing covariates and predictors in the original data set where we save the multiply imputed values and in all auxiliary data sets coming from pooling samples. The updating of values in pooled samples is a very time-consuming process in SAS.

Constructed total household income variables In the EFF data, we also include two constructed total household variables: one corresponds to the earnings obtained in 2001 and the other to the income received in the month in which the interview took place during 2002 or 2003. These two variables are calculated as the sum of the property income from the households' asset holdings as well as the labour and non-labour earnings received by all household members. If there is item non-response in at least one source, the constructed total household income variable is imputed. Moreover, we can also construct alternative income variables as the sum of all these income sources once they have been imputed separately; however, the total household income

variables imputed directly will obviously differ from the total household income variables constructed as the sum of all sources of earnings imputed separately.

Finally, we want to point out some results about the imputation of the income variables broken down by household members and by different income sources. The imputation models impute higher total household income values when we impute total income variables than those obtained when total income is generated as the sum of the different income sources imputed individually and separately for each household member (in each one of their jobs, pensions, unemployment benefits, etc.) and for each type of asset (rent income, interest income, dividends, earnings from fixed-income securities, etc.). We think that this is due to the fact that we have richer information for imputing the constructed variables of total household income. In particular, we can also obtain a lower bound of the imputed total household income constructed. However, we do not have any ranges provided by the respondents when they do not know the exact amount of earnings from one particular source. In the imputation of the two constructed variables of the total household income, we observe that the lower bounds defined by the declared income make the imputation programs draw several random numbers until the imputed income satisfies the lower bound. Unfortunately, we do not have any information for defining the lower bounds of each income source individually.

Multiple responses to one question posed Moreover, the EFF survey also contains questions allowing the households to make multiple responses, for example the question about the different types of commissions to which the outstanding loans taken out for the purchase of each real estate property are subject (question 2.58). The way of imputing multiple responses is also by pooling subsamples defined for each one of the household's multiple responses. For example, to impute question 2.58, we pool the subsamples defined for each of the four potential multiple responses allowed for each of the three most important loans taken out for the purchase of each of the three most important real estate properties owned by the household. These questions allowing multiple responses are usually imputed using hot deck procedures.

6 Evaluating the imputation of the EFF data

As explained in Bover (2004), we implement two procedures specific to the imputation of the EFF data for evaluating the imputed values. This is done due to the fact that we are also concerned about two issues: the first is the evaluation of the imputations of continuous variables, and the second is the convergence of the sample distributions, when we impute stochastically (that is, when we add a disturbance to the predicted value by the regression).

6.1 Evaluating the imputation of continuous variables

Concerning the evaluation of the imputed values of continuous variables, we cannot use goodness of fit statistics like the R^2 for evaluating the imputation; a good within-sample fit does not necessarily ensure a good out-of-sample fit for the non-respondents. Moreover, as we have very different patterns of item missingness across observations, depending on the number of missing cases among the covariates, the values are imputed in the first iteration as if they come from individual regressions done for each household observation; consequently, the R^2 is not a suitable goodness-of-fit statistic. As a result, in order to evaluate whether the imputed values are reasonable or can be considered atypical, we compare each imputed value with both the maximum and the minimum values of the imputed variable over a sample of respondents that are neighbours. For each non-respondent to the variable of interest, we want to evaluate its imputed value, $\widehat{y}_{mis,i}$. Depending on the number of non-missing covariates of the non-respondent i, X_i, the neighbours of i are those observations, j, for which the norm of the differences in their covariate values, normalised by the variance-covariance matrix of the non-missing covariate vector, does not exceed a limit, ε, as follows:

$$\|X_i - X_j\| = \sqrt{(x_{i1} - x_{j1} \dots x_{ik} - x_{jk})\, \widehat{\Omega}_i^{-1}\, (x_{i1} - x_{j1} \dots x_{ik} - x_{jk})'} \leq \varepsilon \qquad (12)$$

The index i denotes the sample observation of the non-respondent to question y, whose imputed value we want to evaluate. The index j denotes the sample observation of

all households being neighbours to household i according to the rule defined in equation (12). The $k \times k$ matrix, $\widehat{\Omega}_i$, denotes the sample variance-covariance matrix of the non-missing covariate vector of non-respondent i; this matrix is obtained using the sample over which the imputation model is constructed, and k denotes the number of non-missing covariates of non-respondent i. The size of the neighbourhood is set by the bound, ε; and the value of this bound differs depending on whether the non-respondent i has missing covariates, as follows:

$$\varepsilon = \begin{cases} \sqrt{\text{no. discrete variables in } X} + \sqrt{\text{no. continuous variables in } X} & \text{if all covariates are observed for non-respondent } i \\ \sqrt{\text{no. non-missing covariates}} & \text{if some covariates of } X \text{ are missing for non-respondent } i \end{cases} \tag{13}$$

The matrix, X, contains the number of sample observations for which we have available both the values of the variable of interest, y, and the values of the K covariates involved by the imputation model. As the covariates of the imputation model may have missing values for some non-respondents of the variable, y, the number of non-missing covariates for non-respondent i, X_i, is k being equal or less than K.

