ILLUMINATE THE UNKNOWN:

**EVALUATION OF IMPUTATION PROCEDURES
BASED ON THE SAVE SURVEY**

Michael Ziegelmeyer

235-2011

Illuminate the unknown:

Evaluation of imputation procedures based on the SAVE Survey[1]

Michael Ziegelmeyer[†]

This version: January 21, 2011

Abstract

Questions about monetary variables (such as income, wealth or savings) are key components of questionnaires on household finances. However, missing information on such sensitive topics is a well-known phenomenon which can seriously bias any inference based only on complete cases analysis. Many imputation techniques have been developed and implemented in several surveys. Using the German SAVE data, this paper evaluates different techniques for the imputation of monetary variables implementing a simulation study, where a random pattern of missingness is imposed on the observed values of the variables of interest. New estimation techniques are necessary to overcome the upward bias of monetary variables caused by the initially implemented imputation procedure. A Monte-Carlo simulation based on the observed data shows the superiority of the newly implemented smearing estimate to construct the missing data structure. All waves are consistently imputed using the new method.

Key words: Imputation methods, Monte-Carlo simulation, imputation evaluation, item-nonresponse, missing data, imputation, retransformation, sample surveys, SAVE

JEL classification: C01, C81, C49

1. Introduction

In the last decades large surveys providing detailed information on households' finance have become a pressing necessity for researchers as well as for policy makers in order to better understand how individuals react to important changes in their economic and institutional environment (such as reforms to the pension or the health care system, or the outbreak of a severe financial crisis). Questionnaires about household finances necessarily touch many sensitive topics like households' income, wealth, and saving. These are known to suffer from very high rates of item-nonresponse, a phenomenon which is generally widespread in micro datasets. Mainly two problems arise: First, if multivariate procedures are used to analyze certain effects, all the variables of each unit (household or individual) must be complete. If there is one missing value in a certain variable, this variable has to be dropped or the sample size has to be reduced by all units containing missing values. This observed-case analysis can lead to a serious reduction of the sample size and the associated loss of efficiency. Additionally, the sample size varies with the question investigated, since different variables are needed for analyses. Second, the variable might not be missing at random and the observation probability could be related to certain characteristics or the environment of the respondent, so that estimations based on observed cases might lead to biased results.

In the SAVE study, a German survey focused on households' saving behavior, the missing values are filled with appropriate substitutes using a *"Markov Chain Monte Carlo multiple imputation procedure"* (Schunk, 2007, 2008). This paper contributes to the methodological literature on the imputation of large scale micro datasets by evaluating different techniques for the imputation of monetary variables. A simulation study is implemented, where a random pattern of missingness is imposed on the observed values of the variables of interest. Using the remaining observed values, the generated missings are imputed applying different imputation models and their ability to replicate the missing data structure is then compared using several criteria. The evaluation clearly shows the superiority of the newly implemented smearing estimate with regard to the various measures used. To the best of my knowledge, this is the first simulation study which evaluates different imputation procedures with regard to monetary variables, which are an integral part of surveys about household finances. All waves are consistently imputed using the new method.

The outline of this article is as follows: a general introduction to item-nonresponse and multiple imputation is given in section 2. Section 3 describes the German SAVE Survey and the multiple imputation method for SAVE (MIMS) implemented by Schunk (2007, 2008). Section 4 deals with the evaluation of different algorithms for the imputation of monetary variables. The previous imputation procedure of monetary variables is described, the bias inherent to the previous imputation procedure is demonstrated, and the new imputation procedure and its implementation are discussed. Finally, section 4 compares the performance of the new and the previous imputation procedure based on a simulation study. Section 5 concludes and gives a perspective for further improvements of the imputation methods of the SAVE dataset.

2. General introduction to item-nonresponse and multiple imputation

2.1 Determinants and patterns of item-nonresponse

Item-nonresponse is an inherent phenomenon of surveys. In contrast to unit non-response, where a household refuses to participate in the survey, item-nonresponse is the failure to respond to one or more questions, although the household agreed to participate in the survey. The determinants of item-nonresponse are complex and range from the unwillingness to provide the information asked for (sensitive information) to difficulties to recall events that occurred in the past and not knowing the correct response. Item-nonresponse increases with the complexity and difficulty of the question and is influenced by the interview mode[2] (face to face or self-administered questionnaires), the topic, and structure of the survey (Rässler & Riphahn, 2006, pp. 219-220; Cameron and Trivedi, 2005, p. 923). The extent of item-nonresponse is not random and often correlated with respondents' characteristics such as age and education.

