BANCO DE ESPAÑA
DG Economics, Statistics, and Research

12.1.2006

The Spanish Survey of Household Finances (EFF) 2002
User Guide

Unit of Microeconomic Information and Analysis

SUMMARY This document describes the files containing the data from the 2002 Spanish Survey of Household Finances (EFF). It also briefly indicates how one may proceed about using these files regarding: (i) multiple imputations that are provided to correct for item-non-response, and (ii) replicate weights that are made available to take into account sample stratification and clustering. A complete description of the 2002 wave and its methods is provided in Bover (2004).

INDEX

1 Data files

1.1 Core data

The files containing the EFF2002 data consist of the following: (i) five imputed data sets, (ii) data set with the shadow variables.

Missing data in the survey have been imputed five times using a multiple imputation procedure. The corresponding data are stored into five separate files: effe_imp1.zip, effe_imp2.zip, effe_imp3.zip, effe_imp4.zip and effe_imp5.zip[1]. All the data files are provided in Stata format.

Each effe_imp*i*.zip (*i*=1, 2, 3, 4 and 5) file contains the following:
other_ sections_imp*i*.dta: all sections of the questionnaire except section 6
section6_imp*i*.dta: section 6[2].

There is also a file, common to all five imputations, containing shadow values of the original variables (sombra.dta). The purpose of this file is to provide as much information as possible about the original state of the variables. Each variable in the survey has a shadow variable that reflects the information content of the primary variable (see below sub-section 2.4 for more details).

The household identifier variable common to all datasets is: h_number. Note that the sample unit is the household.

1.2 Replicate weights

We provide a file with replicate weights (replicate_weights.dta) to enable users taking into account sampling design features in the estimation of the sampling variances (see below some comments about the use of replicate weights). The file contains 999 replicate weights (wt3r_*i*, *i*=1,…, 999) and 999 multiplicity factors (ntimesr_*i*, *i*=1,…, 999)[3].

1.3 Main results: tables and data used

The following files are also available:

(i) File containing tables with the main results (pdf file). A first version of those tables based on preliminary imputations was published in the *Economic Bulletin*[4].
(ii) Definitions of the variables reported in the tables as Stata commands (Word file).
(iii) Data files with the above constructed variables (5 of them, one for each imputed data set).

[1] The data files with the variable labels in Spanish (available from our web site in Spanish) are called eff_imp1.zip, …, eff_imp5.zip.
[2] These are named otras_secciones_imp*i*.dta and seccion6_imp*i*.dta in the Spanish version.
[3] The multiplicity factor indicates the number of times the observation has been selected in the resampling.
[4] Spanish version published in November 04 and English version in January 05.

2 Variables

The EFF was collected using a computer assisted personal interview (CAPI). A paper version of the CAPI questionnaire is provided on the web site (both in the original Spanish wording and in English).

Some variables that appear in the paper questionnaire are not provided either for confidentiality reasons or because the question has not been properly understood by households. These are the following:

(i) place of birth variables: p1_6_*i*, p1_6a_*i,* p1_6b_*i, i*=1,…, 9

(ii) social security contribution base variables: p6_84_*i*, *i*= 1,…, 9

(iii) electronic purse cards variables: p8_8, p8_9, p8_10

2.1 Naming of the questionnaire variables in the Stata files

The questionnaire variables in the data have been named according to some common patterns that should help in identifying the corresponding question.

These patterns are the following:

(i) The variable ps_nn refers to question number nn in section s.

(ii) The variable ps_nn_m refers to question number nn in section s. Position m appears when question ps_nn is asked several times. For example, when the same question is asked to each household member.

(iii) The variable ps_nn_m_r refers to question number nn in section s. The letter m has the same meaning as before and position r appears when the question ps_nn_m is asked several times. For example, when details are asked on the characteristics of each self-employed job for each household member.

Examples:

The variable p2_5 refers to question number 5, section 2.

The variable p2_52_2_1 refers to question number 52, section 2, first loan for second property.

The variable p6_3_2 refers to question number 3, section 6, for the second household member.

The variable p6_13_4_2 refers to question number 13, section 6, second paid-employment job of the fourth household member.

The variable p6_392_3_1 refers to question number 39.2, section 6, first self-employment job of the third household member.

The variable p6_3823_3_1 refers to question number 38.2.3, section 6, first self-employment job of the third household member.

