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 were disregarded, it would lead to biased estimates. For the HFCS data, we therefore used multiple imputation with 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 in terms of maintaining 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. To account for the uncertainty of the missing values, not just one value per missing value is imputed, but several (in the case of the HFCS, five).

Similar surveys – such as the U.S. 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 datasets which are already imputed. 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 possibility 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 describes the imputation procedure used, and in section 5.4 we explain the specification of the imputation model and how the imputations were executed. Finally, some imputation results are presented in section 5.5.

5.2 Item nonresponse

Table 5 shows selected statistics on item nonresponse. On average, each household has 29.9 missing values, which means that item nonresponse was limited to 2.1% of all the questions (variables) addressed to each household. However, the respective percentage for the euro variables amounts to 4.7%. This 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 entails deleting all households that have missing values in any of the variables of interest and basing the analyses solely on complete

[1] *A common related problem that occurs in surveys is unit nonresponse, which means that no questions are answered at all because, for example, a household declined to take part in the survey. This problem is addressed with the construction of HFCS nonresponse weights (chapter 7).*

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

Table 5

Item nonresponse per household (unweighted)

	Mean	Median	Minimum	Maximum
Number of variables asked				
all variables	1,392.0	1,391.0	1,109	1,889
euro variables	63.0	64.0	36	106
Number of variables with missing values				
all variables	29.9	18.0	0	487
euro variables	3.0	2.0	0	36
Share of variables with missing values in %				
all variables	2.1	1.3	0.0	32.0
euro variables	4.7	3.0	0.0	49.2

Source: HFCS Austria 2014, 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.

observations. However, the loss of information resulting from this method leads to two problems: First, it biases estimates if complete observations differ systematically from incomplete ones; second, even if an estimate is unbiased, the 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 6 shows item nonresponse rates across some selected variables.

Table 6 shows that for example when asked about the value of their main residence, 77.1% of households provided a specific amount (column 3). The other

Table 6

Item nonresponse for selected variables (unweighted)

	Household has item		Responses by households that have the item			
	Yes	Unknown	Amount	Range	"Don't know"/ "No answer"	Other missing values[1]
	(1)	(2)	(3)	(4)	(5)	(6)
	%					
Value of main residence[2]	42.9	0.0	77.1	19.4	3.2	0.4
HMR mortgage 1: amount still owed	12.7	0.6	69.4	14.9	15.4	0.3
Monthly amount paid as rent	50.9	0.0	61.7	37.8	0.4	0.1
Other property 1: current value	11.1	0.3	78.4	14.7	6.0	0.9
Other property mortgage 1: amount still owed	1.3	0.3	77.5	10.0	12.5	0.0
Value of sight accounts	99.2	0.0	80.4	9.5	9.5	0.6
Value of saving accounts	84.0	1.2	72.5	14.2	9.6	3.7
Value of publicly traded shares	5.0	0.4	66.2	13.3	18.5	2.0
Amount owed to household	7.6	0.4	94.3	2.6	3.1	0.0
Employment status (main activity) (person 1)	100.0	0.0	100.0	0.0	0.0	0.0
Gross employee income (person 1)	48.7	0.0	85.0	9.9	3.7	1.4
Gross income from unemployment benefits (person 1)	5.5	0.1	89.8	5.4	3.0	1.8
Gross income from financial investments	63.0	15.0	46.6	33.0	19.1	1.3
Gift/inheritance 1: value	26.7	1.3	77.8	9.6	9.0	3.5
Amount spent on food at home	100.0	0.0	98.4	1.4	0.2	0.0

Source: HFCS Austria 2014, OeNB.

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

Note: HMR = household main residence.

