THE IMPACT OF ALTERNATIVE IMPUTATION METHODS ON THE MEASUREMENT OF INCOME AND WEALTH: EVIDENCE FROM THE SPANISH SURVEY OF HOUSEHOLD FINANCES

2008

Cristina Barceló

Documentos de Trabajo
N.º 0829

BANCO DE ESPAÑA
Eurosistema

THE IMPACT OF ALTERNATIVE IMPUTATION METHODS ON THE MEASUREMENT OF INCOME AND WEALTH: EVIDENCE FROM THE SPANISH SURVEY OF HOUSEHOLD FINANCES

THE IMPACT OF ALTERNATIVE IMPUTATION METHODS ON THE MEASUREMENT OF INCOME AND WEALTH: EVIDENCE FROM THE SPANISH SURVEY OF HOUSEHOLD FINANCES

Cristina Barceló (*)

BANCO DE ESPAÑA

(*) I wish to thank Olympia Bover, Elena Martínez-Sanchís and Ernesto Villanueva and an anonymous referee for their suggestions and comments. This paper has also benefited from the contribution of seminar participants at CEMFI (1 Day PEW, 2007), XXXII SAE (2007) and Banco de España (2008). All errors are my responsibility. The first part of the paper circulates under the title "Imputation of the 2002 wave of the Spanish Survey of Household Finances (EFF)", which describes specifically the imputation of the 2002 wave of the EFF. Address for correspondence: Banco de España, Servicio de Estudios, Alcalá 48, 28014 Madrid, Spain (tel.: +34 91 338 5887, fax: +34 91 338 5678, E-mail: barcelo@bde.es).

Documentos de Trabajo. N.º 0829
2008

ISSN: 0213-2710 (print)
ISSN: 1579-8666 (on line)
Depósito legal:
Unidad de Publicaciones, Banco de España

Abstract

The goal of this paper is to emphasise the importance of the way of handling missing data and its impact on the outcome of empirical studies. Using the 2002 wave of the Spanish Survey of Household Finances (EFF), I study the performance of alternative methods: listwise deletion, non-stochastic, multiple and single imputation based on linear-regression models, and hot-deck procedures.

Using descriptive statistics of the marginal and conditional distributions of income and wealth and estimating mean and quantile regressions, listwise deletion brings imprecise and biased estimates, non-stochastic imputation underestimates variance and dispersion and hot deck fails to capture the potential relationships among survey variables.

Keywords: Household wealth surveys, imputation methods.

JEL codes: D10, C81.

1 Introduction

The goal of this paper is to evaluate empirically the most usual ways of handling missing data on income and wealth variables. This is done using the first wave of the *Spanish Survey of Household Finances* (in Spanish *Encuesta Financiera de las Familias*, EFF hereafter). The EFF is a wealth survey that the Banco de España decided to launch in 2001 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] The EFF survey, whose first wave corresponds to 2002, collects information about households' holdings in real and financial assets, debts, different sources of income and 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 wealth surveys. Non-response takes place in a survey in two ways. First, *survey* or *unit non-response* occurs when households do not want to participate in the survey or cannot be located by the interviewers. A typical characteristic of wealth surveys is the existence of a high rate of survey non-response that is not random and depends on income and wealth; the higher household wealth, the higher non-participation in the survey is. This problem is more severe in wealth surveys that oversamples wealthy individuals like the SCF and the EFF. Oversampling is a desirable feature of wealth surveys to allow the feasibility of studies concerning the household portfolio decisions. This is so because the distribution of household wealth is heavily skewed and some types of assets, mainly financial assets, are only held by a low percentage of the population. One way of taking into account that unit non-response is not random is to use weights adjusted by the non-response in order not to bias the potential analysis of the data.

The second type of non-response is called *item non-response* and happens when households do not answer all questions asked by the interviewer, because of lack of understanding of the question, lack of knowledge of the answer, and reluctance and unwillingness to

[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/effe.htm).

disclose some information. This entails the existence of missing data in some parts of the questionnaire completed by households. Item non-response is the kind of non-response that I will address in this paper. Like survey non-response, 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 the EFF, item non-response affects mostly variables on income, wealth, debt and values invested in each type of asset in a non-random way. The problem again becomes more serious in surveys with oversampling of the wealthy since usually the richer the households, the higher the item non-response rates are. Accordingly, the results of all potential analyses based on such surveys ignoring the presence of missing data and not taking into account that the item non-response is not random can be misleading. Moreover, irrespective of whether the item missingness is random or not, Rubin (1996) also suggests another reason why the data base constructors should provide imputations, deleting households with missing data would render some multivariate studies infeasible due to the resulting small sample sizes, as we can see later in the empirical results.

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. There are two characteristics of the imputations provided by the SCF and the EFF. First, several values are imputed for each missing observation, based on the method proposed by Rubin (1976) and explained in more detail in Rubin (1987), Little and Rubin (1987) and Schafer (1997).[3] By providing data imputed multiply, the analysis may take into account the uncertainty about the imputed data. Second, the imputations will preserve the characteristics of the distributions and the relationships among the variables of the survey, if the imputation models allow for a great number of covariates that try to preserve the potential relationships among the imputed variables and the rest of variables of the survey.

Nowadays, many researchers, (see, for instance, Korinek *et al.,* 2005 and 2007, Vazquez Alvarez *et al.,* 1999, and De Luca and Peracchi, 2007), are concerned about how non-

[3]See Kennickell (1991, 1998) for a detailed description of the imputation methods of the *Survey of Consumer Finances* and Barceló (2006) for the EFF.

response may affect their estimates, and they use different parametric or non-parametric techniques to deal with this issue and control for non-response. Korinek *et al.* (2007) exploit the geographic structure of survey nonresponse rates to identify a parametric compliance function that serves to re-weight income data from the US *Current Population Survey* (CPS). Using this method, Korinek *et al.* (2005) encounter that correcting for survey nonresponse increases mean income and inequality. Vazquez Alvarez *et al.* (1999) focus on item nonresponse and follow the Manski's approach to estimate the distribution function and quantiles of personal income; unlike the parametric approach of selection models, without additional assumptions they identify the parameters of interest up to a bounding interval taking into account item nonresponse.

De Luca and Peracchi (2007) investigates whether the missing data mechanisms underlying item and unit non-response on food and expenditure consumption may be considered *missing at random* (MAR) or not using selectivity models and data from the *Survey on Health, Aging and Retirement in Europe* (SHARE). As explained in Section 4, the MAR assumption implies that the distribution of the complete data (observed and missing data) only depends on observed data. This assumption may not be reliable if information is not available about key covariates related to the main determinants of the non-response on income and wealth variables, such as location variables and wealth strata and social status indicators. These variables are key covariates in all imputation models of the EFF data.

David *et al.* (1986) compare alternative imputation methods for the CPS wages and salaries. In particular, they focus on various ways of imputing by hot-deck procedures, regression imputations and different approaches for imputing using single randomised regressions (adding to the predicted value a random term or an empirical residual selected randomly). David *et al.* (1986) can obtain an exact match of the March 1981 CPS data to the *Internal Revenue Service* (IRS) tax records using Social Security numbers of tax filers. They use the IRS data to investigate how the different methods of imputing data underestimate the aggregate and to compare various measures of relative and absolute errors over the aggregate to assess the bias of each method. They conclude that the

imputations based on regression models have slightly smaller mean absolute errors than the hot deck procedures. This paper compares imputation methods not only studying the aggregate, but also focussing on the characteristics of the distributions and estimating some conditional models to analyse the relationships among wealth and income variables.

In this paper, I compare the performance of the multiple imputation method applied in the first wave of the *Spanish Survey of Household Finances* (EFF) against other widely used methods of handling missing data, such as the deletion of the incomplete cases (*listwise deletion*), and other imputation methods, such as non-stochastic linear-regression imputations and stochastic imputation methods based on randomised linear-regression models and hot-deck procedures, and single versus multiple imputations.

For this purpose, I have imputed household total income and wealth held in every asset in two different ways. First, I use the EFF imputation models, mostly randomisation from regression predictions [see Barceló (2006) for details] to impute non-stochastically, without allowing for the randomisation term. Second, I impute the same variables by hot deck using the most economically and statistically significant covariates of the regression prediction models. I highlight the advantages of the multiple imputation models based on a wide range of covariates, looking at the descriptive statistics and the estimates from mean and quantile regressions of some relevant wealth and income variables.

The main results of the paper can be summarised in four points. First, the deletion of incomplete cases usually biases the results and increases the inefficiency of the statistics by reducing the sample size. Second, the non-stochastic imputation method makes the distribution of the imputed variables more peaked around the mean and hence reduces dispersion. Third, hot deck imputation methods cannot preserve all the relationships among the survey variables due to the need of conditioning only on a very small number of covariates. Finally, single imputation may yield misleading information about the significance of the statistics, since it treats the imputed value as it was the actual one and underestimates variances. The desirability of multiple imputation methods versus the deletion of incomplete cases and other *ad hoc* imputation methods have also been addressed in topics on finance by Kofman and Sharpe (2003) and in other disciplines like

health sciences by Longford *et al.* (2000) in their study of alcohol consumption.

The structure of this paper is as follows: Section 2 explains in detail why imputation is useful and which imputation methods would be more appropriate, specially in wealth surveys. Section 3 explains the motivation behind multiple imputation and describes the imputation procedure implemented in the EFF. Section 4 explains more practical issues of the EFF imputation that could be useful when performing imputations in other data sets. Section 5 describes the empirical analysis carried out to evaluate the performance of alternative methods of handling missing data and the main results obtained. Finally, Section 6 summarises the main conclusions of this paper.

2 The choice of the imputation method

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 of interest ("listwise deletion") and to replace 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, the imputation methods and the ways of dealing with missing data mentioned above ("fill-in with means", "listwise deletion" 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 underestimate the variance. Finally, results based on "listwise deletion" may 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.

Only imputation methods based on stochastic imputation can help preserve the distribution of the complete data, since missing information is imputed randomly by hot-deck procedures or by adding a random number to the values 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 and variances.

However, as Rubin (1987, 1996) states, one single stochastic imputation does not take into account the uncertainty about the imputation due to the fact that it treats the imputed value as if it was the actual one; we need to draw several imputed values to

take into account the uncertainty about the imputed values and not to underestimate the standard errors of all statistics used. With only one single stochastic imputed value of the missing data, in the empirical analysis we would have to use complete-data econometric tools as if they were the true data, forgetting that they are not actually observed. The EFF, like the SCF, 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 imputed values under one model for non-response.

Little and Rubin (1987) and Rubin (1996) point out that multiple imputation can reflect two kinds of uncertainty: first, uncertainty about the imputed values under a given model for non-response by drawing stochastically several imputed values (as done by the SCF and the EFF) and, second, uncertainty about the correct model for non-response by drawing stochastically several imputed values not only under one model for non-response, but also across different models for non-response. The combined inferences [explained later and shown in equation (6)] across these models can be contrasted to analyse their sensitivity to the models for non-response. This second kind of uncertainty about the correct model for non-response is not addressed by the multiple imputation methods applied in the SCF, the EFF and in this paper.

However, the multiple imputation method applied by the SCF and the EFF is robust to misspecifications of the imputation models. If the model for non-response is misspecified or poor, the multiple random terms added to the predicted value will differ greatly among themselves and the influence of the randomised part on the total imputed value will be stronger. Moreover, the random terms represent the variation unexplained by the imputation model and come from a distribution whose variance is that of the residuals of the regression model.

Finally, Rubin (1996) gives the two most important reasons why the database constructors should provide imputations of the missing data, instead of letting potential users impute their own data. First, 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 computational resources to obtain acceptable imputations. Second, to preserve confiden-

tiality, users of the data will not receive information about some relevant variables that are 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 to users; these variables are not only very good predictors of many variables, but they are also important factors of item non-response. Thus, users 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*.

However, if users of the EFF data wish to carry out more complex imputation methods or to deal with missing data using maximum likelihood models or other approaches, they may do it. Indeed, all survey values are flagged in such a way that the users know whether they are originally observed or missing and the reason for item missingness (respondent does not know or is not willing to give an answer).

