5 Multiple Imputations

5.1 Introduction

A common problem with voluntary surveys is item nonresponse, i.e. the fact that some survey participants do not answer all questions.[1] This is especially the case with surveys that pose complicated or sensitive questions (e.g. about income or wealth).

If the problem of missing information due to item nonresponse is disregarded, this leads to biased estimates. For the HFCS data, we therefore used a method that addresses this problem, i.e. multiple imputation by chained equations.

The idea behind this approach is to substitute missing values in the dataset with several values that have been estimated based on an iterative Bayesian model. The main aim of this procedure is to impute in such a way that the associations between all variables are preserved or, in other words, to maintain the correlation structure of the dataset. Under this approach, the missing values of each variable are estimated by taking into account a maximum number of available variables. In order to capture the uncertainty behind the missing values, we do not impute just one value per missing value, but several (in the case of the HFCS, five).

Other similar surveys – such as the *Survey of Consumer Finances* (SCF – see Kennickell, 1998) and the *Spanish Survey of Household Finances* (EFF – see Barceló, 2006) – also use the same approach to impute missing data.

As multiple imputation is a very time-consuming process, most institutions that carry out surveys, including the HFCS, provide users with already imputed datasets. This ensures that all users can work with the same imputed datasets. In the case of the HFCS, users can identify every imputed value of any variable by looking at the corresponding flag variable (section 4.5). Thus, they have the opportunity to carry out nonresponse analyses or imputations on their own, or to use other methods for dealing with item-nonresponse in their analyses.

This chapter is structured as follows: In section 5.2, we present data on item nonresponse in the HFCS. Section 5.3 then describes the imputation procedure used, and in section 5.4 we explain the specification of the imputation model and how the imputations were executed. Section 5.5 finally presents some imputation results.

5.2 Item Nonresponse

Table 4 shows selected statistics on item nonresponse. On average, each household has 17.3 missing values, which means that item nonresponse was limited to a mere 2% of all the questions (variables) addressed to each household. However, this rate increases to 6.9% in the case of euro variables, which suggests that questions of this kind might be perceived as sensitive or difficult to answer.

There are different ways of analyzing datasets that include variables with missing values.[2] In most statistical packages, the default method is the complete-case analysis method. This method implies that all households that have missing values in any of the variables of interest are deleted and that analyses are based on complete observations only. However, the loss of information resulting from this

[1] *Another related problem that occurs in surveys is unit nonresponse, which means that no questions are answered at all because, for example, participation in the survey is declined by a household. We address this problem through the construction of HFCS nonresponse weights (chapter 7).*

[2] *For a comprehensive study, see Little and Rubin (2002).*

Table 4

Item Nonresponse per Household (Unweighted)

	Mean	Median	Minimum	Maximum
Number of variables asked				
all variables	826.8	824.0	637	1,242
euro variables	52.1	53.0	17	98
Number of variables with missing values				
all variables	17.3	8.0	0	474
euro variables	3.6	2.0	0	54
Share of variables with missing values in %				
all variables	2.0	1.0	0	39.5
euro variables	6.9	4.2	0	78.8

Source: HFCS Austria 2010, OeNB.

Note: Interval responses are considered as missing values with regard to the corresponding euro variable and are not included as a separate variable. A question addressed to several household members is entered as several variables, one for each household member.

method leads to two problems: First, it biases estimates if complete observations differ systematically from incomplete ones; second, even if an estimate is unbiased, its estimation would be less precise due to the observations lost. To illustrate how significant the loss of information would be in the case of the HFCS, table 5 shows item nonresponse rates across some selected variables.

Table 5 for example shows that – when asked about the value of their main residence – 75.5% of households provided a specific amount (third column). The other 24.5% of households are item nonrespondents, which means that either they

Table 5

Item Nonresponse for Selected Variables (Unweighted)