Table 1 shows item-nonresponse rates for selected variables of the SAVE survey from 2003 to 2008. Basic demographic information have very low missing rates: gender, year of birth, German citizenship, and partnership are almost complete; other variables like the number of children or type of employment have missing rates of up to 2.5% over 2003-2008. The item-nonresponse rates increase for the core questions of the SAVE questionnaire. The average missing rate is 9.8% for annual saving and 17.5% for total net income. The lowest missing

---

[2] Essig and Winter (2003) use an experimental setup of the SAVE survey 2001 to investigate interviewer and mode effects.

rate among the listed asset categories refers to home equity. Other asset categories like checking accounts and cash, saving deposits, life insurance policies, or stock and real estate funds have unconditional missing rates between 10% and 20% over the time period 2003-2008.[3] The missing rates of asset categories conditional on observed ownership are even higher and reach up to 40%.

Table 1: Item-nonresponse rates of selected variables

label of variable	missings in %					
	2003/04	2005	2006	2007	2008	2003-2008
sample size	3,154	2,305	3,474	2,931	2,608	14,472
Basic demographic information						
gender	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
year of birth	0.0%	0.7%	0.0%	0.0%	0.0%	0.1%
german citizen	0.0%	0.4%	0.1%	0.0%	0.0%	0.1%
marital status	0.1%	0.1%	0.3%	0.8%	0.8%	0.4%
living with a partner	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
year of birth - partner	0.3%	0.3%	0.2%	0.4%	0.2%	0.3%
do you have children?	0.2%	0.4%	1.7%	2.1%	0.8%	1.1%
number of children	0.4%	0.5%	2.8%	4.4%	2.4%	2.2%
graduation	0.3%	0.7%	0.2%	0.1%	0.1%	0.3%
graduation - partner	1.4%	0.6%	0.4%	0.2%	0.3%	0.6%
professional training	0.2%	6.6%	0.3%	0.0%	0.0%	1.2%
professional training - partner	1.1%	4.1%	0.5%	0.2%	0.2%	1.1%
type of employment	0.4%	2.8%	4.2%	2.3%	2.6%	2.5%
type of employment - partner?	0.4%	1.3%	3.5%	1.7%	2.0%	1.9%
Saving						
desired amount of savings	6.4%	6.2%	6.3%	8.7%	6.9%	6.9%
annual saving	11.8%	9.4%	8.8%	9.9%	8.7%	9.8%
minimum credit balance - yes/no	1.4%	3.7%	3.6%	3.4%	2.3%	2.9%
amount minimum credit balance	3.8%	7.8%	7.7%	8.1%	6.1%	6.6%
Income						
total net income	30.9%	*	15.9%	11.6%	10.1%	17.5%
Wealth						
flat/house owner - yes/no	3.7%	0.5%	1.2%	1.5%	0.8%	1.6%
market value of flat/house	5.1%	4.6%	6.0%	6.7%	5.8%	5.7%
checking accounts and cash, saving deposits - yes/no	8.0%	5.1%	15.6%	14.4%	14.7%	11.8%
amount checking accounts and cash, saving deposits - beginning of the year	21.7%	16.0%	16.6%	**	**	18.2%
amount checking accounts and cash, saving deposits - end of the year	23.1%	17.4%	17.5%	11.4%	11.3%	16.4%
life insurance policies - yes/no	8.0%	5.1%	9.9%	14.2%	13.7%	10.3%
amount life insurance policies - beginning of the year	19.7%	15.1%	18.5%	**	**	18.0%
amount life insurance policies - end of the year	20.3%	15.3%	18.7%	19.4%	18.6%	18.6%
stock and real-estate funds - yes/no	8.0%	5.1%	11.0%	11.2%	10.9%	9.4%
amount stock and real-estate funds - beginning of the year	14.7%	11.2%	14.0%	**	**	13.6%
amount stock and real-estate funds - end of the year	15.1%	11.6%	14.5%	9.1%	9.2%	12.1%
Wealth conditional on an observed ownership	**2003/04**	**2005**	**2006**	**2007**	**2008**	**2003-2008**
market value of flat/house	4%	8%	10%	10%	9%	8%
amount checking accounts and cash, saving deposits - beginning of the year	24%	19%	19%	**	**	22%
amount checking accounts and cash, saving deposits - end of the year	27%	22%	19%	5%	8%	19%
amount life insurance policies - beginning of the year	42%	37%	41%	**	**	40%
amount life insurance policies - end of the year	44%	38%	41%	20%	21%	35%
amount stock and real-estate funds - beginning of the year	37%	31%	30%	**	**	33%
amount stock and real-estate funds - end of the year	39%	33%	29%	7%	9%	25%