22.9% of households are item nonrespondents: Either they provided a (prespecified or individual) interval (19.4%, column 4), responded with "Don't know" or "No answer" (3.2%, column 5) or their response was set to missing (0.4%, column 6).[3] Nonresponse rates[4] vary substantially across items. Variables with high nonresponse rates include e.g. questions related to the value of publicly traded shares (100%−66.2%= 33.8%) and the household's gross income from financial investments (100%−46.6%=53.4%). With regard to the latter, 33% of households provided at least a range for this type of income, which confirms the importance of asking range questions when a euro question remained unanswered. Range questions provide valuable and often very precise information (see the online appendix and section 2.6.2 for the questionnaire and information on the design of euro loops). A variable with a low nonresponse rate is, e.g., the amount spent on food consumed at home (100%−98.4%=1.6%).

Table 6 (column 2) also shows another aspect of item nonresponse in the HFCS: There are variables known as branch variables (see chart 3 in chapter 4) that may also have missing values due to nonresponses to a higher-order question (head variable) and that are thus set to 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 affirmatively (63%) are then asked the question on the amount of income; the other households, including the 15% of households that did not answer the yes/no question, automatically skip the euro question. As it is unknown, however, whether the 15% 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 considered as second-order (or higher-order) missing values when analyzing nonresponse to a 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 would be considerable also because of 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.

For illustration purposes, table 7 shows a regression of nonresponse to

Table 7

Logit regression of nonresponse in the euro question on value of sight accounts (unweighted)

Covariates	Coefficient
Female (person 1)	−0.0777
	(0.0966)
Age (person 1)	−0.00382
	(0.00338)
Tertiary education level (person 1)	−0.00624
	(0.127)
Employed/self-employed (person 1)	−0.315***
	(0.115)
Residence is in Vienna	−0.642***
	(0.132)
Size of main residence	0.00225**
	(0.000987)
Household size	0.282***
	(0.0440)
Constant	−1.750***
	(0.248)
Observations[1]	2,940

Source: HFCS Austria 2014, OeNB.

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

[1] *The remaining 57 observations of the dataset show missing values in one of the covariates and/or filter missing remarks 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 by subtracting the value in the "amount" column in table 6 from 100%.*

the question regarding the balance of sight accounts ("1" if the value is missing, "0" otherwise) for several explanatory variables. We can see that item respondents differ significantly from item nonrespondents, because respondents live in smaller main residences and in smaller households, they tend to live in Vienna 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 $Y_1, Y_2, ..., Y_P$ to be imputed.
- Step 2: Fill the missing values of $Y_1, Y_2, ..., Y_P$ with random selected values which were actually observed.
- 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 independent variables, which is chosen from among the HFCS variables without missing values and the variables selected in step 1 (except the one being regressed); the regression sample is restricted 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 (true) values and variables with complete observations, each regression model contains a broad set of independent 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 values or those currently being imputed (step 4). It is important that each regression model also contains a broad set of independent variables with *missing* values in order to preserve the joint distribu-

[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 (2014) 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.*

tion of variables with missing values. If t tends to infinity, the imputations of 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 of the imputed values when estimating any variances with imputed variables with missing values. The M imputations of the missing values of $Y_1, Y_2, ..., Y_P$ converge in expectation 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 regressions above 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, separate regressions for each variable reflect the data better, given that the HFCS data contain a large number of variables, many of which have bounds, filter missings, bracketed (i.e. range) responses, interactions or constraints in relation to other variables. This approach thus makes more sense than specifying a joint distribution for all variables together, as is the case for example in the joint modeling approach.[7]

It should be noted 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. In the literature this assumption is referred to as ignorability assumption.

Before running through the five steps above, we need to prepare the data and specify all the parameters of our imputation model: e.g. the selection 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 $Y_1, Y_2, ..., Y_P$ to be imputed. Our strategy is to impute as many variables with missing values as possible, which amounts to around 70% of such variables. The remaining variables with missing values are not imputed with the HFCS imputation procedure due to a lack of sufficient variance or due to a lack of sufficient observations to run a regression.[8]

The imputation of as many variables as possible is intended 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. 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