	Household has item		Responses by households that have the item			
	Yes	Un-known	Amount	Interval	"Don't know"/ "No answer"	Other missing values[1]
	(1)	(2)	(3)	(4)	(5)	(6)
	%					
Value of main residence[2]	49.6	0.0	75.5	15.3	6.4	2.7
HMR mortgage 1: amount still owed	15.1	1.4	63.5	21.2	12.3	3.1
Monthly amount paid as rent	44.1	0.0	97.1	2.3	0.5	0.1
Other property 1: current value	12.6	0.2	74.1	15.3	9.6	1.0
Other property mortgage 1: amount still owed	1.7	0.4	70.7	7.3	14.6	7.3
Value of sight accounts	98.9	0.0	72.0	13.3	14.4	0.3
Value of saving accounts	86.0	1.6	64.6	18.6	16.0	0.8
Value of publicly traded shares	5.4	0.4	71.1	12.5	16.4	0.0
Amount owed to household	9.3	0.5	90.5	5.0	4.5	0.0
Labor status (main employment) (person 1)	100.0	0.0	99.9	0.0	0.1	0.0
Gross cash employee income (person 1)	48.8	0.1	76.7	9.9	3.4	9.9
Gross income from unemployment benefits (person 1)	6.1	0.1	83.3	9.7	6.3	0.7
Gross income from financial investments	70.9	6.6	34.3	40.7	24.0	0.9
Gift/inheritance 1: value	21.4	1.3	71.1	16.3	10.0	2.6
Amount spent on food at home	100.0	0.0	96.3	3.4	0.3	0.0

Source: HFCS Austria 2010, OeNB.

[1] *Missing values due to editing measures and exits from loops.*
[2] *Based on the HB0900 variable.*

Note: HMR = household main residence.

provided a (prespecified or individual) interval (15.3%, fourth column), responded with "Don't know" or "No answer" (6.4%, fifth column) or that their response was edited to a missing value[3] (2.7%, sixth column). Nonresponse rates[4] vary substantially across items. Variables with high nonresponse rates include e. g. questions related to outstanding balances on mortgages that use the main residence as collateral (100% – 63.5% = 36.5%) and the household's gross income from financial investments (100% – 34.3% = 65.7%). With regard to the latter, 40.7% of households provided at least an interval for this type of income, which confirms the importance of asking interval questions if there is a nonresponse to a euro question. Interval questions provide valuable and often very precise information (see the online appendix and section 2.6.2 for information on the design of euro loops). Variables with low nonresponse rates include non-euro variables, such as labor status (100% – 99.9% = 0.1%), or the amount spent on food at home (100% – 96.3% = 3.7%).

Table 5 (second column) also shows another aspect of item nonresponse in the HFCS: There are variables, known as *branch variables* (see chart 3 in chapter 4), which may also have missing values due to nonresponses to a previous question (*head variable*) and which are thus classified as missing. For example, before the euro question on gross income from financial investments is asked, households are asked a yes/no question determining whether they have this type of income or not. Only those that answer in the affirmative (70.9%) are then asked the euro question; the other households, including the 6.6% of households that did not answer the yes/no question, automatically skip the euro question. As it is unknown, however, whether the 6.6% of households that did not answer the yes/no question have a positive gross income from financial investments or not, their nonresponses must also be taken into account as second-order (or higher-order) missing values when analyzing nonresponse to the euro question.

Thus, if a complete-case analysis were to be carried out with the HFCS data, the loss of information and the resulting loss in precision of unbiased estimates could be considerable owing to the large amount of variables with higher-order missing values. Furthermore, as complete observations usually differ systematically from incomplete ones, complete-case analysis would bias the estimates.

Table 6

Logit Regression of Nonresponse in the Euro Question on Value of Sight Accounts (Unweighted)

Covariates	Coefficients
Female (person 1)	0.0775
	(0.0950)
Age (person 1)	–0.0012
	(0.00344)
Tertiary education level (person 1)	–0.259*
	(0.156)
Employed/self-employed (person 1)	–0.195*
	(0.113)
Residence is in Vienna	–0.194
	(0.134)
Size of main residence	0.00274***
	(0.000863)
Household size	0.119***
	(0.0421)
Constant	–1.331***
	(0.256)
Observations[1]	2,330

Source: HFCS Austria 2010, OeNB.

*Note: Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1.*

[1] *The remaining 50 observations of the dataset show missing values in one of the covariates and/or filter „missing" observations in the dependent variable and are thus not included in the regression.*

[3] *See chapter 4 for more details.*

[4] *The nonresponse rate is calculated as 100% minus the value in the "amount" column in table 5.*

For illustration purposes, table 6 shows a regression of nonresponse to the question regarding the value of sight accounts (1 if the value is missing, 0 otherwise) on several explanatory variables. We can see that item respondents differ significantly from item nonrespondents, because respondents live in smaller main residences and smaller households, have higher levels of education and are more often gainfully employed. Thus, a complete-case analysis of the value of sight accounts would bias the estimates toward a population with these household characteristics.