*Source: own calculations based on SAVE 2003-2008. *Question differently asked only in 2005. **Not asked in 2007 and 2008.*

---

[3] Although there are many differences in detail, the results are roughly comparable with the EFF of 2002 (Bover, 2004, table 7) or the SCF (Kennickel, 1991, table 1; 1998, table 1) with respect to the missing rates for those having the item. SAVE has higher missing rates in the first step where the respondents report to have an item or not.

The statistical literature distinguishes three kinds of missing data mechanisms, which were formalized by Rubin (1976) and Little and Rubin (1987, 2002) for the first time: first, missing completely at random (MCAR) describes a missing mechanism which does not depend on observed or unobserved variables; second, the missing process is said to be missing at random (MAR) if the missing mechanism depends only on observed characteristic; finally, the process is not missing at random (NMAR) if the missing mechanism is correlated with variables that are not observed. If no correction is made for missing data mechanisms, the estimates (of means, variances, covariances, coefficients) will be unbiased only in the case of MCAR. However, this is normally not the case as can be seen from the determinants of item-nonresponse.

2.2 Why multiple imputation?

Rässler and Riphahn (2006), as well as Cameron and Trivedi (2005) discuss several methods to deal with item-nonresponse. In particular they describe complete case analysis or available case analysis, weighting, model-based procedures, and imputation techniques, where *"imputation is a generic term for filling in missing data with plausible values"* (Schafer, 1997, p. 1). Imputed datasets are especially appealing since the imputed dataset can be analyzed with complete-data methods (Rubin, 1996, p. 474). Analyses can be based on the same imputed dataset and researchers are spared from the time consuming process of dealing with item-nonresponse by themselves.[4] Also appealing is the fact that the imputation procedure can be completely separated from the analysis (Rässler and Riphahn, 2006, p. 223). There is a wide range of different imputation procedures available, which leads to very different results when constructing the missing data structure (Hu and Salvucci, 2001, pp. 4-21). A good imputation procedure should preserve the complete covariance structure of the dataset and should properly reflect the uncertainty inherent in the imputation process. Since deterministic imputations or single stochastic imputations[5] lead to an underestimation of variances, Rubin (1978) introduces multiple imputation. One cannot impute data without

---

[4] In addition, the data provider might have access to additional information which is not publicly available for confidentiality reasons. This information might help to improve the imputation technique (Rubin, 1996, p. 474; Bover, 2004, p. 20).

[5] For single stochastic imputation the corrected variance estimates can be derived separately (Särndal, 1992; Lee et al., 2002).

making assumptions about the missing data mechanism. Multiple imputation can be applied if the missing data mechanism is said to be ignorable.[6]

The basic idea behind multiple imputation is that each missing value is replaced by several imputed values, normally five.[7] Five complete datasets, which differ in their imputed values, are provided to data users. The advantage of multiple imputation is that the uncertainty about the imputation of the applied model and potential model uncertainty can be properly reflected. The normal procedure is to analyze each dataset separately. Larger changes in the size of coefficients or standard errors indicate the potential influence of the imputation procedure. The next step is to combine the five datasets and calculate coefficients and standard errors according to Rubin's rules (Appendix A or Rubin, 1996, pp. 467-477). According to these rules, coefficient estimates are the average over the coefficients generated by the five different datasets, and the adjusted standard errors take both the within-imputation variance as well as between-imputation variance between the five imputed datasets into account.

Multiple imputation is not necessarily the best imputation method for any given problem. Given sufficient time, resources, and the knowledge about the question of interest, even better estimates could be obtained through weighted estimation or model-based procedures. In real-life applications, time and resources are scarce and the questions one might investigate are not all known in advance. Multiple imputations are easy to use and have good properties as simulation studies[8] like Schafer et al. (1996), Graham, Hofer, and MacKinnon (1996), Graham und Schafer (1999), and Rässler und Riphahn (2006) show.