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:

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

This assumption asserts that the distribution of the non-response conditional on the complete data (the observed and the missing data) is independent of the missing data. This assumption is satisfied when we can control for the determinants of the non-response in the imputation models using the observed data, Y_{obs}. Thus, the lack of some key covariates of the non-response will make the imputations not reliable.

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 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 not related). 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 observa-

tions, 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]

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 to preserve 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 based on a linear-regression model replaces the missing value, y_{mis}, by its best linear predicted value, $X\widehat{\beta}$, plus a random draw, $\widehat{u}$, coming from the 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

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

[5] If the number of households were 2 and 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 (the second variable is missing for the first household and the third for the second household). 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 iterations of the imputation process.

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 imputation 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.

Section 4 explains in more detail the kind of covariates we need to include in the imputation models to preserve the relationships among survey variables and the characteristics of the distribution of the complete data. The matrix X should allow for a large number of covariates in the imputation models: to control for non-response and satisfy the MAR assumption and the ignorable missing data mechanism, to include good predictors of both the variable to impute and the possibly missing covariates, and to use covariates highly related to the variable to impute according to different economic models.

Multiple imputation The EFF, like the SCF, provides multiply imputed values of the missing data instead of one single value, to reflect the uncertainty about the imputed values under one model for non-response. 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 imputed values. For each missing value of each variable k, $y_{mis,ik}$, we have m imputed values, $\widehat{y}^{(1)}_{mis,ik}, \cdots, \widehat{y}^{(m)}_{mis,ik}$. The difference between them depends on the degree of uncertainty 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 in

the EFF, where $m = 5$:

$$\begin{array}{lcc} & Y_{obs} & \widehat{Y}_{mis} \\ \text{Data set 1:} & Y_{obs,1} & \widehat{Y}^{(1)}_{mis,1} \\ & Y_{obs,2} & \widehat{Y}^{(1)}_{mis,2} \\ & \vdots & \vdots \\ & Y_{obs,N} & \widehat{Y}^{(1)}_{mis,N} \end{array} \Bigg\} \rightarrow \widehat{Q}^{(1)}, U^{(1)}$$
$$\vdots \qquad (5)$$
$$\begin{array}{lcc} & Y_{obs} & \widehat{Y}_{mis} \\ \text{Data set 5:} & Y_{obs,1} & \widehat{Y}^{(5)}_{mis,1} \\ & Y_{obs,2} & \widehat{Y}^{(5)}_{mis,2} \\ & \vdots & \vdots \\ & Y_{obs,N} & \widehat{Y}^{(5)}_{mis,N} \end{array} \Bigg\} \rightarrow \widehat{Q}^{(5)}, U^{(5)}$$

Let $\widehat{Q}^{(s)}$ and $U^{(s)}$ be the vector of statistics of interest and its estimated variance-covariance matrix from the complete data set s. Following Little and Rubin (1987), Schafer (1997) 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:

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

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.

Schafer (1997) asserts that in large samples the proper distribution on which to make inferences using the combined statistic, $\overline{Q}$, is not normal for small m. When $\overline{Q}$ is a scalar, its asymptotic distribution is approximated by a t-distribution with ν degrees of freedom, $\nu = (m-1)\left[1+\frac{\overline{U}}{(1+m^{-1})B}\right]^2$. For multivariate estimands and small m, Schafer (1997) says that the between-imputation covariance matrix, B, is very noisy, whereby matrix T is not a proper estimate of total variance. A more reliable estimate of total variance is given by Li *et al.* (1991) as $\widetilde{T} = (1+r_1)\,\overline{U}$, where $r_1 = (1+m^{-1})\,tr\left(B\overline{U}^{-1}\right)\Big/ k$ is the average relative increase in variance due to non-response and k the number of parameters. Li *et al.* (1991) also provide the best approximation for the degrees of freedom of the asymptotic F-distribution, $F_{k,\,\nu_1}$, of the statistic over which to test hypotheses about Q_0 [$\Pr\left(F_{k,\nu_1} > \left(\overline{Q}-Q_0\right)'\widetilde{T}^{-1}\left(\overline{Q}-Q_0\right)/k\right)$]. The degrees of freedom, ν_1, are calculated as follows: $\nu_1 = 4+(t-4)\left[1+(1-2t^{-1})\,r_1^{-1}\right]^2$ where $t = k\,(m-1)$.

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):

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

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 actually known or observed. Then, we start another iteration, $t+1$, repeating both steps until the convergence of the process (when

the imputed data and the parameter estimates are expected to converge in distribution). 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. Appendix A provides more detailed information about this sequential and iterative imputation process.

In all iterations, variables are imputed sequentially; the values imputed for one variable are used to impute the remaining variables in the I-step. Thus, the choice of the order in which the variables are imputed sequentially within the same iteration is not innocuous; 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. The order in which the variables are imputed sequentially matters. For 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.

Functional form of the imputation models Concerning the imputation model, we distinguish three different types of variables using the SCF multiple imputation macro programs written by Arthur Kennickell: 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 of the imputation model, missing information on continuous variables is imputed as follows:

$$\begin{gathered} y = X\beta + u, \quad u \mid X \sim N\left(0, \sigma^2 I\right) \\ \widehat{y}_{mis} = X\widehat{\beta} + \widehat{u},\ \widehat{u} \mid X \sim N\left(0, \widehat{\sigma}^2 I\right) \\ \widehat{\beta} = \left(X'X\right)^{-1} X'y, \quad \widehat{\sigma}^2 = \frac{1}{n}\left(y'y - y'X(X'X)^{-1}X'y\right) \end{gathered} \tag{8}$$

To impute the EFF survey, we do not estimate imputation models by maximum likelihood, non-parametrically or non-linearly due to the enormous costs in terms of effort and

time, since there is a huge number of very different patterns of item missingness among the household covariates in large surveys like the EFF and the SCF. This is the reason why we restrict the imputation of continuous variables to randomised linear-regression models such as that in equation (8), since we can accommodate very easily a huge number of different patterns of item missingness across households, 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:

$$\widehat{\beta} = \begin{pmatrix} x_1'x_1 & x_1'x_2 & x_1'x_3 \\ x_2'x_1 & x_2'x_2 & x_2'x_3 \\ x_3'x_1 & x_3'x_2 & x_3'x_3 \end{pmatrix}^{-1} \begin{pmatrix} x_1'y \\ x_2'y \\ x_3'y \end{pmatrix} \tag{9}$$

However, if the second covariate, x_{i2}, is missing for household i, the imputation model of the variable, $y_{mis,i}$, should be based only 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 (9), as follows:

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

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 (8). For imputing the missing value of household i, $\widehat{y}_{mis,i}$, we reshape these matrices 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 dif-

ferent patterns of item missingness, particularly when the number of covariates is very high (from 100 to 200 in most imputation models).[6] It is as if we implement imputation models individually for each non-respondent household according to its pattern of missing covariates. In the example above, for the non-respondent household i, we need to estimate the parameters, γ, of the imputation model that only uses x_1 and x_3 as covariates and we also need to estimate the variance, ω^2, of the error term, v, implied by the new model.

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 a large survey like the EFF.

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, depending on the sample size, we can use combinations between two or more discrete (or discretized) variables as the two covariates of the imputation model. The hot deck method assigns at random one value among the observed ones for households sharing the identical covariate values. Moreover, when the cell size of such households 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 combinations 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 one value randomly 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

[6] 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.

to each observation. The upper and lower bounds are constructed using the information declared by the households and the missing information already imputed sequentially, 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.

When the imputation model fails 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 More practical issues of the imputation models of the EFF data: covariates and specifications

This section explains more practical issues of the imputation models concerning both their covariates and specifications.

4.1 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 several reasons.

First group of covariates: determinants of the non-response 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*. In order to satisfy both assumptions, we should condition on a set of variables explaining or being related to the non-response together with other covariates, so that the MAR assumption does not become very strong.

For example, to impute wealth held in listed shares, this amount may be positively correlated with the missing household wealth, so the assumption that the distribution of the complete data (observed and missing) only depends on the observed data may be very restrictive. However, if we can use wealth strata indicators as covariates of the imputation models, this assumption becomes more acceptable; conditional on the household wealth strata, the distribution of the complete data may only depend on the observed data, but

not on the missing wealth.

Concerning the EFF data, variables that may be related to the non-response and that should be included in the first group are the following: total household income; 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: good predictors of the outcome variable 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.

To impute the amount of wealth invested in each asset individually, some covariates usually included are 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.[7]

Depending on both the sample size and the fraction of missing information on the values held in each asset, the values invested in the most common assets are generally used as covariates of the imputation model of the rest of assets. The most common assets are the main residence, other real estate properties, stocks, mutual funds and pension schemes.

[7]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.

Third group of covariates: economic variables possibly related to the outcome variable The third group of covariates that the imputation models should allow for is 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.[8] 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: good predictors of missing 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 of the groups of variables. Its role is very important, since variables are imputed sequentially based on both the observed data and the values of the previously imputed variables and there is a very large number of different patterns of item missingness across observations in the EFF.

For this reason, it is necessary to include variables that explain the covariates of the rest of 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, the imputation model will not be very poor, as the matrices, $(X'X)^{-1}$ and $(X'y)$, can be reshaped in equation (8) 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.

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

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 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. The imputation models must be very rich in terms of the number of covariates that we should include to take into account all these questions (determinants of the non-response, good predictors, economic variables possibly related to the variable of interest and good predictors of the missing covariates). For this reason, hot-deck imputations of continuous variables, such as income and wealth variables, may fail to preserve the relationships among the variables of the survey, if we can only control for a very limited number of household characteristics. This problem becomes more severe if the sample size involved in the imputation of the wealth values of some particular assets is very small, since a low percentage of households owns the assets, such as listed and unlisted shares, life insurance policies, mutual funds and businesses.

4.2 Some specifications of the imputation models of continuous variables

To take into account nonlinearity in the imputation models of income and wealth variables, regressors may either be formed by interactions between variables or introduced in logarithms or as polynomials. To impute some euro questions that may have a zero

value, such as the balance of current accounts, the value of the land and buildings of the businesses that the household owns and the market value of the businesses, we first impute a binary variable indicating whether the value is positive or not and then we impute the positive values in logarithms. This is also done when the histogram of continuous variables shows some probability mass points.

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). 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.

Questions asked separately to each particular asset within an asset type Next, I will describe how we impute missing information about questions posed to households concerning their holdings in different mutual funds, pensions schemes, real estate properties, loans, etc. (such as questions 4.31, 5.7, 2.39 and 2.18 of the EFF survey, among others). The way of imputing these variables is to construct a pooling of subsamples defined for each asset within a given type. First, we generate the covariates of the imputation model separately for each asset; 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 are updated in the original data set.

Constructed total household income variables The EFF provides two constructed variables for total household income: 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. These two measures of income are imputed again by alternative imputation methods to study the effects of different ways of handling missing data in Section 5.

5 The effect of alternative methods of handling missing data on income and wealth

In this Section, I will study the effect of different ways of handling missing information by focussing on some relevant variables of income and wealth using the first wave of the EFF data. In particular, I will study four alternative ways of treating missing data: listwise deletion, non-stochastic imputation using linear regressions and stochastic imputation methods, such as single imputations by hot-deck procedures and single and multiple imputations based on randomised linear regressions. I will not only pay attention to the fact of whether the data are imputed stochastically or not and multiply or not, but also to the kind of imputation model, distinguishing among models based on linear regressions or non-parametric models based on hot-deck procedures. The variables on which this empirical analysis is based are:

(i) The total household income earned in 2001 and the income received in the month in which the interview took place.

(ii) The variables on the amounts of debts and wealth held in each of the assets used to construct the households' net wealth and some wealth and debt ratios.

(iii) The non-durable consumption and the current value of the stock of durable goods to construct a measure of total consumption as done in Bover (2005) and a measure of the household saving rate similar to that carried out in Dynan *et al.* (2004).