Therefore, in order to evaluate the imputed value of non-respondent i, we consider that the imputed value, $\widehat{y}_{mis,i}$, is reasonable if it lies inside the range of the values that the variable of interest, y, takes among the neighbours of non-respondent i, as follows:

$$\widehat{y}_{mis,i} \in \left[\min\left(y_{obs,j}\right) - \widehat{\sigma}, \max\left(y_{obs,j}\right) + \widehat{\sigma}\right], \quad j \text{ such that } \|X_i - X_j\| \leq \varepsilon \tag{14}$$

The upper and lower bounds of the range are defined by the maximum and minimum values reached among the sample of neighbours; however, both bounds are disturbed by subtracting and summing respectively the standard error of the regression residuals, $\widehat{\sigma}$, in equation (10).

Accordingly, the imputed values, $\widehat{y}_{mis,i}$, that lie outside the range are considered to be atypical, i.e. if $\widehat{y}_{mis,i} \notin [\min\left(y_{obs,j}\right) - \widehat{\sigma}, \max\left(y_{obs,j}\right) + \widehat{\sigma}]$, and they are reset to the

nearest bound of the range, $\min(y_{obs,j})$ if $\widehat{y}_{mis,i} < \min(y_{obs,j}) - \widehat{\sigma}$ and $\max(y_{obs,j})$ if $\widehat{y}_{mis,i} > \max(y_{obs,j}) + \widehat{\sigma}$. However, we only reset the imputed values considered atypical if the size of the neighbour sample is high enough and if the nearest bound satisfies the constraints of the imputed values given by the information available in the EFF; the required number of neighbours is 15 or more.

If the size of the neighbourhood is smaller than 15, we reimpute the value again using the sample of respondents and non-respondents having "reasonably" imputed values. Then, we evaluate the new imputed value repeating the same procedure of searching for the nearest neighbours. This procedure is repeated up to five times. In the fifth round, the size of the neighbourhood is made larger by increasing the value of the norm bound, ε, in equation (13) by 10. In the last round, we do not reset the imputed values for which we do not find any neighbours or for which the number of neighbours is less than 15.

This way of evaluating whether the imputed values are reasonable or atypical is done just after having finished imputing one continuous variable in the first iteration to ensure reasonable starting values. The reason why we use this procedure based on the Euclidean norm to search for the nearest neighbours and for evaluating the imputed values is that the sample sizes may be very small to condition on the values of all model covariates, given the high number of covariates we need to control for and given the huge number of different patterns of item missingness across observations. Using this procedure, we try to limit the adverse effect of drawing either very large or very small random numbers on the values imputed stochastically.

In practice, the number of values imputed and reset is very tiny; most of the imputed values that lie outside the range defined by their nearest neighbours are those which have to satisfy both an upper and a lower bound imposed either by consistency and logical reasons or by the information collected by the survey. A large number of cases of imputed values being reset according to the nearest neighbour procedure happen to those observations in which the imputation models take several trials to impute one value

satisfying both the upper and lower bounds, since the imputation models must draw unusual values of the random numbers in these cases. When we replace the imputed values by the maximum or minimum values from the neighbour sample with a sufficiently large size, we always take the upper and lower bounds into account. Moreover, we notice that many imputed values cannot be replaced, in order to satisfy these logical upper and lower bounds (several of them have even been imputed equal to one bound). We also observe that the imputed values are replaced by other very similar values to satisfy the bounds.

6.2 Evaluating the convergence of the imputed data across iterations

The convergence of the imputed data across iterations of the imputation process is analysed by studying how both the median, M_y, and the interquartile range, IQR_y, of each imputed variable, y, change across two consecutive iterations, t and $t-1$, as follows:

$$\sqrt{\left(\begin{bmatrix} M_y^{(t)} \\ IQR_y^{(t)} \end{bmatrix} - \begin{bmatrix} M_y^{(t-1)} \\ IQR_y^{(t-1)} \end{bmatrix}\right)' \left(\begin{bmatrix} M_y^{(t)} \\ IQR_y^{(t)} \end{bmatrix} - \begin{bmatrix} M_y^{(t-1)} \\ IQR_y^{(t-1)} \end{bmatrix}\right)} \tag{15}$$

That is to say, we observe a convergence of the sample distributions across iterations when the Euclidean norm of the difference of the vector formed by the median and the interquartile range in two consecutive iterations decreases with the number of iteration, t. We observe how the norm value of each imputed variable decreases with the number of iteration; this value becomes stable, very small and near zero from the second to the last iterations, since the medians and interquartile ranges are very similar across iterations. We define the convergence criterion based on dispersion and position measures instead of using point-wise criteria, due to the fact that one part of the multiply imputed values arises from randomisation and it cannot converge like the point estimates of the model parameters.