5.3 HFCS Imputation Procedure

To impute HFCS data, we have chosen a procedure implemented in the statistics software Stata® by Royston (2004) in which all variables to be imputed are estimated in regression equations (chained equations[5]). It can be summarized in the following steps:[6]

- Step 1: Select the P variables $Y_1, Y_2, ..., Y_P$ that are going to be imputed.
- Step 2: Fill the missing values of $Y_1, Y_2, ..., Y_P$ with random draws from actually observed values.
- Step 3: For each $Y_1, Y_2, ..., Y_P$
  - run a Bayesian regression of the variable to be imputed on a broad set of right-hand side variables that is chosen from among the HFCS variables without missing values and the variables selected in step 1 (except the one being regressed); restrict the regression sample to those observations that are not missing in the dependent variable;
  - randomly draw a vector of regression parameters from their posterior distribution;
  - calculate the corresponding predicted values and use them as the imputed values;
  - replace the missing values of the imputed variable with its imputed values.
- Step 4: Repeat step 3 t times. Each time, replace previous imputed values with updated ones obtained from the latest regression. This creates the first imputation sample (or implicate).
- Step 5: Repeat steps 3 and 4 M times independently to obtain M imputation samples.

The basic idea behind this procedure is to impute missing values for each of the P variables with missing values by drawing predictions based on a Bayesian regression model specific to that variable (step 3). To preserve the associations between variables with missing values and variables with complete observations, each regression model contains a broad set of right-hand side variables with *complete* observations.

Furthermore, the procedure is *multivariate* in the sense that the estimation of the missing values is repeated (t times); variables that are being conditioned in each regression are replaced by the observed or currently imputed values (step 4). It is important that each regression model also contains a broad set of right-hand

[5] *This procedure is also known by several other names, including "stochastic relaxation," "regression switching," "sequential regression," "incompatible MCMC" and "fully conditional specification."*

[6] *Albacete (2012) provides further technical details on the imputation procedure used for the Austrian Household Survey on Housing Wealth, which is identical to that used for the HFCS.*

side variables with *missing* values in order to preserve the joint distribution of variables with missing values. If t tends to infinity, the imputations of the missing values of $Y_1, Y_2, ..., Y_P$ in each cycle are expected to converge to an approximation of a draw from their joint posterior predictive distribution.

In the final step (step 5), the procedure provides *multiple* imputations of each missing value by repeating steps 3 and 4 M times independently. This is done to take into account the uncertainty behind the imputed values when estimating any variances with imputed variables with missing values. The M imputations of the missing values of $Y_1, Y_2, ..., Y_P$ are expected to converge to an approximation of M draws from the joint posterior predictive distribution of the missing values.

Although it is theoretically possible that the sequence of draws based on the above regressions might not converge to a stationary predictive distribution, simulation studies provide evidence that the approach yields estimates that are unbiased (Van Buuren et al., 2006). Furthermore, with data such as those from the HFCS – where there is a large number of variables, many of which have bounds, skip patterns, bracketed responses, interactions or constraints in relation to other variables – using separate regressions for each variable reflects the data better. This approach thus makes more sense than specifying a joint distribution for all variables together, as is the case under the joint modeling approach.[7]

It should be pointed out that the HFCS imputation procedure is based on the assumption that the nonresponse probabilities of variables with missing values are only dependent on observed information – never on unobserved information such as the variables with missing values themselves. This assumption is referred to as *ignorability assumption* in the literature.

Before running the above five steps, we need to prepare the data and specify all the parameters of our imputation model: e.g. the choice of variables to be imputed, the imputation order, the regression model for each variable, the number of cycles t, the number of imputation samples M, etc. The next section describes how this was done.

5.4 Creating the Imputations

5.4.1 Choosing the Variables to Be Imputed

In step 1 of the HFCS imputation procedure, we have to select the variables $Y_1, Y_2, ..., Y_P$ that are going to be imputed. Our strategy is to impute as many variables with missing values as possible, which in our case amounts to around 70% of the variables with missing values. The remaining variables with missing values are not imputed with the HFCS imputation procedure because either they do not have enough variance or there are technically not enough observations to be estimated in a regression.[8]

The imputation of as many variables as possible is supposed to minimize the number of cases in which users are forced to conduct a complete-case analysis with HFCS data because the variables they are interested in have not been imputed (see