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

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

many variables for imputation, we could not use them in the regression models as independent 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 on the HFCS imputation procedure, one of the weaknesses 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 $Y_1, Y_2, \ldots, Y_P$ often aids convergence. Therefore, we order the variables to be imputed by their degree of missingness, starting with the variables with the least missing values and ending with those variables that have the most missing values. Variables with the same degree of missingness are imputed in a fixed random order. Head variables are always imputed before their corresponding branch variables. For example, the variable indicating whether a household has a mortgage or not was always imputed before the mortgage amount was imputed, even if the degree of missingness was 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 four different types of regression models. For continuous variables, we used an interval regression model,[9] because all of our continuous variables are bounded either from above or from below, or both (see section 5.4.6 for more details). For binary variables, we used a logit model; and for ordinal and nominal variables, we used ordered logit and multinomial logit models.[10]

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. In the second wave, we decided to use weights as predictors (section 5.4.7), but not for weighted regressions, as is the current practice in imputation (see e.g. Frumento et al., 2012). The reason stated in the literature is that multiple imputations are only meant to appropriately predict missing values (and their uncertainty). Units should not be weighted until later, when statements about the population are to be made on the basis of an analysis of the final dataset.

[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.*

[10] *The nominal variables on the three-digit International Standard Classification of Occupations (ISCO) and the three-digit European statistical classification of economic activities (Nomenclature of Economic Activities, NACE) classifications, which were difficult to estimate with a multinomial logit model because they contain a very large number of categories (74 and 121, respectively), represent the only exceptions. In these two cases, the predictive mean matching (PMM) procedure was used to first, predict a value by linear regression for each missing value and second, impute the observed value that is closest to the regression-predicted value.*

5.4.5 Variable transformations

Before imputing variables with missing values, we transform several of them, as 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 continuous variables is the result of using the natural logarithm. These types of variables usually have a highly skewed distribution; using the logarithm helps to make the distribution closer to the normal distribution assumption that is necessary for the prediction. Another very helpful transformation for year variables is to impute time periods 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 logarithmic transformation mentioned above is carried out on the durations and not on the years.

Another transformation used for some variables with values between "0" and "1" is the log-odds transformation *(log(y/(1–y)))*, for example for the amount of an outstanding consumer loan. Instead of imputing these variables individually, the original amount of the consumer loan (HC0601 to HC0603) is imputed as a first step. Additionally, an indicator showing whether the amount outstanding is smaller than the original amount of the loan is imputed, and if so, the outstanding amount is imputed as a percentage of the original amount. This share is imputed as a log-odds transformation, considerably improving the quality of the imputed values. Subsequently, the individual variables (HC0801 to HC0803) are calculated from the original loan amounts and shares.

For categorical variables, two types of transformations may be used. First, some of the nominal variables can be 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 are transformed into several binary variables by generating 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 continuous 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 that *"The loan has no set length."* When imputing such a loan-length variable, it would not make sense to run the regression over these observations together with those variables 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 continuous 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 continuous and categorical variables with missing values, is that of individual IDs.[11] Individual variables are modeled and imputed separately for each ID in order to avoid biased imputations (section 5.4.8); this should ensure that people with the same IDs display relatively homogenous characteristics if they are modeled together. For this reason, respondents are grouped into new individual ID categories created specifically for the imputations prior to imputation. The criteria for this categorization are as follows: All male financially knowledgeable persons (FKPs), all male partners of FKPs that were individual 2 and all other FKPs are classified as individual 1 (ID = 1). All female partners of FKPs that were already individual 2 and all women that were individual 1 before their male partners became individual 1 are classified as individual 2 (ID = 2). All other people are ordered by age in descending order and are numbered starting with ID = 3.

In the case of households with members that engage in farming, we use a special transformation of the variables for the value of the household's business(es) (HD0801 to HD0803) and the variable for the value of the household's 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. Then we calculate the individual variables (HD0801 to HD0803 and HB0900) based on the sum and percentages imputed. The reason for using this transformation is that it considerably improves the imputed values, as some households with members that engage in farming did not state separate values for their main residence and their agricultural business but indicated only the combined value (see section 4.6.2.7 for further details).