2.2 Variables from questions with multiple answers[5]

For these questions we generate variables with a pattern equivalent to the previous one but adding after the number of the question the codes cX or sX (ps_nncX_m_r o ps_nnsX_m_r).

The use of these codes is determined as follows:

(i) Variables ps_nncX_m_r correspond to questions where as many dummy variables are generated as alternative answers can be given by the respondent.

(ii) Variables ps_nnsX_m_r correspond to questions where it is assumed that each respondent will answer no more than five options. This way, the variables created for each question of this kind are at most five (ending in s1, s2, s3, s4 and s5 in the Stata file). There is no ordering of the answers when more than one option is chosen by the respondent.

Examples:

Variable p6_392c2_1_1 refers to question number 39.2, section 6, first self-employment job of the first household member and second possible answer ("first property") (question 6.39.2 of the questionnaire, first self-employment job of the first household member). This variable can take two values: 0 and 1 (indicating no and yes, respectively).

Variable p6_392c3_1_1 refers to question number 39.2, section 6, first self-employment job of the first household member and third possible answer ("second property") (question 6.39.2 of the questionnaire, first self-employment job of the first household member). This variable can take two values: 0 and 1.

Variable p2_42s1_4 refers to question number 42.4 in page 15 of the questionnaire, section 2, for the first answer of the household (question 2.42.4 of the questionnaire). This variable can take the values 1, 2, 3, 4, 5, 6, 7, 97. Note that for properties number 1, 2, and 3 p2_42 is not a multiple choice question.

Variable p2_42s2_4 refers to question number 42.4, section 2, for the second answer given by the household (if any) (question 2.42.4 of the questionnaire). This variable can take the values 1, 2, 3, 4, 5, 6, 7, 97.

[5] When multiple answers are allowed, the different possible answers are followed by a coma in the paper questionnaire.

2.3 Constructed total household income variables

Also included in the data are two contructed total household income variables, one corresponding to the whole of 2001 (renthog) and the other to the month (during 2002 or 2003) in which the interview took place (mrenthog).

These variables are calculated as the sum of labour and non-labour income of all household members. When the household fails to provide a value for one of those components we perform a direct imputation of the total. Given that the income components have also been imputed it is also possible to construct an alternative imputation of total income based on the imputed components which obviously differs from directly imputed total income[6].

2.4 Shadow variables

Following the same naming pattern, a series of additional variables have been created (shadow variables) to facilitate the identification of the values that have been imputed. The only difference in the naming of these variables is that they start with "j" instead of "p".

These variables can take the following values: 0, 1, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, and 2055. Their meanings are as follows:

1: complete observation
0: true missing, derived from the answer given by the household on a previous variable in the questionnaire.
2050: imputed value when the answer is 'Don't know'
2051: imputed value when the answer is 'No Answer'
2052: imputed value due to the lack of answer to other preceding variables.
2053: answered by the household but incorrect; value has been imputed.
2054: household answers an option not contemplated in the questionnaire due to interviewer error in question P5.18.
2055: household answers an option not contemplated in the questionnaire due to Capi error in question P5.23.
2049: edited value.
2048: value assigned because the year of birth is not provided.
2047: true missing derived from those variables with shadow values 2048.

Only those observations with shadow values equal or higher than 2050 are to be imputed.

3 Weights

We provide one set of weights (*facine3*) to compensate for (i) unequal probability of the household being selected into the sample given the oversampling of the wealthy in the EFF and geographical stratification, and (ii) differential unit non-response. In the construction of these weights account is also taken of the household composition and therefore the weight is the same for the household and for any of the household members. The sum of weights over all households in the sample is an

[6] Note however that we do not provide imputations for a few of these components (see Appendix 1).

estimate of the total number of households in the population at 2002Q4 (i.e. the weights reported are the inverse of the probability that a household is in the sample). Bover (2004) details how the sample weights are constructed.

Taking into account weights is crucial in obtaining population totals, means, and shares from the EFF data. However, there is some controversy on when weights should be used in regressions [Deaton (1997, Chapter 2) and Cameron and Trivedi (2005, Chapter 24) provide a very useful discussion on these issues]. Each user has to evaluate the situation given the objectives of the analysis at hand.

Note that when analyzing small fractions of the sample, care should be taken in applying weights which have been constructed for the whole sample.

4 Imputation

Imputations are provided for the 'No Answer' or 'Don't Know' replies for all the variables in the survey, except very few variables where the NA/DK category exceeds 60% of the answers to the question or the observations are too few (see Appendix 1 for a list of those variables).