Kennickell (1991, 1994, 1998) was the first who applied a multiple imputation procedure to a large scale micro empirical study about household finances, the Survey of Consumer Finances (SCF). According to Kennickell (1998), the 1989 SCF was the first multiply imputed large-scale survey for all variables. Following Kennickell's example, multiple imputation has also

---

[6] To allow that the relationships between the observed variables are estimated first, and estimates of these relationships are used to predict the missing values, the missing data must fulfill the "ignorable" criteria (Cameron and Trivedi, 2005, pp. 925-927). For that, two assumptions have to hold: first, the MAR (missing at random) assumption makes sure that the probability of a missing value does not depend on the missing value itself after controlling for the other observed variables, which are correlated to the missing value; second, the parameters for the missing values must be unrelated to the parameters which a researcher wants estimate from the data. The MAR assumption is normally not testable, whereas the second assumption is satisfied in most cases. Therefore, the imputation procedure should include all relevant variables and conserve the correlation structure of the dataset when estimating missing values.

[7] Rubin (1987, p. 114) shows that – unless the rate of missing information is very high – five imputated datasets are sufficient to obtain efficient estimates.

[8] For more information about how simulation studies are implemented see subsection 4.4.2.

been implemented in the Spanish Survey of Household Finances (EFF) (Barceló, 2006; Bover, 2004) and the German SAVE survey (Schunk, 2007, 2008).

3. The SAVE dataset 2003-2008

3.1 Introduction

The SAVE survey started in 2001. It was especially designed to better understand the various aspects of the saving behavior of German households. Since 2005 the survey has been repeated on a yearly basis (see figure 1). For a detailed description of scientific background, design, and results the reader is referred to Börsch-Supan et al. (2008). The key contribution of SAVE is the rich set of available control variables out of different areas like health, expectations, attitudes combined with detailed questioning about income, saving, debt, and wealth.

Figure 1: Sample size of SAVE
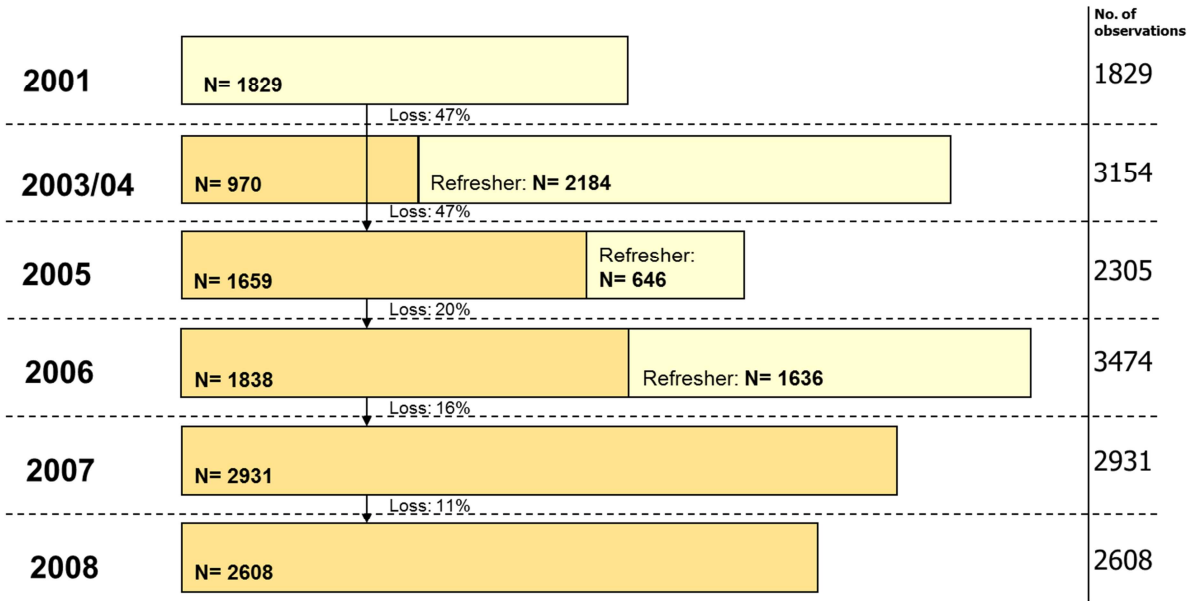
Source: own figure based on SAVE 2001-2008.