[7] *See also Little and Rubin (2002) for an overview of imputation techniques.*

[8] *A very small fraction (around 2%) of these variables that could not be imputed with the HFCS imputation procedure were imputed with ad hoc methods such as hotdeck after the HFCS procedure had been completed. This is because their imputation is considered very important as they are used, for example, to build important aggregate variables, such as total household income.*

the description of the disadvantages of complete-case analysis in the introduction). Another important reason for adopting this strategy is that we do not want to bias the correlation structure of the data with our imputations. If we were to reject many variables for imputation, we could not subsequently use them in the regression models as right-hand side variables with missing values either, and we would thus bias the associations between the unimputed variables with missing values and the imputed ones.

5.4.2 Imputation Order

As mentioned in section 5.3, a weakness of the procedure is that it does not enable us to prove, in theoretical terms, that the sequence of drawn predictions based on the Bayesian regressions converges to a stationary predictive distribution. In practice, however, it has been found that choosing a particular order of $Y_1,Y_2,...,Y_P$ often aids convergence. Therefore, we order the variables to be imputed by their degree of missingness, starting imputation with the variables with the least missing values and ending it with those variables that have the most missing values. Variables with the same degree of missingness are processed in a random order, but always in the same order. Head variables are always imputed before their corresponding branch variables. For example, the variable indicating whether a household has a mortgage or not is always imputed before imputing the mortgage amount, even if the degree of missingness is the same for both variables.

5.4.3 Types of Regression Models

In step 3, we defined a regression model for each variable to be imputed. Depending on the type of the variable, we choose from among four different types of regression models. For quantitative variables, we use an interval regression model[9] because all of our continuous variables are bounded either from above or from below, or from both above and below (section 5.4.6). For binary variables, we use a logit model; and for ordinal and nominal variables, we use ordered logit and multinomial logit models.

5.4.4 Use of Weights in Regressions

Generally speaking, there is little debate about the need to use weights for the estimation of descriptive parameters (means, proportions, totals, etc.). There is, however, some debate about the use of weights when fitting regression models to survey data. This issue also arises when fitting the regressions in step 3 of the HFCS imputation procedure. We have decided to use weights for the following reasons. If the regression models are misspecified, weighted regression is better than unweighted regression because it provides unbiased estimates of the regression coefficients. Weighted regression should only be forgone if regression models are correctly specified, because then it would result in inflated standard errors. As we do not know the data-generating process, we prefer to be on the safe side and use weights to arrive (with certainty) at unbiased imputations at (maybe) the cost of slightly inflated standard errors. This is especially important, as the dataset will

[9] *The interval regression model is a generalized version of the Tobit model. It is used to account for censoring from below and/or above. See Cameron and Trivedi (2005) for more details.*

become publicly available and we anticipate that most users will work with the imputed datasets.[10]

5.4.5 Variable Transformations

Before imputing variables with missing values, we transform several of them, because this has proved to be extremely helpful in improving the imputed values of these variables and, hence, in improving the quality of the imputed values in general. Once the imputations are finished, we transform all variables back into their original measure.

One important transformation of quantitative variables involves using the logarithm. Usually, these types of variables have a highly skewed distribution; using the logarithm helps to make the distribution look closer to the normal distribution assumption which is necessary for the forecast. Another very helpful transformation for year variables is to impute durations instead of years. For example, instead of imputing the purchase year of a house, we impute the time elapsed since the house was purchased. In such cases, the above-mentioned logarithmic transformation is carried out on the durations and not on the years.

For categorical variables, two types of transformations may be used. First, some of the nominal variables are transformed into ordinal variables by reordering categories. This improves the stability of the imputation model, as fewer parameters need to be estimated for ordinal regression models than for multinomial regression models. Second, multiple response variables that are usually nominal are transformed into several binary variables by creating one binary variable for each response category (1 if the category applies, 0 otherwise). This makes it possible to impute more than one response category for the same question per imputation sample.

A transformation that is done for both quantitative variables with missing values and categorical variables with missing values involves splitting the original variable into head and branch variables; this is done when there is a certain heterogeneity in the original variable. For example, some loan-length variables have the value –4 indicating *loan has no set term*. When imputing such a loan-length variable it would not make sense to run the regression over these observations together with those that do provide a loan-length value. In such cases, the variables are split into two: (1) a binary head variable indicating whether the loan has a set term or not (imputed with a logit regression model); and (2) a quantitative branch variable indicating the loan length if the loan has a set term (imputed with interval regression).