Subsection 5.1 explains the analysis and how I impute the income, wealth and consumption variables non-stochastically and by hot-deck procedures. Subsection 5.2 shows descriptive statistics of the marginal and conditional distributions of income and wealth variables that are obtained with different methods of handling missing data. Finally, in Subsection 5.3 I discuss the main results obtained when using income, wealth and consumption variables imputed in various ways to estimate linear and quantile regression models. The empirical approach I follow to estimate the models is very similar to that implemented by Dynan *et al.* (2004) to study the relationship between the household saving rate and income using median regression estimates.

5.1 Empirical implementation

In this analysis, I impute the two income variables constructed in the EFF and the amounts of debts and wealth values held in each asset conditional on its ownership, but not the indicators of whether the household has the asset (or the debt). However, there is also item non-response in some ownership indicators, such as in accounts, deposits, shares and mutual funds participations, as well as missing information on many covariates used in the imputation models. I do not consider missing information on these variables due to the large number of item missingness patterns we should take into account and which would make the implementation of this study very complicated unnecessarily.

Under the impossibility of an exact match of the EFF data to registered data to make comparisons [as David *et al.* (1986) do], the approach I follow is to consider the first data set imputed multiply as if it was the "true" data sample, in which there is only item non-response in the total income variables, in the consumption of different goods (non-durables, vehicles and housing equipment), in the value of both debts pending to repayment and asset holdings that the households declare to have or that are previously imputed to hold in the EFF. More precisely, using this sample I impute non-stochastically using linear-regression models and stochastically by hot-deck procedures and compare the results to the stochastic imputation regression procedures already obtained for the EFF, as if the latter would provide the true data.[9] The analysis of the different ways of handling missing data is divided into three steps:

(i) Compare the estimates using the listwise deletion method with the estimates that come from the multiple imputation regression procedures already obtained for the EFF.

(ii) Compare the estimates coming from both the single and multiple randomised regressions already imputed for the EFF. The single imputations correspond to the ones in the first of the multiply imputed data sets in the EFF.

[9] To impute the two constructed total household income variables provided by the EFF, I restrict the imputed income values to exceeding the income declared by the household in the EFF. When the non-stochastic and hot-deck imputed values lie under this lower bound, I impute them the declared income.

(iii) Compare the estimates obtained using non-stochastic regressions and single hot-deck imputation methods with the estimates coming from the single imputation regression procedure.

The non-stochastic and hot-deck imputations obtained for the first data set could be also evaluated and compared with the multiple imputation of the EFF data methods using the combined statistic in equation (6). As far as the multiply imputed samples have converged in distribution during the EFF data imputation process, the five data sets imputed multiply will provide consistent estimates of the statistics obtained using single and multiple imputations. However, the estimates obtained using non-stochastic and hot-deck imputations should not be compared directly with the estimates of the listwise deletion method, since these two imputation methods and the listwise deletion are applied on different data samples (the imputation methods on the first data set imputed stochastically and the listwise deletion on the sample of observed data).

For the non-stochastic imputation, I use the same specifications and covariates included in the EFF stochastic imputation models, but without adding the random number from the normal distribution in equation (4). For hot-deck imputation, I include the most significant categorical covariates using combinations of variables. Note that hot-deck imputation is also a stochastic imputation method, since the value imputed is chosen randomly from the cell formed by households having the same characteristics as the non-respondent household.[10] In this empirical analysis, when I say multiple imputation and single imputation, I always refer to continuous variables that are imputed stochastically as hot-deck procedures do, but using linear regression models that allow for a wide range of covariates.

In non-stochastic and hot-deck imputations, income, debt, wealth and consumption variables are imputed sequentially, but not iteratively as done in the algorithm of the EFF multiple imputation [see equation (7)]. In the first data set imputed multiply by the EFF, I convert the imputed values for these continuous variables into missing values, and then I reimpute them sequentially by the alternative methods. In the non-stochastic

[10] In this empirical analysis, hot-deck imputation is done in Stata using the ado files written by Mander and Clayton (STB_54: sg116.1).

imputation method based on linear regression models, I impute each missing observation individually. This is done by estimating as many linear regression models as needed to accommodate the different patterns of item missingness among the income, wealth and consumption covariates of the imputation models, in a similar way to that explained in the example of Section 3.

As the sample used to impute by these alternative imputation methods is the first data set imputed multiply, I expect that the differences obtained in the sample distributions of wealth and income variables between the various imputation methods are only a lower bound of those one would encounter if applying the imputation methods to all variables at all stages of the imputation process. The differences would have been much larger if I had imputed by hot deck procedures using the imputation algorithm in Section 3 and by non-stochastic imputation using the subsample of observed data. Finally, the imputation process by hot-deck is not done iteratively, since the sample data used to impute have already converged in distribution across the iterations of the multiple imputation process for the EFF. Therefore, the gain of imputing by hot-deck procedures iteratively in this empirical implementation is very small.

5.2 Descriptive statistics

To construct savings and net worth variables, the wealth and debt variables are defined in a similar way to that in Banco de España (2005a). Total consumption is obtained as the sum of non-durable expenditures and the consumption of fixed proportions of the current values of the stock of durable goods (house equipment and vehicles), as Bover (2005) does. All monetary questions are deflated in 2001 euros. The household saving rate is constructed in a similar way to Dynan *et al.* (2004).

I use an active saving measure that excludes capital gains. Annual saving is constructed as the difference between the monthly average income earned in the year in which the interview took place and the monthly average total consumption, multiplied by 12. The household saving rate is obtained by dividing saving by the annual average of the household total income earned in the two periods reported in the 2002 wave of the EFF [2001 and the year of interview (2002/03)]. Following Dynan *et al.* (2004), in the denominator

of the saving rate, I use the annual average of the total household income, since it may be a better proxy for permanent income due to the fact that the average is less affected by income transitory shocks than the earnings during a particular year. However, this measure of household saving is very noisy, as wealth surveys underestimate and are not meant to exhaustively measure the total household consumption, compared with diary surveys like the *Continuous Family Expenditure Survey* (in Spanish, *Encuesta Continua de Presupuestos Familiares*) in Spain [see Browning *et al.* (2003) for a discussion].

The performance of the alternative methods of dealing with missing data, analysed here, is studied looking at: first, descriptive statistics of the marginal and conditional distributions of the income and wealth variables, and second, the estimates of mean and quantile regressions of some common wealth and debt ratios, such as the saving rate, the net worth to income ratio, the financial burden ratio and the loan to value ratio.

Marginal distributions of income and wealth variables Figure 1 shows Epanechnikov kernel density estimates of the logarithms of the household total income in 2001, wealth held in real assets, financial wealth and net wealth. The kernels are evaluated in the same points of the variable across the four methods for handling missing data; for the multiple imputation regression method, the kernel is obtained by averaging the kernel function values of the five multiply imputed data sets in each point.

The kernel estimates show three features. First, the non-stochastic and hot-deck imputation methods only differ significantly from the multiple imputation in the kernel estimates for the total income and wealth held in financial assets. This is due, first, to the fact that I have imputed using the EFF final data set instead of the subsample of observed data, and second, to the fact that imputation methods seem to make less of a difference for variables having lower percentages of missing information (28.7% in wealth held in real assets, 39.4% in financial wealth and 51.9% in total income). The proportion of missing information in net wealth is also high (51.2%), but the shape of the distribution of the net wealth is similar to the pattern of the distribution of wealth in real assets due to the fact that a large proportion of the Spanish households' wealth is invested in real estate, which exhibits a low rate of item non-response. Around the 79.2% of the total

gross wealth is invested in main residences and other real estate properties, according to the estimates from the EFF [see Box 5 on page 62-63, Banco de España (2005b) and Bover *et al.* (2005)]. Another reason why the kernel estimates of the wealth held in real assets do not differ greatly across alternative imputation methods is that the information available for imputing real assets is much more detailed and rich than that available for imputing financial assets.

The second feature highlighted by the kernel estimates is that the distribution of the multiply imputed wealth and income variables are skewed to the right with respect to that of the observed data, mainly for the total income, which is consistent with the fact that item non-response rates on income and wealth variables are not random and are positively correlated with the households' wealth and income. The comparatively high skewness of the income distribution may be also the result of the high rate of item non-response for total income with respect to its counterpart in wealth held in real assets. Finally, the third feature is that non-stochastic imputation seems to make the distributions more peaked around the mean, mainly in financial wealth and income, while hot-deck imputation seems to preserve the dispersion of the marginal distributions of variables.

Figure 2 shows the kernel estimates of the logarithm of the average total household income, consumption, savings and monthly loan payments. The kernel density of the average household income does not only reproduce the same pattern as that of the total income earned in 2001, but also that of the estimated density of the total income received in the year of the interview (not shown in the paper). The estimated kernel density of this income variable also varies across alternative methods of handling missing data, as the kernel densities of household savings and total consumption show (saving is constructed as the income earned at the year of the interview minus total consumption). On the contrary, the density of total consumption and monthly loan payments estimated using the different imputation methods do not differ considerably, due to their low rates of item non-response (19.1% and 8.7%, respectively).

Finally, Figure 3 shows the kernel estimates of the wealth and debt ratios (in logarithms) that will be included in the regression estimates to analyse the performance of the

different methods of treating missing data. These are the net wealth to average income ratio, the household saving rate, the per cent ratio of the loan value to gross wealth (the loan to value ratio), and the financial burden ratio (defined as the percentage of the annual loan payments over the average total household income). Multiple imputation based on randomised regressions shifts the distribution of the ratios towards one of the tails with respect to the distribution of the observed data; it is shifted considerably to the right for the saving rate and to the left tail for the rest of ratios, mainly for the financial burden ratio. All the differences in the distributions are mainly caused by income and wealth. Indeed, the distributions of the loan value pending repayment and the loan payments are very similar across alternative methods of handling missing data due to the extremely low item non-response rates (below 13.5%) and to the fact that the loan characteristics, such as the initial value of the loan and the period length until the loan is fully repaid, are very helpful in imputing these variables. Thus, alternative imputation methods are expected not to make a difference.

Table 1 shows various descriptive statistics for all these variables. Descriptive statistics are weighted to compensate for the unequal probability of the household being selected due to the oversampling of the wealthy, geographical stratification and unit non-response. Results are presented as ratios of percentiles over the median. Larger ratios of the percentiles below the median mean that these percentiles are nearer the median and there is less dispersion; the same applies to the percentiles over the median when their ratio takes smaller values. To compare descriptive statistics across the various methods of dealing with missing data, I contrast the equality of estimates obtained using the multiple imputation method with the estimates obtained by each of the rest of methods using 500 bootstrap replications and taking into account that the samples across alternative methods are dependent.

Generally, the dispersion of the distribution of the variables imputed by hot-deck is very similar to that of their counterparts imputed multiply using linear regression models; the dispersion obtained by hot deck only seems to be higher in the upper tails of the loan to value ratio and the saving rate, and lower in both measures of total household

income shown in Table 1. On the contrary, as expected, the dispersion diminishes in non-stochastically imputed variables, being significant in the total household income earned in 2001, in the loan to value ratio and in the lowest tail of the financial burden ratio. In non-stochastic and hot-deck imputation methods, the descriptive statistics of saving rate are statistically different from those obtained by the multiple imputation method. Comparing the listwise deletion and multiple imputation methods, the dispersion of the distributions is quite similar or lower in the imputed variables (except for the financial burden ratio and the top tail of the distribution of the total household income). However, the medians of the total income and wealth variables are considerably larger for imputed data.

Moreover, when we study income and wealth inequality looking at the shares of the top 1%, 5% and 10% of the richest population (in terms of the variable of interest) and Gini indices, these measures of inequality increase significantly after imputing the data (mainly in net wealth with multiple randomised regression and hot-deck imputation methods). This indicates that item non-response rates on income and wealth variables are not random, but positively correlated with them, whereby the listwise deletion method may bring a non-sensible measure of the inequality in income and wealth distributions. Thus, the way of dealing with missing data may yield very different results in our empirical studies in terms of income and wealth inequality and dispersion of the distributions.