3.2 The imputation algorithm for SAVE

The SAVE dataset was imputed every year from 2003 onwards using a "Markov Chain Monte Carlo Multiple Imputation Procedure" to fill the missing values with plausible substitutes. The imputation algorithm is shortly described as follows.[9]

Figure 2: Overview of the multiple imputation method for SAVE
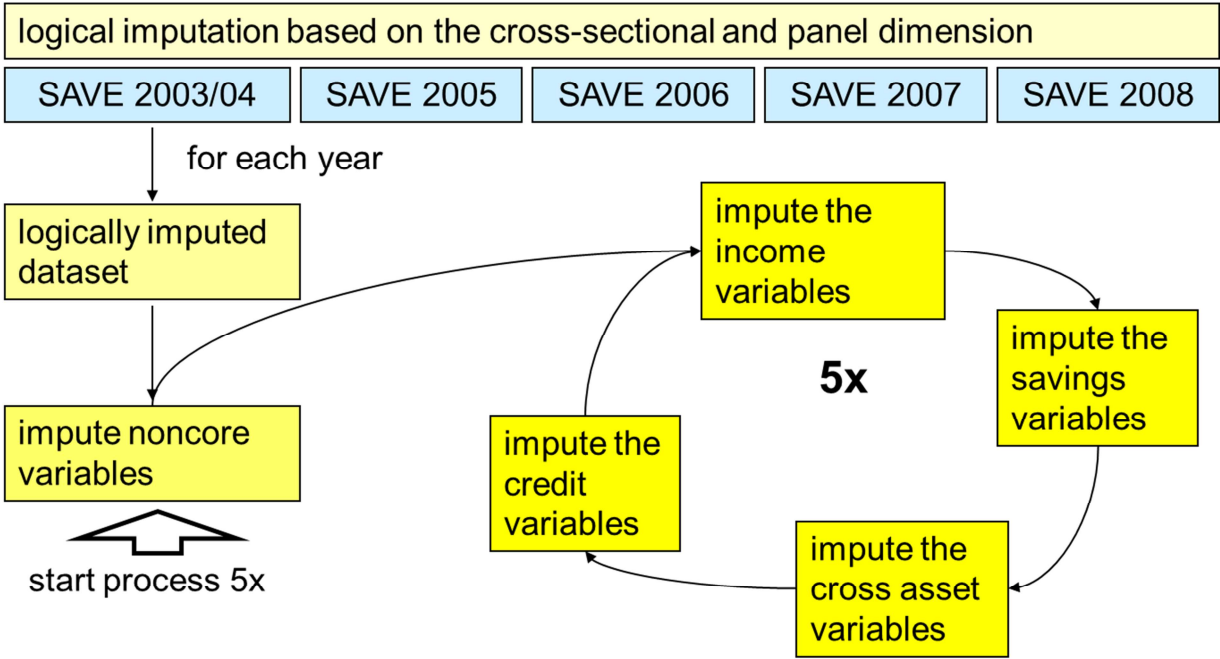
Source: own figure based on Schunk (2008).

Before the stochastic imputation starts, a logical imputation based on the available cross-sectional information is carried out whenever the data structure allowed a unique identification of missing values. Ziegelmeyer (2009a) extends the logical imputation using the panel structure from 2003-2008. The logical panel imputation of the SAVE dataset decisively reduces the number of missing values for some variables. In some cases more than 50% of all missing values can be replaced. Noncore variables as defined in Schunk (2007, 2008) with low missing rates (mainly socio-demographic, psychometric, expectations, and health variables) are stochastically imputed first. Then core variables (income, saving, asset, and credit variables)[10] are imputed making use of the additional information of the already imputed noncore variables. After all gaps are filled, the imputation procedure is repeated for the core variables with a maximum set of covariates because now all variables can be

---

[9] For a detailed description of the whole procedure and the implementation see Schunk (2007, 2008).
[10] These were the variables of main interest when SAVE was set up.

included in the analysis based on the fact that there are no missing values left. The procedure is repeated five times to fulfill convergence criteria. After five loops, the procedure stops and one complete dataset is obtained (for a good description of the iteration process see Barceló, 2006, pp. 19-21). The overall procedure is repeated five times generating five datasets with different imputed values. Figure 2 illustrates the multiple imputation method for SAVE.