A further transformation, which is carried out both for quantitative and categorical variables with missing values, is that of person IDs.[11] Person variables are

[10] *Another possibility for deciding whether to use weights or not is to impute with and without weights and then compare both models. If we find significant differences between the parameter estimates of both approaches, the results suggest the use of weights to at least obtain unbiased estimates. If no significant changes between the parameter estimates are found, but instead large changes in standard errors, then this suggests an appropriate specification and minimal problems with using unweighted models. However, running the HFCS imputation procedure twice, once weighted and once unweighted, is a process which would be very time-consuming and has not been done yet. This is left for future research.*

[11] *In the dataset, financially knowledgeable persons are designated with the ID = 1 by default; all other persons are ordered according to their age.*

modeled and imputed separately for each person ID in order to avoid biased imputations (section 5.4.8); this should ensure that persons with the same IDs show relatively homogenous characteristics if they are jointly modeled. For this reason, respondents are grouped into new person ID categories, created specifically for the imputations, prior to imputation. The criteria for this categorization are the following: All male financially knowledgeable persons (FKPs), all male partners of FKPs that were second person and all other FKPs are classified as first persons (person ID = 1). All female partners of FKPs that previously already were second persons and all women that were first persons before their male partners became first persons are classified as second persons (person ID = 2). All other persons are ordered and numbered according to their age (from old to young).

In the case of households with members that engage in farming, we use a special transformation of the variables regarding the value of the household's business(es) (HD0801–HD0803) and the variable regarding the value of the household main residence (HB0900). Instead of imputing these variables individually, we first impute the sum of these variables and, additionally, the percentage of this sum that is attributable to the farm. In a second step, we calculate the individual variables (HD0801–HD0803 and HB0900) based on this sum and these percentages. The motivation behind this transformation is that it considerably improves the imputed values, as some households with members that engage in farming did not state the value of their main residence separately from the value of their farm but indicated the sum of both (see section 4.6.2.8 for further details).

5.4.6 Bounds

As mentioned above, we use interval regression models to impute quantitative variables in step 3 because all these variables are bounded either from above or from below, or from both above and below. These bounds are used to avoid the imputation of values that are not defined or that are inconsistent with other variables in the survey. We distinguish between *general* bounds and *individual* bounds.

General bounds are the same for all households and persons, and are used to avoid imputing values that are not defined or very unrealistic. Examples of this type of bound include nonnegativity constraints on quantitative or count variables (e.g. income or age). The lower bound for these variables is zero for all households. Furthermore, for each quantitative variable, we use the following rule: for every household, the lower bound is equivalent to half of the smallest value observed for the variable, whereas to calculate the upper bound, we take the largest observed value and double it. This helps to avoid the imputation of extreme outliers without biasing results. More examples of general bounds include share variables (e.g. share of homeownership), where we set the lower bound to zero and the upper bound to 100, or some year variables (e.g. the purchase or inheritance year of the household main residence), where the upper bound is 2011, i.e. the year in which the latest survey interviews were carried out.

Unlike general bounds, individual bounds take different values depending on each household or person; they usually ensure consistency with other variables of the same household. Most of the HFCS bounds fall into this category. For example, when imputing the amount spent on food eaten at home, we set the total consumption expenditure estimated by the household as the upper bound. Or looked at from the opposite angle, when imputing the total estimated consump-

tion expenditure, we set the sum of the amounts spent on food and drink consumed at home and outside of the home as the lower bound. Individual bounds are also used when a household provides an interval (either prespecified or individual) in a euro question instead of a specific value. Such intervals are asked for after every euro question that is left unanswered; they prove very useful for imputation purposes as they yield valuable and precise information on the missing value in the euro question (see also section 5.2 in connection with table 5).

Individual bounds in the HFCS are, for example, also used when imputing rents (e.g. rent including utilities is used as an upper bound for net excluding utilities and vice versa), loans (e.g. the initial amount of a loan is used as an upper bound for the outstanding amount of the loan and vice versa), or when imputing several count variables (e.g. the birth year of the oldest household member is used as a lower bound for the year of acquisition of the main residence). If an observation has more than one lower and/or upper bound (e.g. general and individual bounds), we take the lower and/or upper bound that is the most restrictive.