Conditional distributions of income and wealth variables When looking at income and wealth distribution broken down by percentiles and conditional on household characteristics, such as the family head's age, education and labour status, I find that income and wealth variables generally follow similar patterns across methods of handling missing data, although the differences encountered in the marginal distributions remain in the conditional distributions. Table 2 shows weighted descriptive statistics of the conditional distribution of total income; dispersion is measured by the ratio of the interquartile range to the 25th percentile $[\frac{(p75-p25)}{p25}]$, i.e. the number of times that the difference between the 75th and 25th percentiles is in terms of the 25th percentile. Medians are higher with multiple imputation method than with the rest of methods of treating missing data, and

the dispersion is almost always lower with non-stochastic imputation and is more similar with the hot-deck imputation. Moreover, the pattern of median income according to household characteristics bears more similarities in the samples imputed multiply and non-stochastically than those in the sample imputed by hot-deck procedures. Particularly, the imputation method by hot deck fails to reproduce the inverted-U shape of the age-income profile and shows a flat income pattern among households with family heads aged under 65.

The reason why hot-deck may fail to impute higher income values is that it cannot support a large number of covariates to take into account observed household heterogeneity, in contrast to the other two imputation methods. Household income, family head's age, education and labour status are usually included as covariates in the hot-deck imputations; however, if the sample size is very small, some of these covariates must be excluded from the model. This may explain why sometimes hot-deck imputation does not seem to follow the same pattern across household characteristics, for example, in the conditional distribution of the financial burden ratio across household net wealth (see Table 3).

5.3 Regression results

In this subsection, I investigate the consequences of the alternative methods of handling missing data in the estimation of some conditional models based on the mean and the 0.25, 0.50 and 0.75 quantiles. Mean regressions robust to heteroskedasticity and quantile regressions are estimated for the following saving and debt variables: the ratio of the household's net worth to average income, the saving rate, the financial burden ratio and the loan to value ratio.

$$\begin{aligned} \text{Mean} \quad &: \quad y_j = X_j\beta_j + v_j, \\ \alpha^{th} \text{ quantile} \quad &: \quad y_j = X_j\beta_{\alpha,j} + \varepsilon_{\alpha,j}, \\ \alpha \quad &= \quad 25,\, 50,\, 75,\; j = LD,\, MI,\, S,\, NS,\, HD \end{aligned}$$

Listwise deletion estimates are denoted by LD, multiple imputation by MI, single imputation by S, non-stochastic imputation by NS and hot-deck by HD; y_j is the $N \times 1$

dependent variable where missing data have been handled according to method j; N is the number of observations; X_j is a $N \times k$ matrix containing the set of k regressors; β denotes the parameter vector in each regression; v_j and $\varepsilon_{\alpha,j}$ are the error terms in the mean regression and in the αth quantile regression for method j, respectively. Multiple imputation estimates combine the estimates made separately over the five samples imputed multiply using the formulae in equation (6).

To analyse the saving and debt variables, the empirical approach I follow is similar to the one in Dynan *et al.* (2004) who estimate median regression models to study the relationship between income and household saving rates. To focus on a homogeneous group of population having a similar behaviour towards saving and debt, the sample is restricted to households with family heads aged 30-59 and with household annual earnings over 1,000. Earnings and other monetary variables are deflated with consumer price indices based in 2001.

Following Dynan *et al.* (2004), the estimation strategy consists of a two-stage procedure. In the first stage, current income measured by the annual average income of both periods (2001 and 2002/03) is regressed on proxies for permanent income and family head age bands. The age groups considered are 40 to 49, 50 to 59 and the omitted category is of 30 to 39. Fitted values from this first-stage regression are used to place households into predicted permanent income quintiles that will be included as explanatory variables in the second stage. This is done to remove measurement errors and biases from the estimation of the relationship between income and wealth in the second stage. The quintiles of predicted permanent income are constructed separately for each age band.

In the second stage, the mean and quantile regressions for the household saving rates, net wealth to income ratio, financial burden and loan to value ratio are estimated. Independent variables include an intercept, predicted income quantiles (omitted the first), age bands (omitted the age group of 30-39) and predicted income (divided by 10,000). This two-stage procedure is estimated using the data obtained from each of the five methods of handling missing data analysed ($j = LD$, MI, S, NS, HD). Standard errors are obtained by bootstrapping the two-stage process with 500 replications. The standard errors

and t-ratios of the combined estimators for the multiply imputed samples are constructed according to the procedure proposed by Li *et al.* (1991). Tables 4 to 6 show unweighted regression estimates, since the sample over the regressions are estimated has been selected and the weights are only constructed to represent the whole sample. There is a public discussion about the use of weights in such cases [see Deaton (1997) and Cameron and Trivedi (2005)].

Median of the net worth to income ratio Table 4 shows the results of unweighted median regression for the ratio of the household net worth to income. Comparing the listwise deletion estimates with those obtained with multiple imputation, I find that, when the multiple imputation estimates are significant, the t-ratios associated with listwise deletion are significantly lower (this is true for all mean and quantile regression models estimated here). For example, the listwise deletion estimates would indicate that the median of the household worth to income ratio follows a flatter pattern according to household head's age, since the coefficient estimates associated with the age group of 40-49 is not significantly different from the omitted group of 30-39. However, the multiple imputation estimates show a profile increasing in age, being significant at the 1% level. The less precise estimates obtained with the listwise deletion method may be due to the very reduced sample size after deletion; in this case we are left with 18.5% of the sample available in the case of randomised regression models for imputing multiply. With such a small size, most of the explanatory variables are not significant, and therefore the listwise deletion method could not support a wide range of empirical studies concerning household income and portfolio choice.

Moreover, the listwise deletion and multiple imputation methods estimate different patterns of the net worth to income ratio according to income. The net worth ratio increases monotonically with income at the 10% level in the sample with missing cases deleted. However, the multiple imputation method shows that the median of this ratio only increases slightly with income only at the intermediate quintiles, but not monotonically.

Regarding the non-stochastical linear regression imputation method, this seems to

provide artificially more precise estimates than multiple imputation because it ignores the uncertainty about the values predicted to replace missing data. In this way, this method makes the distribution of the complete data (observed and imputed) more peaked around the conditional mean. As a consequence, standard errors associated to some variables are underestimated.

In contrast, the hot-deck imputation method estimates a different relationship between the net worth ratio and household income. The estimates exhibit a more significant and clear relationship between the net worth ratio and the income quintiles and also suggest that the net worth ratio increases monotonically with income, unlike the multiple imputation method based on randomised regressions. This may be due to the fact that hot-deck procedures impute randomly and non-parametrically missing data using a limited number of covariates. Household income and age of household head are two covariates present in most of the hot-deck procedures for imputing income and wealth. However, a large number of other household characteristics must be left out of the imputation model or are not included very exhaustively, such as the household composition, portfolio composition, spouse's demographic characteristics and other variables related to labour market and labour history of the couple.

Concerning stochastic imputation methods, Table 4 shows that the values estimated with multiple imputation and single imputation (using the first EFF data set) are very similar, the only difference being that standard errors are usually lower in single imputation, since it takes the imputed value as if it was actually observed and does not take into account the uncertainty about the imputed value [the between-imputation variance, B, in equation (6)]. For example, this may lead to consider that the median of the net worth ratio increases steadily with income from the third quintile using single imputation and also grows monotonically with income, when actually the income pattern is rather flatter according to the multiple imputation estimates.

Finally, Table 4 shows Hausman statistics that test for the equality of the parameter estimates between two different methods. Under the null hypotheses, $H_0: \beta_{MI} = \beta_j,\ j = LD,\ S$, the test statistics contrast the equality of the estimates obtained using the samples imputed multiply with those using the subsample of observed data ($H_0: \beta_{MI} = \beta_{LD}$) and

the single imputation method based on randomised regressions ($H_0: \ \beta_{MI} = \beta_S$). The other two Hausman statistics test for the equality of the parameter estimates obtained using single imputation based on randomised regressions with those using non-stochastic imputation ($H_0: \ \beta_S = \beta_{NS}$) and hot-deck procedures ($H_0: \ \beta_S = \beta_{HD}$). These are compared to single imputation estimates, due to the fact that the non-stochastic and hot-deck imputation methods are applied to the sample that arises from the first data set after converting into missing the imputed values of continuous variables of total income, debt, consumption and wealth. The covariance matrix between the two estimators for the Hausman test, $V = E[\left(\widehat{\beta}_j - \widehat{\beta}_S\right)\left(\widehat{\beta}_j - \widehat{\beta}_S\right)' \mid X_j, \ X_S]$, is estimated using bootstrap.

The values of these tests provide evidence of the equality of the single and multiple imputation estimates based on randomised regressions (as expected), but reject the null hypotheses of equality among the estimates obtained by the other methods of handling missing data except for the listwise deletion method. The estimates obtained by the listwise deletion seem to follow the same pattern as those by multiple imputation, but the estimates are very imprecise. This may be the reason why Hausman statistics do not reject the null hypothesis of equality of estimates.

The 25^{th} and 75^{th} quantile regressions of the net worth to income ratio In order to investigate whether the alternative methods of handling missing data typically differ at different quantiles, Table 5 shows the estimates of the quantile regressions for the 25 and 75 quantiles using the same specification as for the median regression. Quantile regressions are not estimated jointly due to the small sample size. Once again the listwise deletion method shows less precise estimates and not sensible estimates of the patterns between variables compared with the multiple imputation. Moreover, the 25th quantile estimates provide evidence that the listwise deletion may bias the results, since the listwise deletion estimates indicate that the net worth ratio increases monotonically with one additional euro of earnings, while the multiple imputation estimates show that the net worth ratio grows with the income quintile, but not monotonically within quintile. In both quantiles, the listwise deletion method does not estimate a steadily increasing pattern of the net worth ratio according to age.

Once again, the non-stochastic imputation method provides more precise estimates with frequently higher t-ratios in the significant explanatory variables than those obtained in the data imputed by randomised regressions. Moreover, the absolute values of the estimated coefficients with the non-stochastic imputation are also higher than those with the randomised imputations. This is due to the fact that the non-stochastic imputation method imputes missing data with conditional mean values, which makes the distribution of the data more peaked around the mean. This leads the non-stochastic method to estimate artificially stronger links among variables than those that arise from the stochastic imputation methods.

Concerning hot-deck imputations, this method seems to fail to reproduce the pattern of the net worth ratio for income and age. In both quantiles, the increasing age profile of the net worth ratio is stronger. Regarding income at the 25^{th} quantile, net worth ratio increases with income quintile and monotonically within quintile, while the multiple imputation estimates indicate that it does not vary with household income.

Finally, as single imputation treats the imputed values as if they were observed ones and does not take into account the uncertainty associated with them, sometimes it fails to establish the significance of the variables suitably. For example, in the 25^{th} quantile regression, the estimates also indicate profiles increasing in income for the net worth ratio at the 1% level.

Various variables: tests of equality of the coefficient estimates across the different methods of handling missing data To investigate further the rejection of the equality of coefficient estimates across alternative methods of handling missing data, Table 7 shows the p-values of Hausman tests applied to the mean and quantile regressions. Both the weighted and unweighted regressions for net worth to income ratio, saving rate, financial burden ratio and loan to value ratio have the same specifications as above. The unweighted mean and median regression estimates of these ratios are shown in Table 6 for the multiple imputation method. In general, the mean regressions of wealth and debt ratios on predicted income quintiles and age bands are poor, mainly for debt ratios. These estimates suggest that the means of these debt ratios seem not to depend greatly

on income nor to vary with income quintiles. One typical feature of these wealth and debt ratios is the presence of a high proportion of extreme values due to measurement errors, which makes it more convenient to estimate quantile regressions. The quantile regressions reveal that these ratios vary considerably with the income quintiles and the relationship is different across quantiles.

The household saving rate defined as in this paper seems not to be a good measure or proxy of the household's savings and wealth due to the presence of large measurement errors in the savings variable. In almost all estimates obtained with the different methods of handling missing data, age is not significant in explaining the household saving rate. On the contrary, the estimates indicate that the financial burden ratio and the loan to value ratio are decreasing with family head's age and the decline is more steady at higher quantiles. The financial burden ratio also decreases with household income quintile, but the loan to value ratio does not depend on the household income, except for the 75^{th} quantile of the households more indebted (not shown in the paper) that exhibits the same decreasing income profile.