The use of imputed values enables the analysis of the data with complete-data methods. However, the user is free to ignore the imputations we provide and obtain his own or work with explicit probability models for non-response (imputed values are identifiable through the corresponding shadow variable, as described above). For an introduction to the reasons for imputation and the choice of the imputation method used, see Bover (2004), and for a detailed description of imputation in the EFF, see Barceló (2005).

For each missing value (i.e. NA/DK answer) we provide five imputed values. These imputations are stored as five distinct datasets (five 'implicates'). One distinct advantage of using multiple imputations (MI) is to be able to assess the uncertainty associated with the imputation process [see Rubin (1987)].

To make inferences from the five multiply imputed datasets one has (1) first to analyze each of the five datasets by complete-data methods and (2) then combine the results.

Suppose the interest lies in a point estimate of some parameter Q (e.g. mean, median, regression parameter) and that for each of the five imputed datasets we have obtained an estimate of Q (using standard complete-data methods), denoted $\hat{Q}_i$. The MI point estimate of Q, $\overline{Q}$, is the average of the five complete data estimates

$$\overline{Q} = \frac{1}{5}\sum_{i=1}^{5}\hat{Q}_i$$

The variance associated with this estimate $\overline{Q}$ has two components:

(i) the within imputation sampling variance W which is the average of the five complete-data variance estimates $(\hat{V}_i)$:

$$W = \frac{1}{5}\sum_{i=1}^{5}\hat{V}_i$$

(ii) the between imputations variance which reflects the variability due to imputation uncertainty and is the variance of the complete data point estimates:

$$B = \frac{1}{4}\sum_{i=1}^{5}(\hat{Q}_i - \overline{Q})^2$$

The total variance for $\overline{Q}$ is given by:

$$T = W + (6/5)B$$

In practice, to obtain MI estimates of the type just described, the user may find useful some of the following alternatives:

(i) if only means or similar statistics are of interest, an alternative to analyzing separately the five datasets and combining the results is to construct a dataset containing the five imputed datasets successively (i.e. a unique dataset where the number of observations is five times the actual number of respondents), divide the weight variable (*facine3*) by five, and calculate the statistic.

(ii) Stata users may find helpful to download and use the procedures described in Carlin et al. (2003) for manipulating and analyzing MI datasets.

(iii) Finally, for general modelling outcomes, the user has to perform the analysis five times and combine them following the formulae above. To help see the simplicity of combining the results from the five datasets we include below few lines of Stata code that would provide the combined results (MI point estimate and its standard error) from inputting the five point estimates and five standard errors.

Usually it may suffice to do the exploratory analysis with one or two of the MI datasets and only use all of the five datasets for final results.

```
------------------------------------------------------------------------------
*   OVERALL ESTIMATES ;

use c:\input.dta;
*the file input.dta should contain five observations and two variables which are the point estimate (called here bmean) and the standard error (called here bsemean) for each of the five datasets;
gen ni=5;
set type double;
gen varmean=bsemean*bsemean;
egen w=mean(varmean);
egen qbar=mean(bmean);
gen dev=(bmean-qbar)*(bmean-qbar);
egen be=sum(dev);
replace be=be*(1/(ni-1));
gen totvar=w+(1+(1/ni))*be;
gen sqrttotvar=sqrt(totvar);
* qbar denotes the overall point estimate, totvar the overall variance (within and between component), and sqrttvar the overall standard error;
format qbar totvar sqrttotvar %12.1f;
list;
------------------------------------------------------------------------------
```

5 Sampling error: calculation of variances in each implicate

Samples designed for surveys rarely consist in simple random sampling from the population. They usually involve some (i) stratification and/or (ii) clustering. To calculate the sampling variance of estimates of interest one needs to take into account these characteristics of the sample design. Stratification may increase the precision of estimates over simple random sampling if, for example, means are different across strata. Some clustering (i.e. sampling first clusters or primary sampling units – *secciones censales* – and then choosing households from within each cluster) is usual sampling practice in order to reduce costs but it may diminish precision if household characteristics are similar within clusters. Therefore, the use of standard random sample formulas for evaluating the sampling variance may be misleading.