5.4.7 Selecting Predictors

As mentioned above, one of the main goals of imputation is to preserve the association between variables with missing values and variables with complete observations – and also that among variables with missing values themselves. Therefore, when choosing predictors for the imputation model, it is not sufficient to select the most accurate predictors for each variable to be imputed. Such an approach could bias the correlation structure between the variable to be imputed and the excluded variables. Furthermore, ignoring variables that are determinants of nonresponse with respect to the variable to be imputed makes the ignorability assumption on which our imputation model relies (section 5.3) less plausible.

Thus, we choose as many predictors as possible (broad conditioning approach). In a large dataset such as that of the HFCS, with several hundred variables, it is, however, not feasible to include all of them, as this may lead to both multicollinearity problems and computational problems. In line with Van Buuren et al. (1999) and Barceló (2006), we have adopted the following strategy for selecting predictor variables:

1. Include the variables that are determinants of nonresponse. These are necessary to satisfy the ignorability assumption on which our imputation model relies (section 5.3). Variables included as typical determinants of nonresponse in the HFCS imputation model are, for instance, variables that describe the household (e.g. estimated household income, household size, number of children), variables that describe household members (e.g. age, education, sex and labor status of the household's first person and the latter's partner), stratification variables (e.g. province, municipality size), information provided by the interviewers (e.g. standard of living, type of neighborhood, type of building, interview atmosphere, etc.). The latter pieces of information (paradata) were extremely important for the imputations since they provided plausible explanations for item nonresponse for many variables.
2. In addition, include variables that are very good at predicting and explaining the relevant variable to be imputed. This is the classic criterion for predictors and it helps us to reduce some of the uncertainty surrounding the imputations. These predictors are identified by their correlation with the variable to be

imputed. For example, when imputing credit variables, we typically use the original loan amount (as mentioned above), the repaid loan amount or principal outstanding as predictors because, in most regressions, these variables can explain a considerable amount of variance. Usually, these variables are logically connected (e.g. outstanding principal is the original loan amount minus the sum of all repaid loan amounts). However, in the course of imputation, it is not possible to preserve all of these logical connections, in particular if several of these variables are being imputed.

3. Remove the aforementioned predictor variables that have too many missing values within the subsample of missing observations of the variable to be imputed and substitute them with more complete predictors of these predictors. As a rule of thumb, predictors with percentages of observed cases within this subsample below 50% are removed and substituted by more complete predictors. This criterion helps to make the imputations more robust. Typical predictors of predictors include essential household characteristics, such as household size, the number of children, region and age, as well as the labor and marital status of the first person.
4. Include all variables that appear in the models that will be fitted to the data after imputation. In other words, think about different economic theories that might be tested with the data and include the variables as predictors that are expected, according to these theories, to affect or explain the variable to be imputed. Failure to do so will tend to bias the results of potential users of the data when testing the hypothesis of one particular model. For example, the HFCS data provide detailed information on different components of households' wealth, e.g. real assets or financial assets. This information is used for the analysis of wealth effects on consumption. Therefore, we use these variables both for the imputation of consumption expenditure and for the imputation of asset variables.

Obviously, many variables in the survey – for example, income, age or education of the first person – fulfill more than one criterion for predictor selection.

In all regression models we also include an interaction term and a main effect dummy for each one of the above predictor variables that was not inquired about in the case of every household that was asked about the variable to be imputed. For example, suppose that we want to impute a household's consumption expenditure using mortgage amount as one of our predictors. While every household in the sample was asked about consumption expenditure amounts, not all of them were asked about mortgage amounts. If, for those households that do not have a mortgage, we just set the mortgage amount to zero (corresponds to an interaction term), the estimates would be biased, because the information on whether a household has a mortgage or not would be omitted. This information should thus be additionally included as a main effect dummy in the regression model. But again, not all households were asked whether they have a mortgage, just homeowners. Thus, we should also include a homeowner dummy in the regression.

Finally, the number of predictors is restricted by the size of the subsample over which the regression is estimated. In cases where the subsample size is smaller than the number of predictors selected according to the above strategy, we use the Akaike information criterion to choose the subset of predictors which best fits the data, ensuring that each one of the above four predictor categories is represented

in each regression equation (to the extent possible). Typically, the number of predictors used for each regression model is around 20% of the number of observations for the variable to be imputed for small subsamples. For large subsamples, the number of predictors usually lies between 5% and 10%. For more details on the specification of subsamples, see the next section.