5.4.6 Bounds

As mentioned above, we use interval regression models to impute continuous variables in step 3 because all such variables are bounded either from above or from below, or both. 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 are very unrealistic. Examples of this type of bound include nonnegativity constraints on continuous or count variables (e.g. income or age). For all households the lower bound for these variables is zero. For some continuous variables, we assume that a value above or below a particular general bound cannot occur in practice. As a case in point, the lower bound for the year a loan was taken out (HB1301 to HB1303) is 1945. We assume that no loan in Austria was taken out, renegotiated or refinanced more than 70 years ago. The use of such "empirical" bounds helps avoid imputing extreme outliers of these variables without providing biased results. More examples of general bounds include percentage variables (e.g. share of homeownership for part-owners), where we set the lower bound to zero and the upper bound to 100, or some year variables (e.g. the purchase year of the household's main residence), where the upper bound is 2015, i.e. the year in which the last survey interviews were carried out.

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

Unlike general bounds, individual bounds take different values depending on each household or individual; they usually ensure consistency with other variables from 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. Inversely, when imputing the total estimated consumption 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 a range (either prespecified or individual) in a euro question instead of a specific value. Such ranges are requested if respondents do not provide specific amounts in response to euro questions; they prove very useful for imputation purposes, as they yield valuable and precise information on the missing value from a euro question (see also section 5.2 in connection with table 6).

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 rent excluding utilities 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 distribution among 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 for the variable to be imputed makes the ignorability assumption on which our imputation model relies (see 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 containing several hundred variables, it is, however, not feasible to include all variables, as this may lead to both multicollinearity problems and computational problems. In line with Van Buuren et al. (1999) and Barceló (2006), we have therefore 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 (see 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, gender and employment status of the household's first individual and his/her partner), stratification variables (e.g. province, municipality size), information provided by the interviewers (e.g. standard of living, type of neighborhood, building condition, interview atmosphere, etc.). The latter pieces of information (para-

data) were extremely important for the imputations, since they provided plausible explanations for item nonresponse for many variables.

2. In addition, include variables that are well suited to predicting and explaining the relevant variable to be imputed. This is the classic criterion for using predictors, and it helps to reduce the statistical uncertainty surrounding the imputations. These predictors are identified by their correlation with the variable to be imputed. For example, when imputing loan 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. When imputing the market value of various types of real estate property, we usually include the purchase value, the length of time (in years) for which the household has already owned the respective property and the total value of real estate property owned by the household. Usually, these variables are connected logically (e.g. outstanding principal is the original loan amount minus the sum of all loan repayments). 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 simultaneously.
3. Remove the aforementioned predictor variables that have too many missing values in 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 where the percentage of observed cases within this subsample is below 50% are removed and replaced 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, age, as well as the employment and marital status of the first individual.
4. Include all variables that appear in the models that will be applied to the data after imputation. In other words, consider which different economic theories might be tested based on the data and include those variables as predictors that are expected, according to these theories, to influence or explain the variable to be imputed. Failure to do so will tend to bias the results of potential data users 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 wealth variables.

Obviously, many variables in the survey – for example, the income, age or education of the first individual – fulfill more than one criterion for selecting predictors.

We also include the final survey weights in all regression models (see the discussion in section 5.4.4) and an interaction term, as well as a main effect dummy for each of the above-mentioned predictor variables that households that were asked about the variable to be imputed were not asked about. For example, suppose that we want to impute a household's consumption expenditure using the mortgage amount as one of our predictors. While every household in the sample was asked about consumption expenditure, not all of them were asked about mort-

gage amounts. If, for those households that do not have a mortgage, we just set the mortgage amount to zero (which 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 for 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, if possible, each of the above four predictor categories is represented in each regression equation. Typically, the number of predictors used for each regression model is around 20% of the number of observations for the variable to be imputed. More details on the specification of subsamples can be found in the next section.

5.4.8 Specification of subsamples

Each regression in step 3 is estimated over a subsample consisting of all households and individuals 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 mortgages are ordered by importance, which would introduce a bias to our estimates.[12]

A further example is the imputation of individual variables. These are also only regressed over the subsample of people that share the same ID. To ensure the homogeneity of people with the same IDs, respondents are grouped into new ID categories created specifically for the imputation (see section 5.4.5), and which then form the mentioned subsamples. When imputing question by question, as we do, the bias will be very small, 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 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 cycles. Given the large computational effort required for the HFCS imputation model, we set the cycle number for

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

the HFCS imputation model at $t = 10$. Other similar surveys, like the SCF (Kennickell, 1998) and the EFF (Barceló, 2006) use $t = 6$.