7 Conclusions

This paper describes the imputation of the first wave of the Spanish Survey of Household Finances (EFF) implemented in 2002. One typical feature of wealth surveys like the EFF is the presence of high rates of item non-response due to the lack of knowledge or due to the unwillingness of households to reveal certain information about their income and wealth. As the item non-response rates are not random but depend on household characteristics, all studies carried out by the potential users of the EFF data may be misleading if missing information is not imputed. However, the potential users of the public data sets may neither know the imputation techniques nor devote enough time, effort and computational resources to obtain acceptable imputations. Accordingly, the EFF database constructor provides the imputed data as well as the flags of all survey variables explaining the origin of the data (whether observed or missing). For imputing the EFF data, we have been very fortunate to benefit from the programs written by Arthur Kennickell for the SCF multiple imputation. This paper also tries to make transparent how the imputation of the EFF data has been conducted for users. For this purpose, the paper explains the specific features of EFF data imputation, such as the way of defining the shadows values that flag the EFF survey appropriately, the covariates included in the imputation models, the specifications of the imputation models, the order in which variables are imputed and the means of evaluating both the imputed data and the convergence of the imputation process. Finally, we expect that this description of the EFF imputation procedure may facilitate improvements in the imputation of future waves of the survey.

References

Albarran, P. (2000). *Income Uncertainty and Precautionary Saving: Evidence from Household Rotating Panel Data,* CEMFI, Working Paper No. 0008.

Attanasio, O. P. (1999). "Consumption", Edited by: J. B. Taylor and M. Woodford, *Handbook of Macroeconomics,* volume 1B, North-Holland, Amsterdam, pp. 741-812.

Banco de España (2005a). "Survey of Household Finances (EFF): Description, Methods, and Preliminary Results", *Economic Bulletin,* January.

– (2005b). "Quarterly Report on the Spanish Economy", *Economic Bulletin,* April.

Bover, O. (2004). *The Spanish Survey of Household Finances (EFF): Description and Methods of the 2002 Wave,* Occasional Paper No. 0409, Banco de España.

Browning, M. and A. Lusardi (1996). "Household Saving: Micro Theories and Micro Facts", *Journal of Economic Literature,* vol. 34 (4), pp. 1797-1855.

Cameron, A. C., and P. K. Trivedi (2005). *Microeconometrics: Methods and Applications,* Cambridge University Press, Cambridge.

Carroll, C. D. (1994). "How Does Future Income Affect Current Consumption?", *The Quarterly Journal of Economics,* vol. 109 (1), pp. 111-147.

Dynan, K. E. (1993). "How Prudent are Consumers?", *The Journal of Political Economy,* vol. 101 (6), pp. 1104-1113.

Kennickell, A. B. (1991). *Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation.*

– (1998). *Multiple Imputation in the Survey of Consumer Finances.*

Korinek, A., J. A. Mistiaen and M. Ravallion (2005). *Survey Nonresponse and the Distribution of Income,* Policy Research Working Paper No. 3543, World Bank.

Little, R. J. A., and D. B. Rubin (1987). *Statistical Analysis with Missing Data,* John Wiley & Sons, New York.

Rubin, D. B. (1976). "Inference and Missing Data", *Biometrika*, 63 (3), pp. 581-592.

– (1987). *Multiple Imputation for Nonresponse in Surveys,* John Wiley & Sons, New York.

– (1996). "Multiple Imputation After 18+ Years", *Journal of the American Statistical Association,* vol. 91 (434), pp. 473-489.

Vazquez Alvarez, R., B. Melenberg and A. van Soest (1999). *Nonparametric Bounds on the Income Distribution in the Presence of Item Nonresponse,* Discussion Paper No. 9933, Tilburg University, Center for Economic Research (CentER).

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data,* Chapman & Hall, London.

Tanner, M. A., and W. H. Wong (1987). "The Calculation of Posterior Distributions by Data Augmentation", *Journal of the American Statistical Association,* vol. 82 (398), pp. 528-540.