5.4.8 Specification of Subsamples

Each regression in step 3 is estimated over a subsample which consists of all households and persons that were asked the question pertinent to the variable to be imputed. For example, if a household has two mortgages and we want to impute the outstanding amount of the second mortgage, then we impute this missing value by regressing over the subsample of households that have at least two mortgages. If we also included the households that only have one mortgage when imputing the second mortgage amounts, we would ignore systematic differences between the first and second mortgages. For example, we would ignore the fact that the outstanding amount of the first mortgage is always higher than the second one, because households order mortgages by importance, which would introduce a bias in our estimates.[12]

A further example is the imputation of person variables. These are also only regressed over the subsample of persons that share the same person ID. To ensure the homogeneity of the persons with the same IDs, respondents are grouped into new person ID categories, which are specifically created for the imputation, prior to imputation (section 5.4.5), and which then form the mentioned subsamples. When imputing question by question, as we do, the bias will be very limited, though at the cost of precision because, consequently, the subsample sizes are often small.

5.4.9 Number of Cycles

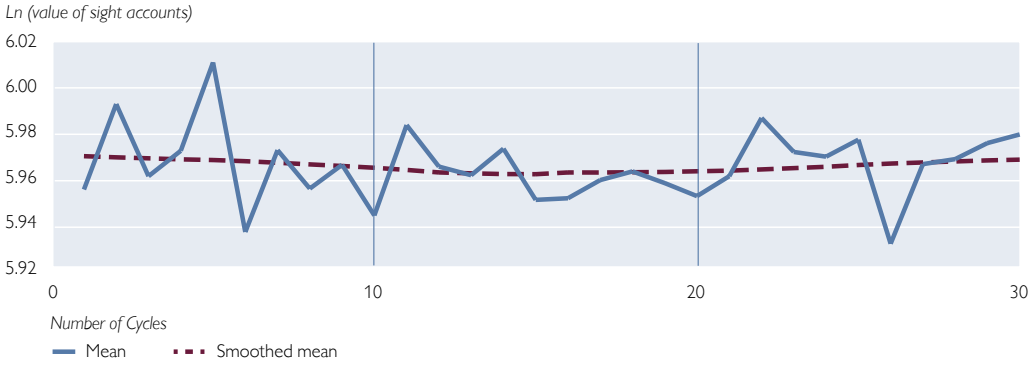
In step 4, the number of cycles (or iterations) t determines how often step 3 is repeated. As t tends to infinity, the imputed values should converge to a draw from the joint posterior predictive distribution of the variables with missing values. However, according to Van Buuren et al. (1999), in practice, convergence in these models usually occurs very quickly during the first few iterations. Given the large computational effort required for the HFCS imputation model, and following the number of iterations used in other similar surveys, e.g. the SCF (Kennickell, 1998) and the EFF (Barceló, 2006), we set the iteration number for the HFCS imputation model at $t = 6$.

Typically, we check convergence graphically by plotting the mean of the imputed values against the iteration number t; chart 4, for example, shows this for the value of sight accounts (variable HD1110). Convergence is judged to have occurred as soon as the pattern of the imputed means becomes random. In chart 4, this seems to be the case very early on: from the first iteration on, it is no longer possible to recognize any clear trend in the smoothed curve of the imputed means of the value of sight accounts. Furthermore, chart 4 shows that the fluctuation range of the imputed means is very small, which is a further indicator of conver-

[12] *Of course, in such cases, we could introduce a large number of interaction terms into our model to reduce the bias, but there still might be unobserved differences between the two groups.*

Chart 4

Mean Value of Sight Accounts across Iterations (Unweighted)



Source: HFCS Austria 2010, OeNB.

gence. Of course, checks of this kind can never confirm convergence (nor can any other check under the chained equations approach), but they can highlight weaknesses in the imputation model or other unusual outcomes that could be indicators of nonconvergence.

5.4.10 Number of Imputation Samples

In a last step (step 5), we choose the number of realizations $m = 1,2,...,M$ that we want to have from the joint posterior predictive distribution of the missing data or, put more simply, the number of samples to be generated through multiple imputation. Setting M too low leads to standard errors of the estimates that are too low and to p-values that are too low. However, Schafer and Olsen (1998) show that the gains in efficiency of an estimate rapidly diminish after the first few M imputation samples. They claim that good inferences can already be made with $M = 3$ to 5. In line with the international standards set by the ECB and other similar surveys (like the SCF or EFF), we set the number of imputations at $M = 5$.