Table 7 shows that Hausman statistics provide evidence for the equality between the estimates obtained by multiple and single imputations based on randomised regressions. As shown by the p-values of the stastistics, Hausman test cannot reject the equality of the estimates in the subsample of observed data, mainly for the net worth ratio regressions. However, the listwise deletion method does not provide statistically significant estimates due to the small sample size again. The estimates across alternative imputation methods seem to differ greatly with single imputation in terms of Hausman statistics, except for the loan to value ratio. Non-stochastic imputation makes the variables involved in the models more dependent among themselves and also yields more significant estimates, since this method imputes missing data with conditional means and underestimate the dispersion of the complete data (observed and imputed). Finally, the differences between hot-deck and randomised regression imputations may be due to the inability of the former to capture the correlation of the variable of interest with a large number of covariates in the imputation models, which distorts the relationships among the variables.

6 Conclusions

One typical feature of wealth surveys is the presence of high rates of item non-response due to the lack of knowledge or to the unwillingness of households to reveal certain information about their income and wealth. As item non-response rates are not random but depend on household characteristics, studies carried out by users of the data may be misleading if this missing information is not imputed. Using the 2002 wave of the *Spanish Survey of Household Finances* (in Spanish, *Encuesta Financiera de las Familias*, EFF), I analyse the performance of alternative methods of handling missing data, specifically, listwise deletion, non-stochastic imputation and hot-deck imputation as compared to the multiple imputation method used in the first wave of the EFF.

The goal of this paper is to emphasise the importance of the way of handling missing data and the impact of the method used on the outcome of empirical studies. Indeed, the results obtained with alternative methods may be very different in terms of the inequality of income and wealth, dispersion of the distributions, and the potential relationships among variables.

For this purpose, using the first data set of the EFF multiply imputed, I impute non-stochastically household income, wealth, debt and consumption variables using the same specifications and covariates of the EFF multiple imputation models in order to evaluate the performance of the stochastic versus non-stochastic imputation methods. Similarly, I choose the most significant covariates of the models in order to impute income and wealth variables by hot-deck procedures. The number of covariates selected depends on the fraction of missing information and the sample size available to impute the variable of interest, since the cells in which the covariates split the sample must not be excessively small. Thus, the main drawbacks of the hot-deck imputations are the necessity of selecting carefully few key covariates, because of the inability of including a wide range of covariates to preserve the relationships and correlations among variables. The listwise deletion method removes from the sample all those observations having missing data in the variables used in the empirical analysis.

To compare the various methods of treating missing data, I show weighted and un-

weighted descriptive statistics of the marginal and conditional distributions of the household total income, net wealth, financial wealth, value of holdings in real assets, and debt ratios. Moreover, I also show some mean regressions and quantiles regressions, in order to analyse the similarities and differences of the estimates across methods.

There are five main results highlighted in this paper. First, the listwise deletion may bring imprecise estimates due to small sample sizes after deletion, if the fraction of missing information is high. Second, more seriously, the listwise deletion method may lead to a severe bias and non-sensible analyses if the item non-response rates are not random and correlated with the variables of interest, income and wealth in this case. Third, non-stochastic imputation method makes the distribution of the complete data (observed and imputed) more peaked around the means of variables and underestimates the dispersion of the distributions, which leads to overestimate more significant and stronger relationships among variables. Fourth, the hot-deck imputation method helps preserve the dispersion of the distributions, but fails to reproduce some patterns of income and wealth variables according to household characteristics. This is probably because the hot-deck imputation models cannot allow for a large number of covariates in order to preserve the relationships among the variables. This fact seems to be crucial for imputing the household total income. Finally, single imputation may yield misleading information about the significance of the regressors. In particular, it underestimates variances since it treats the imputed values as if they were the actual ones and therefore does not take into account the uncertainty about the imputation models used (the between-imputation variance).

A Appendix

The imputation process has a sequential and iterative structure based on data augmentation algorithm (see Tanner and Wong, 1987) and Markov chain Monte Carlo method, as explained in Section 3 and shown in Equation (A.1):

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

First iteration of the imputation process The first iteration of the imputation process basically differs from the rest of iterations in the way that the starting values of the parameters, $\widehat{\theta}^{(0)}$, are chosen, due to the fact that estimates of the model parameters are not available from a preceding iteration. Instead, the starting values, $\widehat{\theta}^{(0)}$, 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:

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

In the I-step of the first iteration, the initial value of θ_1, $\widehat{\theta}_1^{(0)}$, corresponds to the estimates from the 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 $\widehat{y}^{(1)}_{mis,1}$, respectively. This sequential process continues until imputing the last variable of the survey, $\widehat{y}^{(1)}_{mis,K}$. If the survey has K variables, the way of imputing in the first iteration is as follows:

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

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

$$\theta' \quad = \quad (\theta'_1 \, \theta'_2 \ldots \theta'_K) \tag{A.3}$$

In the first iteration of the imputation process, the imputed values of one variable are not only used to impute the remaining variables within the iteration (the I-step), but also to estimate the parameters of the imputation models of successive variables with missing information [see equation (A.2)]. In the I-step of whatever iteration, the imputed data are treated as if they were actually observed for imputing the remaining variables. The parameter vector, θ, collects all the parameter subvectors, θ_i $i = 1, \ldots K$, implied by the imputation model of each variable.

After the I-step, we implement the P-step as in equation (A.3): once all variables are imputed, the parameter vector of the imputation model, θ, is estimated for imputing missing information in the next iteration. Then, we start the second and the remaining iterations following the two steps described in equation (A.1). 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.

Finally, we implement two procedures specific to the implementation of the imputation of the EFF to ensure reasonably starting values of continuous variables that come from randomisation in the first iteration of the imputation process and to evaluate the convergence of the sample distributions (imputed stochastically) across iterations of the imputation process [see Bover (2004) and Barceló (2006) for details].

References

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

ATTANASIO, O. P. (1999). "Consumption", edited by: J. B. Taylor and M. Woodford, *Handbook of Macroeconomics,* vol. 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.

BARCELÓ, C. (2006). *Imputation of the 2002 wave of the Spanish Survey of Household Finances (EFF),* Occasional Paper No. 0603, Banco de España.

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

– (2005). *Wealth Effects on Consumption: Microeconometric Estimates from the Spanish Survey of Household Finances,* Working Paper No. 0522, Banco de España.

BOVER, O., C. MARTÍNEZ-CARRASCAL and P. VELILLA (2005). "The wealth of Spanish households: a microeconomic comparison with the United States, Italy and the United Kingdom", *Economic Bulletin,* July (English version) or April (Spanish version), Banco de España.

BROWNING, M., T. F. CROSSLEY and G. WEBER (2003). "Asking Consumption Questions in General Purpose Surveys", *Economic Journal,* 113, pp. F540-F567.

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.

DAVID, M., R. J. A. LITTLE, M. E. SAMUHEL and R. J. TRIEST (1986). "Alternative Methods for CPS Income Imputation", *Journal of the American Statistical Association,* vol. 81 (393), pp. 29-41.

DE LUCA, G., and F. PERACCHI (2007). *On estimating models with unit and item nonresponse from cross-sectional surveys,* mimeo.

DEATON, A. (1997). *The Analysis of Household Surveys,* The World Bank, The John Hopkins University Press.

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

DYNAN, K. E., J. SKINNER and S. P. ZELDES (2004). "Do the Rich Save More?", *Journal of Political Economy,* 112, pp. 397-444.

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.*

KOFMAN, P., and I. G. SHARPE (2003). "Using Multiple Imputation in the Analysis of Incomplete Observations in Finance", *Journal of Financial Econometrics,* 1, pp. 216-249.

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

– (2007). "An econometric method of correcting for unit nonresponse bias in surveys", *Journal of Econometrics,* 136, pp. 213-235.

LI, K. H., T. E. RAGHUNATHAN and D. B. RUBIN (1991). "Large-Sample Significance Levels from Multiply Imputed Data Using Moment-Based Statistics and an

F Reference Distribution", *Journal of the American Statistical Association,* 86, pp. 1065-1073.

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

LONGFORD, N. T., M. ELY, R. HARDY and M. E. J. WADSWORTH (2000). "Handling Missing Data in Diaries of Alcohol Consumption", *Journal of the Royal Statistical Society. Series A (Statistics in Society),* 163, pp. 381-402.

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.

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.

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).

Figure 1: Kernel density estimates of some variables of the households' income and wealth.

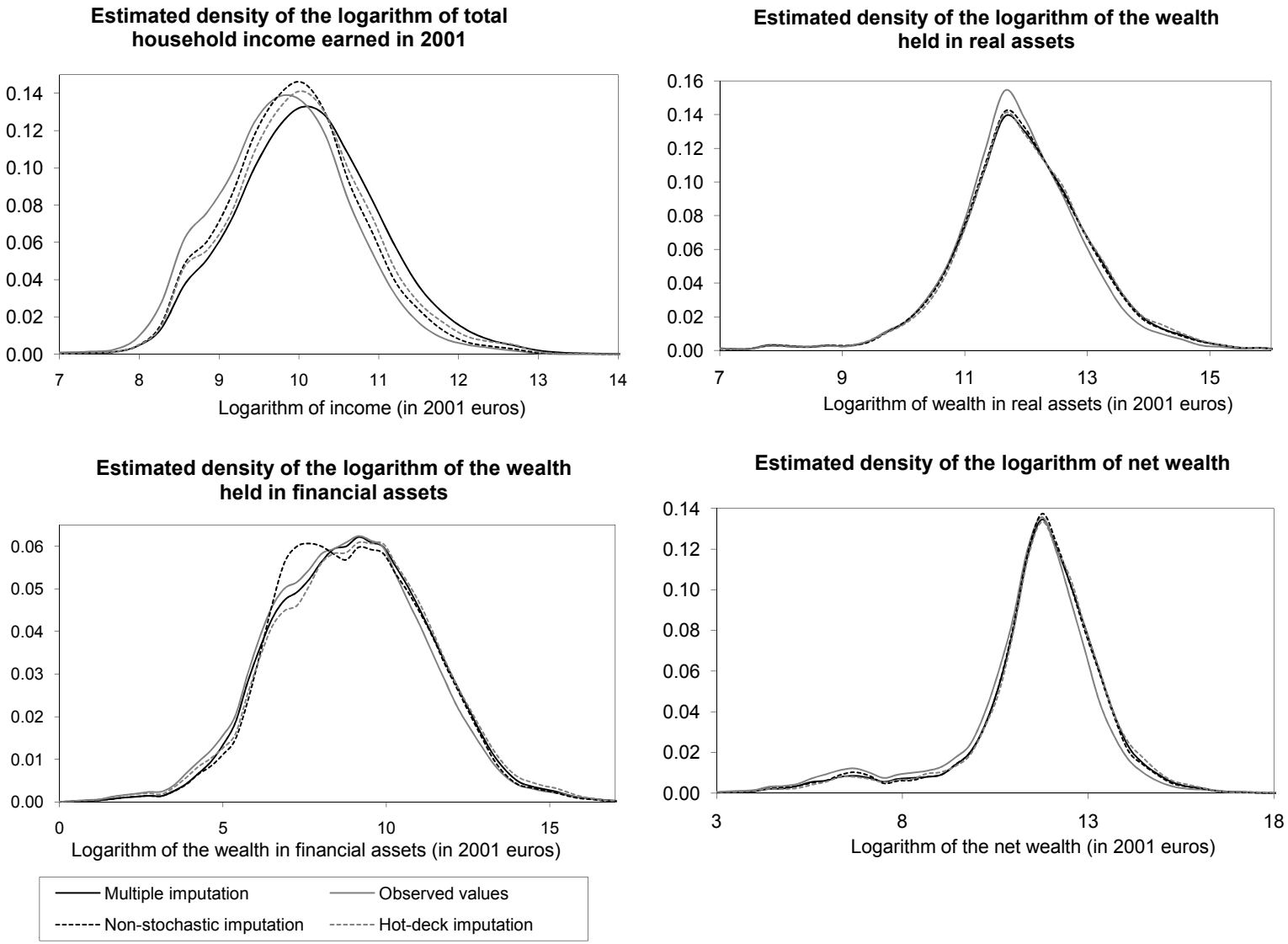

Figure 2: Kernel density estimates of some relevant variables of household wealth and debts.