4. Regression based stochastic imputation

4.1 Initial implementation of ownership and amount imputations

In an extensive questionnaire like SAVE - with a special focus on saving, income, and wealth - many questions ask for euro amounts. In a first step, the respondent is asked whether the amount is zero or positive.[11] The second step asks for the exact amount. For the first step, a Probit model is estimated for the binary variable, and missing values are predicted.[12] For all respondents with an observed or imputed ownership, the exact amount has to be imputed for all missing values (hurdle model also referred to as two part model: Probit followed by an OLS regression). In the first as well as in the second step, the correlation structure of the data should be maintained. Hence, the imputation method should be able to capture all relevant relationships between variables. This is done by including as many conditional variables as possible.[13] This includes all possible determinants of the variable, all their powers, and interactions (Little and Raghunathan, 1997), as well as potential predictors of missingness (Schafer, 1997).

Since improvements are mainly related to the second step, the imputation of the exact euro amounts is explained in more detail. If one assumes a simple linear relationship between the dependent variable (y) and independent variables (X), ordinary least squares can be used to obtain the estimates of the coefficients (β) based on n observations and k conditioning variables (Barceló, 2006, p. 16; Schunk, 2008, pp. 105-106):

$$y_{obs} = X\beta + u \ , \qquad u|X \sim N\left(0, \sigma^2 I\right). \tag{1}$$

---

[11] For assets, the first question asks about the ownership of a certain asset category. In the case of saving, the first questions ask whether a household saves at all, has a saving goal or not, or whether the household saves for precautionary reasons and so on.

[12] The same procedure was implemented in the EFF (Barceló, 2006) and the SCF (Kennickell, 1998).

[13] It is very important that the imputation model does not impose restrictions on parameters (e.g. the effect of a certain variable is assumed to be zero), which might be later part of an analyst's estimation model. If the imposed restrictions are wrong, then the inference based on imputed data is biased (Schafer, 1997, pp. 139-143).

Assuming that all necessary variables are included to assure the MAR assumption to hold, unbiased estimates of the coefficients ($\hat{\beta} = (X'X)^{-1}(X'y)$) based on the observed values of dependent variable (y_{obs}) can be estimated. Additionally, all the predictor variables (X) must be non-missing (or had themselves been imputed) for both observed and missing cases of the variable to be imputed. In a second step the missing values are stochastically imputed:

$$\hat{y}_{mis} = X\hat{\beta} + \hat{u}, \qquad \hat{u}|X \sim N(0, \hat{\sigma}^2 I). \qquad (2)$$

The missing values are replaced by their best linearly predicted values, $X\hat{\beta}$, plus a random draw $\hat{u}$. It is assumed that $\hat{u}$ is drawn from a normal distribution with mean zero and variance $\hat{\sigma}^2 = \frac{1}{n-k}\left(y'y - y'X(X'X)^{-1}X'y\right)$ (mean squared error of equation 1). The better the fit of equation 1, the lower the variance of the added random draw.[14] The estimation of the coefficients by ordinary least squares are particularly appealing, since maximum likelihood estimation techniques might not converge. Especially in an iterative imputation procedure with a wide range of covariates as in SAVE, not converging equations make the procedure burdensome.[15]

4.2 Bias of the initial imputation method

The procedure to impute euro amounts as described above has two main advantages: the hurdle model is easy to implement, and it allows a maximum amount of covariates (constrained only by the degrees of freedom). However, it also comes with a serious drawback: the model might produce predictions which are out of range of the observed values. In most cases this means that missing values are predicted to be negative. This problem mainly applies to censored metric variables, such as extraordinary incoming payments or inheritances, minimum credit balance, annual saving, precautionary saving, amounts in different asset categories and so on. This problem becomes even more serious if the random draw is added to the predicted value, since this tends to stretch out the distribution and forces even more values to become negative. To overcome this problem in SAVE as well as in other surveys like the EFF (Barceló, 2006, pp. 24-25) and the SCF (Kennickel, 1991, pp. 17-20; Kennickell, 1997, p. 6; Kennickell, 1998, p.8), a so-called "shooting" procedure is

---

[14] The added random variable is censored to the maximum or minimum of +/- 1 (or 1.5) standard deviation around mean zero.