5.5 Selected Results

After imputation, the HFCS dataset is five times bigger than before, because it consists of $M = 5$ multiple imputation samples (also referred to as "implicates"). Table 7 provides first insights into the imputation output. It shows the weighted means of selected euro variables in both the multiple imputation samples and the original unimputed sample.

One interesting result is that the means of most variables are, on average, higher after imputation than before imputation. If imputations are close to the true values, the result suggests that the households that do not respond to the relevant variables tend to be households with higher (unobserved) amounts in these variables. For example, the mean value of the first gift/inheritance (without main residence) is EUR 88,019 before imputations. After imputations, it increases to EUR 110,526 in $m=2$, EUR 94,873 in $m=3$, EUR 190,532 in $m=4$, and EUR 125,350 in $m=5$. In $m=1$ the mean decreases slightly to EUR 87,819. Thus, on average the imputations increase the mean value of the first gift/inheritance from EUR 88,019 to EUR 121,820, i.e. by 38% (about half of the values imputed in this context are based on interval responses by households). This suggests that house-

Table 7

Means for Selected Variables before and after Multiple Imputation (Weighted)

	Mean before imputation	Multiple imputation sample means				
	m = 0	*m = 1*	*m = 2*	*m = 3*	*m = 4*	*m = 5*
	EUR					
Value of main residence[1]	246,203	261,468	271,337	266,286	275,096	268,015
HMR mortgage 1: amount still owed	55,745	94,427	84,670	47,135	58,619	47,657
Monthly amount paid as rent	363	334	332	335	330	334
Other property 1: current value	231,583	195,953	193,173	265,195	198,880	218,387
Other property mortgage 1: amount still owed	68,300	58,141	90,563	70,627	74,631	67,420
Value of sight accounts	2,406	3,343	3,255	3,130	2,908	3,220
Value of saving accounts	21,989	28,230	29,700	33,696	29,781	28,905
Value of publicly traded shares	30,440	23,554	36,887	22,753	28,553	22,573
Gross cash employee income (person 1)	25,871	25,075	25,254	26,517	26,230	29,403
Gross income from unemployment benefits (person 1)	6,263	6,361	6,880	6,300	6,225	6,295
Gross income from financial investments	836	800	763	730	771	787
Gift/inheritance 1: value	88,019	82,842	130,673	95,338	90,959	94,404
Amount spent on food at home	379	381	380	381	380	380

Source: HFCS Austria 2010, OeNB.

[1] Based on the HB0900 variable.

Note: All means are estimated over the observations "Household has item = yes." The number of these observations may vary across the differenent imputation samples m if we impute whether households have the relevant item or not. HMR = household main residence.

holds with more valuable inheritances tend to respond to this question less often than households with less valuable inheritances. The largest increases in comparison to the unimputed sample occur when imputing financial assets (e.g. the market value of stocks). Households' interval responses again play an important part here, as they provide very precise information for the imputations (see also table 5).

However, for some variables, the mean does not change significantly; in some cases, it even decreases. For example, the mean amount spent on food eaten at home does not change significantly after imputation, due to the low item nonresponse rate of this variable (table 5). The mean gross income from financial investments is even lower after imputation than before imputation, which suggests that nonrespondents with regard to this variable tend to have lower income from financial investments.

Finally, table 7 also shows that the uncertainty of imputations can vary a lot depending on the variables. For some variables (e.g. mortgage 1 that uses the main residence as collateral), the means show a relatively high variance among the five multiple imputation samples, signaling the uncertainty of the imputed values due to the lower number of observations for these variables. For other variables (e.g. gross income from unemployment benefits or the monthly amount paid as rent) the means show a relatively low variance among the five multiple imputation samples, which in turn signals a higher precision of the imputed values. Had we conducted a single imputation of the variables – with only one imputation sample – instead of multiple imputations, the variance of the estimates would be too low, since the uncertainty behind the imputed values would be disregarded, and they would thus be treated like true values.

5.6 Concluding Remarks

We have seen that imputation is necessary for analyzing the HFCS dataset because, compared with complete-case analysis, it decreases the nonresponse bias of estimates when complete observations differ systematically from incomplete ones. It also decreases the loss of information in analyses because no observations need to be deleted. We chose a multiple imputation procedure known as *multiple imputation by chained equations* to create five multiple imputation samples. For information on analyzing multiply imputed data in Stata®, please see the HFCS User Guide (chapter 9).