Typically, we check convergence graphically by plotting the mean of the imputed values against the cycle number t. Convergence is judged to have occurred as soon as the pattern of the imputed means becomes random and a definite trend can no longer be observed.

In the second wave of the HFCS, we additionally examined the convergence of selected variables using the Gelman-Rubin convergence diagnostic, which is used very frequently in literature (for more details, see e.g. Cowles and Carlin, 1996). According to this diagnostic, convergence of a variable is reached when the variance of an estimate of this variable (e.g. the mean, median or other percentiles) is relatively small between the multiple imputation samples compared to the variance of the same estimate between the cycles.[13] All variables examined in the second wave of the HFCS meet this criterion.[14]

Of course, such tests (just like any other diagnostic test to assess chain convergence) can never confirm the existence of convergence (see section 5.3). But they are useful for pointing out weaknesses of the imputation model or other unusual results that could indicate nonconvergence.

5.4.10 Number of imputation samples

In the 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 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 $M = 5$. In line with the international requirements and 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 8 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 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 87,202 before imputation. After the respective imputations, it

[13] *The Gelman-Rubin diagnostic is the root of [(t−1)/t) + (BV/WV)], with BV denoting the between-chain variance and WV the within-chain variance. If the Gelman-Rubin values are below 1.2 to 1.1, they are usually considered to denote convergence.*

[14] *The following important variables were tested: HB0900, HB1701, HB2801, HB4400, HD1110, HD1210, HD1510, HI0100, HI0200 and HI0310.*

Table 8

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]	285,996	290,833	290,995	292,706	290,210	292,890
HMR mortgage 1: amount still owed	73,205	80,468	86,705	81,603	81,151	85,650
Monthly amount paid as rent	407	393	399	401	396	391
Other property 1: current value	249,384	237,947	248,696	258,458	233,246	246,517
Other property mortgage 1: amount still owed	78,480	81,357	70,089	74,470	67,089	74,713
Value of sight accounts	2,623	2,689	2,695	2,624	2,612	2,528
Value of saving accounts	23,201	26,925	27,293	26,375	26,526	27,389
Value of publicly traded shares	27,584	26,222	32,490	25,038	26,693	31,007
Gross cash employee income (person 1)	27,319	27,677	27,587	27,695	27,509	27,560
Gross income from unemployment benefits (person 1)	6,437	6,504	6,502	6,363	6,482	6,664
Gross income from financial investments	706	523	564	553	596	587
Gift/inheritance 1: value	87,202	92,620	91,502	92,076	100,621	97,088
Amount spent on food at home	373	374	373	373	373	373

Source: HFCS Austria 2014, 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 different imputation samples m if we impute whether households have the relevant item or not. HMR = household main residence.

increases to EUR 92,620 in *m = 1*, EUR 91,502 in *m = 2*, EUR 92,076 in *m = 3*, EUR 100,621 in *m = 4*, and EUR 97,088 in *m = 5*. Thus, on average the imputations increase the mean value of the first gift/inheritance from EUR 87,202 to EUR 94,781, i.e. by 9%. Additionally about one-third of the values imputed in this context are based on interval responses by households, which suggests that households with more valuable inheritances tend to answer with an interval response to this question less often than households with less valuable inheritances. The largest increases in comparison to the unimputed sample occur when imputing savings account holdings and mortgage loans for the main residence. Households' interval responses again play an important part here, as they provide valuable and often very precise information for the imputations (see also table 6).

However, for some variables, the mean does not change significantly and, 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 (see table 6). 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 8 also shows that the uncertainty of imputations can vary a lot depending on the variables. For some variables (e.g. other property 1), 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 mean values 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 shown 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 with 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).