[15] For advantages and disadvantages of other estimation techniques see Schunk (2007, pp. 14-15). Barceló (2006, pp. 22-23) comments on the advantages of linear regression models.

applied. The imputed values above (below) the maximum (minimum) observed values are "shot" with a random value of appropriate sign (i.e. negative (positive) if the imputed value is above (below) the maximum (minimum) observed value), drawn from the same distribution as $\hat{u}$. The "shooting" procedure continues until the obtained value lies within the observed range of values. In other words, "the model draws from the estimated conditional distribution until an outcome is found that satisfies any constraints that may apply" (Kennickell, 1997, p. 6).

Figure 3 visualizes the effect of the initial imputation method using the question about annual saving as an example (see Appendix B for a demonstration of the initial imputation method on other variables).[16]

Figure 3: Distribution of annual saving by each imputation step
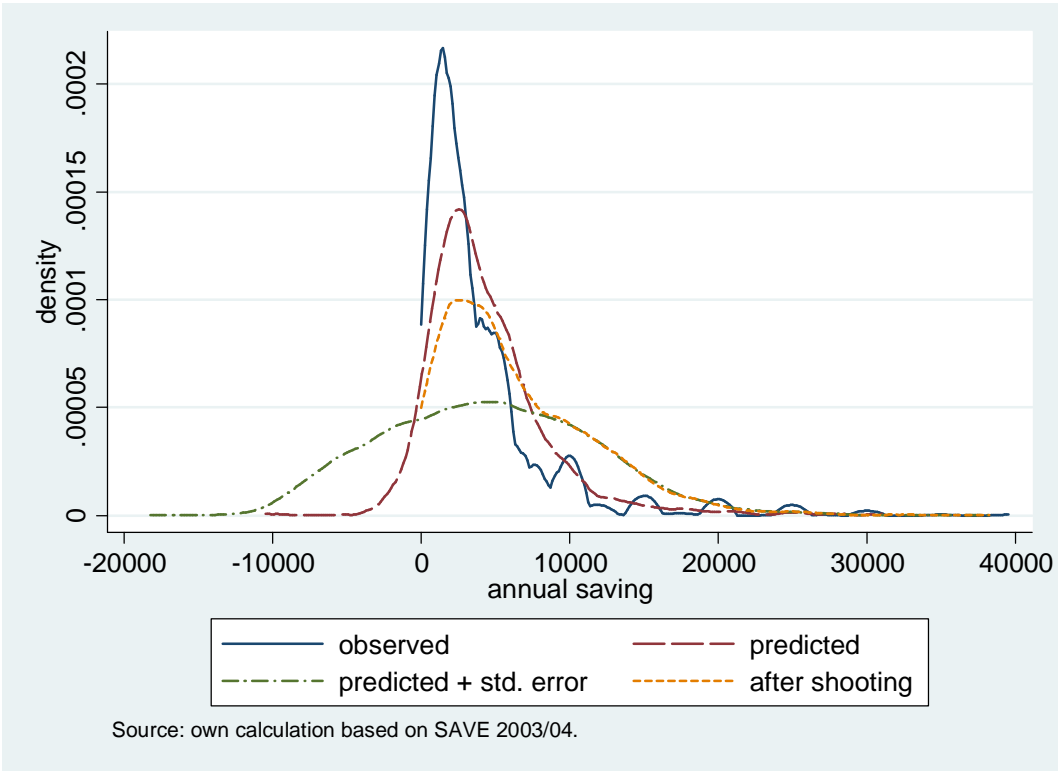
Source: own calculation based on SAVE 2003/04.

Table 2: Mean and median annual saving by each imputation step

	observed	predicted	pred. + error	shooting
mean	4513	4513	4663	6549
median	2500	3519	4452	5136

Source: own calculations based on SAVE 2003/04.

The observed values are imputed (in sample prediction) to compare the distribution of the imputed values to the distribution of the observed ones. The blue line (solid) shows the observed values conditional on respondents having positive saving. The red line (long dashes) displays the predicted values without adding a standard error. Approximately 7% of the predicted values are negative and not plausible based on the question asked. The problem fortifies if a random draw is added to the prediction (green line (dash dot)). As a consequence, the fraction of negative values increases to 28%, and the mean of the positive values scales up, whereas the overall mean remains roughly constant.[17] The positive values are kept and many draws[18] are needed to make the negative values positive by adding additional error terms according to the shooting process described above. Thus, adding the error term and the subsequent shooting process are responsible for a substantial increase in the mean and the median of the imputed values (see table 2).[19] The finally imputed values (yellow line; short dashes) often differ remarkably from values originally estimated by the first regression.