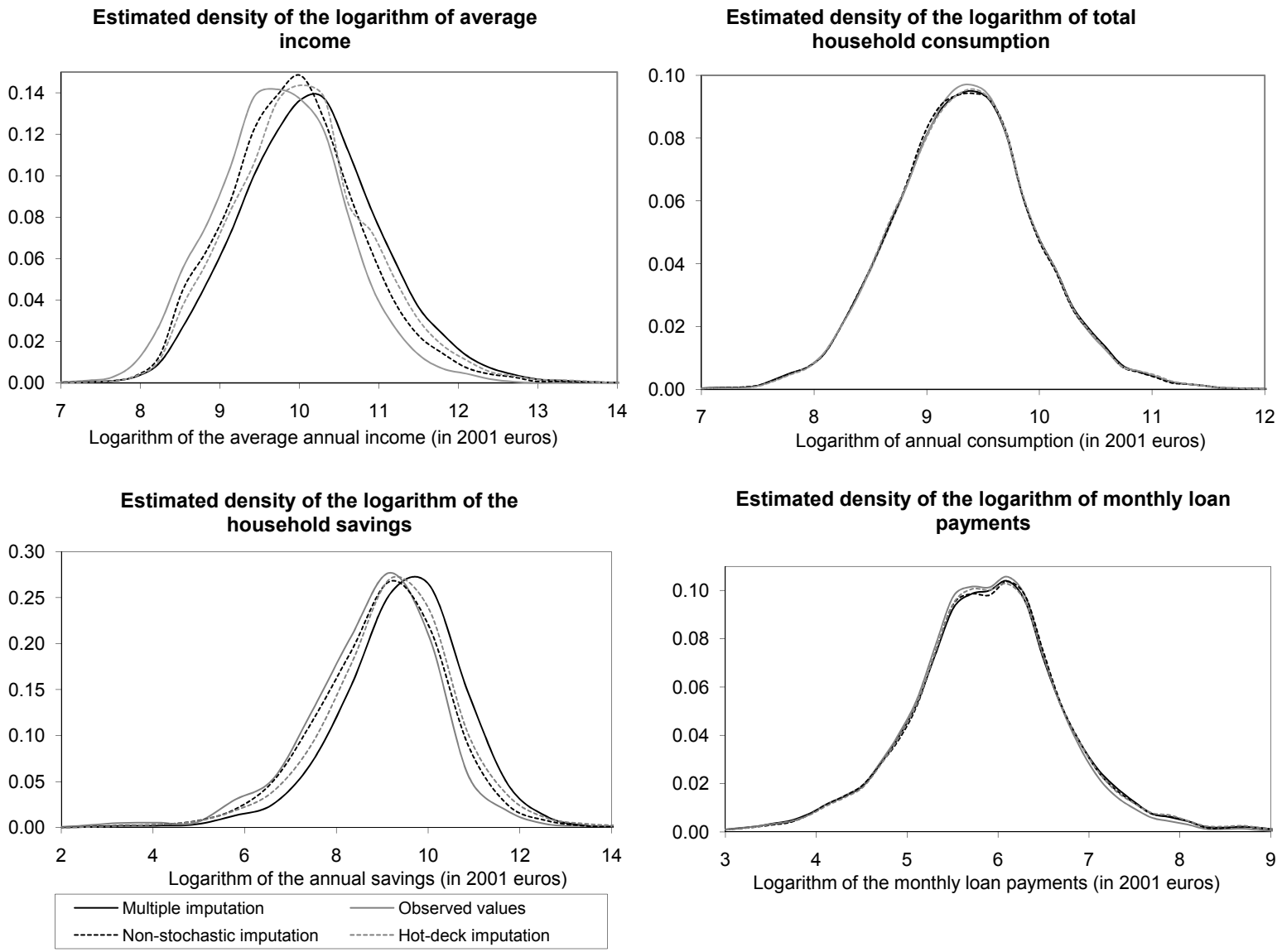

Figure 3: Kernel density estimates of average total household income, total consumption, saving and loan payments.

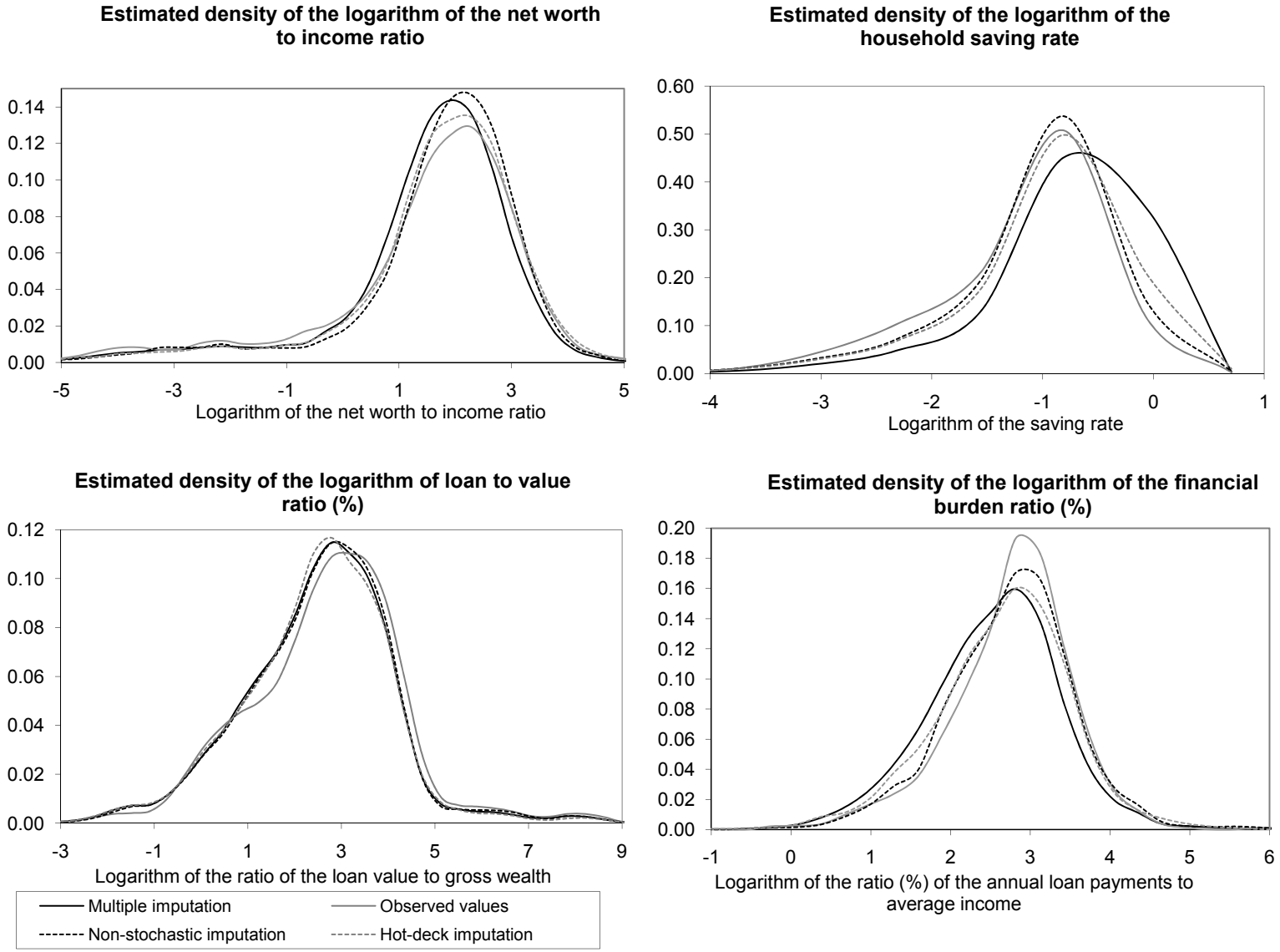

Table 1: Weighted descriptive statistics of household income and wealth variables.

		Linear-regressions		Hot-deck
	Listwise deletion	Multiple imputation	Non-stochastic imputation	Single imputation
Net wealth:				
Quantile/median ratios:				
10th percentile	0.01	0.03	0.02	0.04
25th percentile	0.36	0.42	0.42	0.42
75th percentile	1.90	1.88	1.86	1.90
90th percentile	3.42	3.36	3.32	3.42
Median	78,363	91,273	90,263	93,299
Shares:				
Top 1%	10.46	15.29	14.70	18.93
Top 5%	27.04	31.46	30.83	34.27
Top 10%	40.21	43.74	43.15	46.10
Gini index:	56.88	58.33	57.75	60.12
Total income in 2001:				
Quantile/median ratios:				
10th percentile	0.34	0.35	0.36	0.34
25th percentile	0.58	0.59	0.63	0.60
75th percentile	1.61	1.65	1.58	1.55
90th percentile	2.42	2.60	2.40	2.38
Median	17,350	22,067	18,843	20,328
Shares:				
Top 1%	6.61	6.75	6.50	7.18
Top 5%	18.48	19.56	18.29	19.46
Top 10%	29.20	30.62	28.93	29.96
Gini index:	39.92	41.63	38.95	40.11
Financial wealth:				
Quantile/median ratios:				
10th percentile	0.08	0.07	0.11	0.07
25th percentile	0.25	0.23	0.28	0.23
75th percentile	4.04	3.55	4.06	3.79
90th percentile	11.32	10.33	11.71	10.69
Median	3,477	4,169	3,465	4,452
Shares:				
Top 1%	24.16	33.26	31.77	44.10
Top 5%	50.43	55.81	55.31	63.66
Top 10%	66.06	69.34	69.14	74.79

Table 1: Weighted descriptive statistics of household income and wealth variables (Cont.).

	Listwise deletion	Linear-regressions: Multiple imputation	Linear-regressions: Non-stochastic imputation	Hot-deck: Single imputation
Average total income:				
Quantile/median ratios:				
10th percentile	0.35	0.36	0.37	0.37
25th percentile	0.62	0.60	0.62	0.61
75th percentile	1.57	1.60	1.56	1.53
90th percentile	2.30	2.52	2.38	2.41
Median	16,760	22,703	18,619	20,009
Shares:				
Top 1%	6.33	6.53	6.68	6.36
Top 5%	18.83	18.96	18.69	18.60
Top 10%	29.03	29.89	29.24	29.29
Net wealth to average income ratio:				
Quantile/median ratios:				
10th percentile	0.02	0.04	0.03	0.04
25th percentile	0.35	0.44	0.44	0.43
75th percentile	2.06	1.91	1.88	1.97
90th percentile	4.11	3.29	3.16	3.39
Median	3.99	3.91	4.71	4.56
Saving rate:				
Quantile/median ratios:				
10th percentile	−0.68	0.05	−0.11	−0.14
25th percentile	0.14	0.55	0.44	0.42
75th percentile	1.66	1.42	1.51	1.47
90th percentile	2.29	1.77	1.92	1.87
Median	0.29	0.49	0.37	0.41
Loan to value ratio (%):				
Quantile/median ratios:				
10th percentile	0.11	0.12	0.12	0.13
25th percentile	0.37	0.35	0.36	0.37
75th percentile	2.27	2.25	2.20	2.37
90th percentile	3.86	3.82	3.67	4.03
Median	20.24	17.79	18.45	16.77
Financial burden ratio (%):				
Quantile/median ratios:				
10th percentile	0.39	0.33	0.38	0.34
25th percentile	0.66	0.58	0.59	0.58
75th percentile	1.41	1.52	1.49	1.55
90th percentile	2.09	2.21	2.18	2.25
Median	18.38	14.25	17.05	16.31

Table 2: Descriptive statistics of the total household income earned in 2001 according to the main household characteristics and across methods of handling missing data.