BANCO DE ESPAÑA PUBLICATIONS

OCCASIONAL PAPERS

0304 ALBERTO CABRERO, CARLOS CHULIÁ AND ANTONIO MILLARUELO: An assessment of macroeconomic divergences in the euro area. (The Spanish original of this publication has the same number.)
0305 ALICIA GARCÍA HERRERO AND CÉSAR MARTÍN MACHUCA: La política monetaria en Japón: lecciones a extraer en la comparación con la de los EEUU.
0306 ESTHER MORAL AND SAMUEL HURTADO: Evolución de la calidad del factor trabajo en España.
0307 JOSÉ LUIS MALO DE MOLINA: Una visión macroeconómica de los veinticinco años de vigencia de la Constitución Española.
0308 ALICIA GARCÍA HERRERO AND DANIEL NAVIA SIMÓN: Determinants and impact of financial sector FDI to emerging economies: a home country's perspective.
0309 JOSÉ MANUEL GONZÁLEZ-MÍNGUEZ, PABLO HERNÁNDEZ DE COS AND ANA DEL RÍO: An analysis of the impact of GDP revisions on cyclically adjusted budget balances (CABS).
0401 J. RAMÓN MARTÍNEZ-RESANO: Central Bank financial independence.
0402 JOSÉ LUIS MALO DE MOLINA AND FERNANDO RESTOY: Recent trends in corporate and household balance sheets in Spain: macroeconomic implications. (The Spanish original of this publication has the same number.)
0403 ESTHER GORDO, ESTHER MORAL AND MIGUEL PÉREZ: Algunas implicaciones de la ampliación de la UE para la economía española.
0404 LUIS JULIÁN ÁLVAREZ GONZÁLEZ, PILAR CUADRADO SALINAS, JAVIER JAREÑO MORAGO AND ISABEL SÁNCHEZ GARCÍA: El impacto de la puesta en circulación del euro sobre los precios de consumo.
0405 ÁNGEL ESTRADA, PABLO HERNÁNDEZ DE COS and JAVIER JAREÑO: Una estimación del crecimiento potencial de la economía española.
0406 ALICIA GARCÍA-HERRERO AND DANIEL SANTABÁRBARA: Where is the Chinese banking system going with the ongoing reform?
0407 MIGUEL DE LAS CASAS, SANTIAGO FERNÁNDEZ DE LIS, EMILIANO GONZÁLEZ-MOTA AND CLARA MIRA-SALAMA: A review of progress in the reform of the International Financial Architecture since the Asian crisis.
0408 GIANLUCA CAPORELLO AND AGUSTÍN MARAVALL: Program TSW. Revised manual. Version May 2004.
0409 OLYMPIA BOVER: The Spanish Survey of Household Finances (EFF): description and methods of the 2002 wave. (There is a Spanish version of this edition with the same number.)
0410 MANUEL ARELLANO, SAMUEL BENTOLILA AND OLYMPIA BOVER: Paro y prestaciones: nuevos resultados para España.
0501 JOSÉ RAMÓN MARTÍNEZ-RESANO: Size and heterogeneity matter. A microstructure-based analysis of regulation of secondary markets for government bonds.
0502 ALICIA GARCÍA-HERRERO, SERGIO GAVILÁ AND DANIEL SANTABÁRBARA: China's banking reform: an assessment of its evolution and possible impact.
0503 ANA BUISÁN, DAVID LEARMONTH AND MARÍA SEBASTIÁ BARRIEL: An industry approach to understanding export performance: stylised facts and empirical estimation.
0504 ANA BUISÁN AND FERNANDO RESTOY: Cross-country macroeconometric heterogeneity in EMU.
0505 JOSÉ LUIS MALO DE MOLINA: Una larga fase de expansión de la economía española.
0506 VICTOR GARCÍA-VAQUERO AND JORGE MARTÍNEZ: Fiscalidad de la vivienda en España.
0507 JAIME CARUANA: Monetary policy, financial stability and asset prices.
0601 JUAN F. JIMENO, JUAN A. ROJAS AND SERGIO PUENTE: Modelling the impact of aging on Social Security expenditures.
0602 PABLO MARTÍN-ACEÑA: La Banque de France, la BRI et la création du service des Études de la Banque d´Espagne au début des années 1930.
0603 CRISTINA BARCELÓ: Imputation of the 2002 wave of the Spanish Survey of Household Finances (EFF).

BANCO DE ESPAÑA

Unidad de Publicaciones
Alcalá, 522; 28027 Madrid
Telephone +34 91 338 6363. Fax +34 91 338 6488
e-mail: Publicaciones@bde.es
www.bde.es