For simple sample designs and simple statistics appropriate variance formulas can be derived using Taylor approximations. Alternatively, bootstrap is a more computer intensive method widely used [first introduced in Efron (1979); see Horowitz (2001)]. Bootstrap samples repeatedly from the original sample with replacement. The drawing of these repeated samples is done taking into account the sample design. At each resampling the statistic of interest is evaluated and stored. The variability of these resampling statistics is used as a measure of the sample statistic.

However, taking stratification and clustering sampling features into account, either analytically or by bootstrapping, requires the availability of stratum and cluster indicators. Generally, Statistical Offices or survey agencies do not make them available for confidentiality reasons.

Alternatively, to enable more accurate variance estimates with the EFF data without disclosing stratum or cluster information we provide a file with 999 replicate weights[7]. This number of replicates is regarded as sufficient to estimate the tails of the distribution. For variance estimation a smaller number would be needed.

With a set of replicate weights the variance can be estimated from repeated estimation of the statistic of interest for each of the 999 replicate weights. This is an alternative to 999 bootstrap resampling estimates using stratum and cluster indicators (and a unique weight, *facine3*).

The replicate weights provided for this EFF wave take into account the sample design but not postratification (including non-response) or raking.

Below we include some Stata code as an example on how one could proceed to estimating the variance for the first implicate data set, i.e. V_1 in the notation of the previous section[8].

[7] These are the variables wt3r_*i*, *i*= 1, …, 999 in the replicate weights file, as described in sub-section 1.2.
[8] This should be repeated for the rest of the implicates to obtain $\bar{W}$.

```
* STANDARD ERRORS USING REPLICATE WEIGHTS (FOR THE FIRST IMPLICATE)
* HERE, FOR EXAMPLE, FOR THE MEDIAN

* A.- Statistic of interest: original sample weighted median of the variable riquezanet;
* To calculate the statistic of interest we use here the Stata procedures described in Carlin et al. (2003);
miset using c:\eff;
mido pctile medvivpr=riquezanet [pweight=facine3];
mido list medvivpr in 1/2;
mido drop if _n>1;
mici, indiv: medvivpr;
clear;

* B.- OBTAINING THE STANDARD ERROR (FOR ONE OF THE FIVE IMPLICATE DATA SETS).
* FIRST IMPLICATE;
* We first merge our data with the replicate weights file;
use c:\eff1;
sort n_cuest;
merge n_cuest using c:\replicate_weights\wdata.dta;
tab _merge;
drop _merge;
save c:\eff1wdata.dta;

* First bootstrap sample and its weighted median;
pctile medhp=riquezanet [pweight=wt3r_1];
list medhp in 1/2;
drop if _n>1;
save c:\loop1, replace;
clear;

* Reps-1 bootstrap samples and their weighted medians;
set output error;
forvalues s=2/999 {;
use c:\eff1wdata.dta;
pctile medhp=riquezanet [pweight=wt3r_`s'] ;
drop if _n>1;
append using c:\loop1;
save c:\loop1, replace;
drop _all;
        };
set output proc;
use c:\loop1;
*The sum command will provide the sampling standard error of the median in the first imputed data set, (V_1)^{1/2} ;
sum;
clear;
```

REFERENCES

BARCELÓ, C. (2005). *Imputation of the 2002 Wave of the Spanish Survey of Household Finances (EFF),* mimeo.

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

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

CARLIN, J. B., N. LI, P. GREENWOOD, and C. COFFEY (2003). 'Tools for analyzing multiple imputed datasets', *The Stata Journal,* 3, pp. 226-244.

EFRON, B. (1979). 'Bootstrap methods: another look at the jackknife', *Annals of Statistics,* 7, pp. 1-26.

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

HOROWITZ, J. L. (2001). 'The Bootstrap', in *Handbook of Econometrics, Volume 5,* edited by J. J. Heckman and E. Leamer, Elsevier Science.

RUBIN D. B. (1987). *Multiple Imputation for Nonresponse in Surveys,* Wiley.

APPENDIX 1:

VARIABLES THAT HAVE NOT BEEN IMPUTED

The variables for which no imputation is provided for NA/DK answers are the following:

1) P.4.8.1
2) P.4.8.3
3) P.4.40
4) P.6.28d
5) P.6.28f
6) P.6.51b
7) P.6.57b
8) P.6.59b
9) P.6.60b
10) P.7.4b
11) P.7.8b
12) P.7.10

For these variables, observations whose values should have been imputed but imputation was judged not reliable are marked with a -9999 value.