	Listwise deletion		Imputation based on linear regression: Multiple		Imputation based on linear regression: Non-stochastic		Hot-deck imputation Single	
	Median	IQR/p25	Median	IQR/p25	Median	IQR/p25	Median	IQR/p25
Percentile of income:								
Lower than 20	5,880	0.52	7,616	0.74	6,720	0.61	6,867	0.68
Between 20 and 40	11,900	0.29	14,722	0.26	12,982	0.22	13,767	0.23
Between 40 and 60	17,350	0.21	22,039	0.20	18,843	0.20	20,300	0.21
Between 60 and 80	25,053	0.21	32,538	0.23	26,600	0.21	28,560	0.23
Between 80 and 90	34,370	0.15	47,898	0.19	37,437	0.18	40,000	0.17
Between 90 and 100	54,002	0.42	74,682	0.51	57,412	0.40	59,406	0.54
Percentile of net wealth:								
Lower than 25	13,205	1.63	16,278	1.64	13,611	1.46	15,400	1.55
Between 25 and 50	15,000	1.50	18,793	1.59	16,393	1.47	16,800	1.46
Between 50 and 75	17,948	1.43	23,680	1.54	21,000	1.36	21,600	1.36
Between 75 and 90	23,352	1.46	30,570	1.56	25,792	1.23	26,498	1.45
Between 90 and 100	38,200	1.83	48,062	1.57	39,100	1.35	37,443	1.47
Family head's age:								
Below 35	19,500	1.34	23,671	1.19	19,986	1.02	23,480	1.01
Between 35 and 44	19,206	1.31	24,378	1.48	21,240	1.15	21,840	1.24
Between 45 and 54	21,950	1.32	28,904	1.60	24,554	1.31	25,030	1.23
Between 55 and 64	21,150	1.50	25,797	1.70	22,210	1.39	23,592	1.60
Between 65 and 74	12,000	1.90	16,535	1.92	14,000	1.71	14,445	1.66
75 or over	7,700	1.22	10,388	1.66	8,239	1.38	9,240	1.63
Family head's education:								
Below secondary	13,595	1.73	17,523	1.77	15,120	1.56	15,930	1.56
Secondary	21,668	1.19	25,860	1.29	22,400	1.11	24,100	1.13
University	29,765	1.31	38,639	1.62	33,636	1.42	34,770	1.40
Family head's labour status:								
Employee	22,196	1.11	27,025	1.31	23,730	1.06	24,400	1.13
Self-employed	21,950	1.78	31,118	1.72	25,165	1.25	26,000	1.35
Retired	12,600	1.68	16,848	1.72	14,101	1.57	15,076	1.56
Inactive or unemployed	8,200	1.97	12,425	2.31	9,927	2.17	12,020	2.45

Notes from Table 2 to Table 3: IQR is the interquartile range and p25 denotes the 25th percentile.

Table 3: Descriptive statistics of the ratio (%) of loan payments to average total household income according to the main household characteristics and across methods of handling missing data.

	Listwise deletion		Imputation based on linear regression: Multiple		Imputation based on linear regression: Non-stochastic		Hot-deck imputation Single	
	Median	IQR/p25	Median	IQR/p25	Median	IQR/p25	Median	IQR/p25
Percentile of income:								
Lower than 20	22.34	3.58	26.32	1.29	28.95	1.23	29.78	1.40
Between 20 and 40	24.10	0.80	19.11	1.18	21.25	1.21	20.32	1.30
Between 40 and 60	20.64	0.85	16.40	1.23	19.95	1.12	19.19	1.65
Between 60 and 80	17.01	0.81	12.99	1.32	15.79	1.12	14.66	1.25
Between 80 and 90	17.02	1.12	10.85	1.64	13.13	1.57	13.24	1.57
Between 90 and 100	10.75	1.40	8.52	1.73	11.39	1.48	10.27	2.32
Percentile of net wealth:								
Lower than 25	19.42	0.96	16.31	1.86	18.60	1.47	17.83	1.99
Between 25 and 50	20.12	0.77	15.97	1.26	18.69	1.00	17.61	1.30
Between 50 and 75	15.57	1.23	12.85	1.55	15.56	1.43	13.62	1.46
Between 75 and 90	17.30	1.35	12.33	2.18	15.67	1.71	16.64	1.85
Between 90 and 100	12.21	1.36	12.41	1.99	13.78	2.28	14.55	2.73
Family head's age:								
Below 35	20.00	1.30	16.72	1.42	20.27	1.19	19.10	1.44
Between 35 and 44	18.79	0.95	15.60	1.35	18.49	1.29	18.08	1.50
Between 45 and 54	17.30	1.93	11.91	1.97	15.09	1.56	14.10	1.58
Between 55 and 64	16.65	1.48	12.27	2.15	14.80	1.91	14.33	2.25
Between 65 and 74	15.28	1.01	11.17	1.61	13.70	1.48	13.94	1.60
75 or over	17.47	2.06	15.16	1.17	16.72	1.12	15.85	0.93
Family head's education:								
Below secondary	19.12	0.98	15.61	1.80	18.02	1.55	17.42	1.66
Secondary	18.57	1.49	14.14	1.54	17.72	1.40	15.85	1.72
University	15.57	0.92	11.36	1.45	13.67	1.14	14.06	1.33
Family head's labour status:								
Employee	17.75	1.21	13.80	1.51	16.65	1.43	16.29	1.50
Self-employed	23.70	1.16	17.33	2.02	22.52	1.81	20.60	2.26
Retired	18.13	1.03	12.77	1.85	15.59	1.67	14.96	1.80
Inactive or unemployed	16.97	1.30	14.64	1.71	16.65	1.89	16.23	1.74

Table 4: Unweighted median regressions of the household's net worth to income ratio across alternative methods of handling missing data.

Dependent variable: net worth to average income ratio

Median (q50) : $y_j = X_j\beta_{50,j} + \varepsilon_{50,j};\ j = LD,\ MI,\ S,\ NS,\ HD$

		Imputation by linear-regression models			Hot-deck
	Listwise deletion (β_{LD})	Multiple (β_{MI})	Single (β_S)	Non-stochastic (β_{NS})	Single imputation (β_{HD})
	q50	q50	q50	q50	q50
Quintile 2	0.029	0.559	0.512	0.742	0.645
	(0.03)	(1.10)	(1.01)	(1.26)	(1.45)
Quintile 3	0.514	0.927	0.958	1.073	1.198
	(0.39)	(1.72)	(1.93)	(1.73)	(2.40)
Quintile 4	−0.072	1.346	1.169	1.706	1.505
	(−0.04)	(2.35)	(2.03)	(2.81)	(2.59)
Quintile 5	−1.735	1.262	1.025	1.567	1.440
	(−0.65)	(1.47)	(1.35)	(1.97)	(1.71)
Age 40-49	1.044	1.258	1.391	1.491	1.317
	(1.51)	(4.32)	(5.49)	(4.41)	(3.97)
Age 50-59	2.264	2.561	2.455	3.014	2.818
	(2.64)	(7.19)	(7.79)	(7.52)	(7.22)
Income/10^4	1.518	0.166	0.197	0.168	0.300
	(1.85)	(1.14)	(1.59)	(1.13)	(1.60)
Intercept	−0.596	1.319	1.199	1.788	1.359
	(−0.53)	(2.82)	(2.56)	(3.15)	(3.04)
Pseudo R^2	0.05	0.04	0.04	0.03	0.03
Test: $\beta_{\alpha,j} = \beta_{\alpha,S}$	-	-	-	0.00	0.00
Test: $\beta_{\alpha,j} = \beta_{\alpha,MI}$	0.72	-	0.47	-	-
Sample size	448	2426	2426	2426	2426

Notes: Figures in parentheses are t-ratios. The median is denoted by q50. Standard errors of multiple imputation estimates are calculated as proposed by Li *et al.* (1991). In all estimates, the standard errors are computed over 500 bootstrap replications.

Table 5: Unweighted quantile regression estimates of the household's net worth to income ratio across alternative methods of handling missing data.

Dependent variable: net worth to average income ratio

Quantile α^{th} (qα) : $y_j = X_j\beta_{\alpha,j} + \varepsilon_{\alpha,j}$; $\alpha = 25, 75$; $j = LD, MI, S, NS, HD$

			Imputation by linear-regression models						Hot-deck	
	Listwise deletion (β_{LD})		Multiple (β_{MI})		Single (β_S)		Non-stochastic (β_{NS})		Single imputation (β_{HD})	
	q25	q75	q25	q75	q25	q75	q25	q75	q25	q75
Quintile 2	−0.076	−2.229	0.470	−0.690	0.378	−0.918	0.337	−1.116	0.758	−0.069
	(−0.14)	(−1.05)	(1.53)	(−0.80)	(1.22)	(−1.27)	(0.88)	(−1.40)	(2.16)	(−0.08)
Quintile 3	0.386	−0.582	0.925	0.033	0.941	−0.002	1.123	−0.871	1.063	0.768
	(0.57)	(−0.26)	(3.39)	(0.04)	(3.73)	(0.00)	(3.42)	(−1.08)	(3.31)	(0.88)
Quintile 4	0.530	0.581	1.342	0.268	1.201	−0.029	1.829	0.926	1.441	2.239
	(0.63)	(0.19)	(4.76)	(0.27)	(5.49)	(−0.03)	(6.56)	(1.00)	(4.81)	(1.96)
Quintile 5	−0.150	−2.209	1.329	−0.499	1.028	−1.374	1.890	0.785	1.725	0.259
	(−0.09)	(−0.40)	(2.56)	(−0.33)	(2.79)	(−1.03)	(4.40)	(0.42)	(3.47)	(0.13)
Age 40-49	0.438	−0.096	0.822	1.111	0.822	1.290	0.938	1.881	0.871	1.228
	(1.18)	(−0.08)	(3.85)	(2.14)	(4.29)	(2.62)	(3.71)	(3.07)	(3.53)	(2.06)
Age 50-59	1.008	4.264	1.530	3.143	1.500	3.028	1.988	4.289	1.744	3.891
	(1.90)	(2.61)	(6.44)	(4.70)	(7.93)	(5.30)	(8.11)	(6.99)	(7.16)	(5.52)
Income/10^4	0.857	1.450	0.123	0.594	0.159	0.676	0.149	0.200	0.167	0.794
	(2.16)	(0.99)	(1.24)	(2.00)	(2.26)	(2.50)	(1.39)	(0.51)	(1.46)	(1.90)
Intercept	−1.085	4.103	−0.155	4.215	−0.201	4.148	−0.144	6.284	−0.190	3.973
	(−2.18)	(1.78)	(−1.05)	(5.99)	(−1.83)	(6.30)	(−1.01)	(8.88)	(−1.14)	(4.98)
Pseudo R^2	0.07	0.06	0.05	0.04	0.05	0.03	0.05	0.03	0.03	0.03
Test: $\beta_{\alpha,j} = \beta_{\alpha,S}$	-	-	-	-	-	-	0.00	0.00	0.00	0.00
Test: $\beta_{\alpha,j} = \beta_{\alpha,MI}$	0.16	0.42	-	-	0.81	0.76	-	-	-	-
Sample size	448	448	2426	2426	2426	2426	2426	2426	2426	2426

Notes: Figures in parentheses are t-ratios. The 25th and 75th quantiles are denoted by q25 and q75, respectively. Standard errors of multiple imputation estimates are calculated as proposed by Li *et al.* (1991). In all estimates, the standard errors are computed over 500 bootstrap replications.

Table 6: Combined estimates of the unweighted mean and median regressions of the household net worth to income ratio, saving rate, financial burden ratio and loan to value ratio.

$$\begin{aligned} \text{Mean} &: \quad y_{MI} = X_{MI}\beta_{MI} + v_{MI}, \\ \text{Median (q50)} &: \quad y_{MI} = X_{MI}\beta_{50,MI} + \varepsilon_{50,MI} \end{aligned}$$

	Net worth ratio		Saving rate		Financial burden ratio (%)		Loan to value ratio (%)	
	Mean	q50	Mean	q50	Mean	q50	Mean	q50
Quintile 2	−0.152	0.559	−0.036	−0.048	1.022	−0.094	−31.111	0.830
	(−0.24)	(1.10)	(−0.86)	(−1.01)	(0.24)	(−0.05)	(−0.26)	(0.18)
Quintile 3	0.150	0.927	0.030	0.010	−2.877	−3.127	−27.734	−1.327
	(0.23)	(1.72)	(0.91)	(0.24)	(−0.91)	(−1.94)	(−0.24)	(−0.38)
Quintile 4	0.397	1.346	0.094	0.078	−0.372	−3.790	−122.819	−4.249
	(0.54)	(2.35)	(2.83)	(2.06)	(−0.08)	(−2.19)	(−1.40)	(−1.34)
Quintile 5	−0.043	1.262	0.104	0.092	−4.505	−6.024	−115.974	−6.132
	(−0.03)	(1.47)	(2.18)	(1.68)	(−1.25)	(−3.27)	(−1.53)	(−1.60)
Age 40-49	0.982	1.258	0.023	0.007	−3.256	−3.532	57.991	−10.775
	(2.32)	(4.32)	(1.28)	(0.25)	(−2.54)	(−3.68)	(1.14)	(−4.50)
Age 50-59	2.665	2.561	0.013	0.014	−0.164	−3.719	−1.560	−16.189
	(5.02)	(7.19)	(0.58)	(0.43)	(−0.05)	(−3.14)	(−0.04)	(−7.36)
Income/10^4	0.630	0.166	0.012	0.009	−0.614	−0.010	−4.020	0.049
	(2.32)	(1.14)	(1.73)	(1.09)	(−1.08)	(−0.04)	(−0.67)	(0.12)
Intercept	2.577	1.319	0.420	0.465	23.453	19.475	145.180	28.809
	(4.80)	(2.82)	(15.59)	(15.47)	(10.40)	(13.75)	(1.68)	(8.32)
Pseudo R^2	0.06	0.04	0.05	0.03	0.02	0.04	0.01	0.02
Sample size	2426	2426	2426	2426	1257	1257	1254	1254

Notes: Figures in parentheses are t-ratios. The median is denoted by q50. Standard errors of multiple imputation estimates are calculated as proposed by Li *et al.* (1991). In all estimates, the standard errors are computed over 500 bootstrap replications.

Table 7: P-values of test statistics for the equality of the parameter estimates in mean and quantile regressions in which the income and wealth variables have been imputed by different methods.

	Unweighted regressions				Weighted regressions			
		Quantiles				Quantiles		
	Mean	$q25$	$q50$	$q75$	Mean	$q25$	$q50$	$q75$
Net worth to average income ratio:								
$H_o: \beta_{LD} = \beta_{MI}$	0.61	0.16	0.72	0.42	0.98	0.69	0.95	0.99
$H_o: \beta_S = \beta_{MI}$	0.86	0.81	0.47	0.76	0.72	1.00	0.24	0.68
$H_o: \beta_{NS} = \beta_S$	0.00	0.00	0.00	0.00	0.00	0.45	0.00	0.00
$H_o: \beta_{HD} = \beta_S$	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00
Saving rate:								
$H_o: \beta_{LD} = \beta_{MI}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$H_o: \beta_S = \beta_{MI}$	1.00	0.88	0.34	0.59	0.68	0.97	0.60	0.74
$H_o: \beta_{NS} = \beta_S$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$H_o: \beta_{HD} = \beta_S$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Financial burden ratio:								
$H_o: \beta_{LD} = \beta_{MI}$	0.33	0.00	0.02	0.28	0.21	0.00	0.02	0.25
$H_o: \beta_S = \beta_{MI}$	1.00	0.46	0.93	0.78	0.98	0.65	0.73	0.75
$H_o: \beta_{NS} = \beta_S$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$H_o: \beta_{HD} = \beta_S$	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00
Loan to value ratio:								
$H_o: \beta_{LD} = \beta_{MI}$	0.00	0.48	0.08	0.07	0.55	0.50	0.15	0.02
$H_o: \beta_S = \beta_{MI}$	1.00	0.98	0.95	0.99	0.97	1.00	0.75	0.95
$H_o: \beta_{NS} = \beta_S$	0.92	0.86	0.81	0.96	0.96	0.75	0.52	0.66
$H_o: \beta_{HD} = \beta_S$	0.98	0.60	0.95	0.89	0.98	0.89	0.58	0.69

Notes: The 25th, 50th and 75th quantiles are denoted by q25, q50 and q75, respectively. The parameter vector of the regressions (OLS or quantile regressions) are denoted indistinctly by β across alternative methods of dealing with missing data: NS = non-stochastic imputation, S = single imputation and HD = hot-deck imputation.
Standard errors of multiple imputation estimates are calculated as proposed by Li *et al.* (1991). In all estimates, the standard errors are computed over 500 bootstrap replications.

BANCO DE ESPAÑA PUBLICATIONS

WORKING PAPERS[1]

0721 CLAUDIA CANALS, XAVIER GABAIX, JOSEP M. VILARRUBIA AND DAVID WEINSTEIN: Trade patterns, trade balances and idiosyncratic shocks.
0722 MARTÍN VALLCORBA AND JAVIER DELGADO: Determinantes de la morosidad bancaria en una economía dolarizada. El caso uruguayo.
0723 ANTÓN NÁKOV AND ANDREA PESCATORI: Inflation-output gap trade-off with a dominant oil supplier.
0724 JUAN AYUSO, JUAN F. JIMENO AND ERNESTO VILLANUEVA: The effects of the introduction of tax incentives on retirement savings.
0725 DONATO MASCIANDARO, MARÍA J. NIETO AND HENRIETTE PRAST: Financial governance of banking supervision.
0726 LUIS GUTIÉRREZ DE ROZAS: Testing for competition in the Spanish banking industry: The Panzar-Rosse approach revisited.
0727 LUCÍA CUADRO SÁEZ, MARCEL FRATZSCHER AND CHRISTIAN THIMANN: The transmission of emerging market shocks to global equity markets.
0728 AGUSTÍN MARAVALL AND ANA DEL RÍO: Temporal aggregation, systematic sampling, and the Hodrick-Prescott filter.
0729 LUIS J. ÁLVAREZ: What do micro price data tell us on the validity of the New Keynesian Phillips Curve?
0730 ALFREDO MARTÍN-OLIVER AND VICENTE SALAS-FUMÁS: How do intangible assets create economic value? An application to banks.
0731 REBECA JIMÉNEZ-RODRÍGUEZ: The industrial impact of oil price shocks: Evidence from the industries of six OECD countries.
0732 PILAR CUADRADO, AITOR LACUESTA, JOSÉ MARÍA MARTÍNEZ AND EDUARDO PÉREZ: El futuro de la tasa de actividad española: un enfoque generacional.
0733 PALOMA ACEVEDO, ENRIQUE ALBEROLA AND CARMEN BROTO: Local debt expansion… vulnerability reduction? An assessment for six crises-prone countries.
0734 PEDRO ALBARRÁN, RAQUEL CARRASCO AND MAITE MARTÍNEZ-GRANADO: Inequality for wage earners and self-employed: Evidence from panel data.
0735 ANTÓN NÁKOV AND ANDREA PESCATORI: Oil and the Great Moderation.
0736 MICHIEL VAN LEUVENSTEIJN, JACOB A. BIKKER, ADRIAN VAN RIXTEL AND CHRISTOFFER KOK-SØRENSEN: A new approach to measuring competition in the loan markets of the euro area.
0737 MARIO GARCÍA-FERREIRA AND ERNESTO VILLANUEVA: Employment risk and household formation: Evidence from differences in firing costs.
0738 LAURA HOSPIDO: Modelling heterogeneity and dynamics in the volatility of individual wages.
0739 PALOMA LÓPEZ-GARCÍA, SERGIO PUENTE AND ÁNGEL LUIS GÓMEZ: Firm productivity dynamics in Spain.
0740 ALFREDO MARTÍN-OLIVER AND VICENTE SALAS-FUMÁS: The output and profit contribution of information technology and advertising investments in banks.
0741 ÓSCAR ARCE: Price determinacy under non-Ricardian fiscal strategies.
0801 ENRIQUE BENITO: Size, growth and bank dynamics.
0802 RICARDO GIMENO AND JOSÉ MANUEL MARQUÉS: Uncertainty and the price of risk in a nominal convergence process.
0803 ISABEL ARGIMÓN AND PABLO HERNÁNDEZ DE COS: Los determinantes de los saldos presupuestarios de las Comunidades Autónomas.
0804 OLYMPIA BOVER: Wealth inequality and household structure: US vs. Spain.
0805 JAVIER ANDRÉS, J. DAVID LÓPEZ-SALIDO AND EDWARD NELSON: Money and the natural rate of interest: structural estimates for the United States and the euro area.
0806 CARLOS THOMAS: Search frictions, real rigidities and inflation dynamics.
0807 MAXIMO CAMACHO AND GABRIEL PEREZ-QUIROS: Introducing the EURO-STING: Short Term INdicator of Euro Area Growth.
0808 RUBÉN SEGURA-CAYUELA AND JOSEP M. VILARRUBIA: The effect of foreign service on trade volumes and trade partners.
0809 AITOR ERCE: A structural model of sovereign debt issuance: assessing the role of financial factors.
0810 ALICIA GARCÍA-HERRERO AND JUAN M. RUIZ: Do trade and financial linkages foster business cycle synchronization in a small economy?

---

1. Previously published Working Papers are listed in the Banco de España publications catalogue.

0811 RUBÉN SEGURA-CAYUELA AND JOSEP M. VILARRUBIA: Uncertainty and entry into export markets.
0812 CARMEN BROTO AND ESTHER RUIZ: Testing for conditional heteroscedasticity in the components of inflation.
0813 JUAN J. DOLADO, MARCEL JANSEN AND JUAN F. JIMENO: On the job search in a model with heterogeneous jobs and workers.
0814 SAMUEL BENTOLILA, JUAN J. DOLADO AND JUAN F. JIMENO: Does immigration affect the Phillips curve? Some evidence for Spain.
0815 ÓSCAR J. ARCE AND J. DAVID LÓPEZ-SALIDO: Housing bubbles.
0816 GABRIEL JIMÉNEZ, VICENTE SALAS-FUMÁS AND JESÚS SAURINA: Organizational distance and use of collateral for business loans.
0817 CARMEN BROTO, JAVIER DÍAZ-CASSOU AND AITOR ERCE-DOMÍNGUEZ: Measuring and explaining the volatility of capital flows towards emerging countries.
0818 CARLOS THOMAS AND FRANCESCO ZANETTI: Labor market reform and price stability: an application to the Euro Area.
0819 DAVID G. MAYES, MARÍA J. NIETO AND LARRY D. WALL: Multiple safety net regulators and agency problems in the EU: Is Prompt Corrective Action partly the solution?
0820 CARMEN MARTÍNEZ-CARRASCAL AND ANNALISA FERRANDO: The impact of financial position on investment: an analysis for non-financial corporations in the euro area.
0821 GABRIEL JIMÉNEZ, JOSÉ A. LÓPEZ AND JESÚS SAURINA: Empirical analysis of corporate credit lines.
0822 RAMÓN MARÍA-DOLORES: Exchange rate pass-through in new Member States and candidate countries of the EU.
0823 IGNACIO HERNANDO, MARÍA J. NIETO AND LARRY D. WALL: Determinants of domestic and cross-border bank acquisitions in the European Union.
0824 JAMES COSTAIN AND ANTÓN NÁKOV: Price adjustments in a general model of state-dependent pricing.
0825 ALFREDO MARTÍN-OLIVER, VICENTE SALAS-FUMÁS AND JESÚS SAURINA: Search cost and price dispersion in vertically related markets: the case of bank loans and deposits.
0826 CARMEN BROTO: Inflation targeting in Latin America: Empirical analysis using GARCH models.
0827 RAMÓN MARÍA-DOLORES AND JESÚS VAZQUEZ: Term structure and the estimated monetary policy rule in the eurozone.
0828 MICHIEL VAN LEUVENSTEIJN, CHRISTOFFER KOK SØRENSEN, JACOB A. BIKKER AND ADRIAN VAN RIXTEL: Impact of bank competition on the interest rate pass-through in the euro area.
0829 CRISTINA BARCELÓ: The impact of alternative imputation methods on the measurement of income and wealth: Evidence from the Spanish survey of household finances.

BANCO DE **ESPAÑA**
Eurosistema

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