This bias as a result of the initial imputation process is particularly large in SAVE in comparison with similar surveys. Although the same imputation procedure is used for continuous variables in the SCF, the monetary questions are often followed by (unfolding) bracket questions in the SCF if no exact amount can be given. In most cases where the exact amount is missing, at least a range response provides additional information (Kennickell, 1998, table 1). The ranges limit the outcomes allowed and reduce the bias inherent to the imputation procedure outlined in subsection 4.1.

---

[17] The fraction and the results of table 2 are based on the realization of one random draw. The results do not change qualitatively if another random draw is realized. Quantitatively there are differences. An easy to understand setup was chosen to demonstrate the bias. More sophisticated methods are applied in subsection 4.5.

[18] E.g. around 10-20 draws are needed in case of annual saving. Around 80-100 draws are needed in case of checking accounts.

[19] See Rick (2010, pp. 47-51) and Ziegelmeyer (2009b, pp. 39-41) for an initial description of the bias.

4.3 Log-level regression and the retransformation problem

4.3.1 Duan's smearing estimate

Log-level regressions were implemented to overcome the problems of the initial imputation method. The idea behind the implemented change is very simple. The positive values of the dependent variable are transformed taking the normal logarithm of this variable. Thus, the prediction of negative values is no longer possible. An additional advantage of this transformation is that variables of monetary amounts are usually highly skewed to the right. The log transformation reduces, if not eliminates, this skewness. Moreover, when only positive values are observed, which is the case in the second equation of the hurdle model, the dependent variable log(y) often satisfies the assumptions of the classical linear model more closely than models using the level of y. Finally, taking the logarithm of the dependent variable reduces the sensitivity due to outliers since taking logs narrows the range of a variable. Wooldridge (2003, pp. 184-185) suggests as a rule of thumb to take logs when a dependent variable is a positive money amount.

I denote the observations of the untransformed variable as (y), and the transformed observations by (η), where $\eta_i = \ln(y_i)$. Again a linear regression model is applied to the transformed variable:

$$\eta_i = x_i \beta + \varepsilon_i, \tag{3}$$

where $E(\varepsilon_i | x_i) = 0$. The linear relationship is estimated again by ordinary least squares, which is appealing as noted at the end of subsection 4.1. The estimates of the transformed model are always positive, normally more precise, and robust.

However, these advantages are associated with additional costs, which are caused by the so-called retransformation problem (for an introductory discussion see Wooldridge (2003, pp. 202-204) as well as Cameron and Trivedi (2009, pp. 103-104); for a more detailed discussion see Duan (1983) and Manning (1998)). The estimates of the log scale are of no interest. The imputed values must be in the original scale. It may seem natural to use the inverse function of natural logarithm to retransform the transformed scale prediction. The expectation of the individual's response is obtained by:

$$E(y_i | x_i) = E(e^{\eta_i}) = E(e^{x_i \beta + \varepsilon_i}) = e^{x_i \beta} E(e^{\varepsilon_i}) \neq e^{x_i \beta}. \tag{4}$$

$\exp(x_i\hat{\beta})$ is a biased estimate of $E(y_i|x_i)$ even if the true parameters β are known.[20] To obtain an unbiased estimate of $E(y_i|x_i)$ one has to consider the structure of the error term ε_i appropriately. Rewriting equation (4) leads to:

$$E(y_i|x_i) = e^{x_i\beta} \int e^{\varepsilon_i} dF(\varepsilon_i) = e^{x_i\beta} \phi, \tag{5}$$

where $F(.)$ is the cumulative density function of the error term ε_i and ϕ a homoscedastic distribution robust retransformation factor. If one additionally assumes that the error term is log normally distributed and homoscedastic with $\mathrm{var}(\varepsilon_i) = \sigma^2$, term (5) simplifies to:

$$E(y_i|x_i) = e^{x_i\beta + 0.5\sigma^2} = e^{x_i\beta} e^{0.5\sigma^2} > e^{x_i\beta}. \tag{6}$$

In the case of a homoscedastic error term, the correction to obtain unbiased estimates of $E(y_i|x_i)$ are easily calculated by multiplying $e^{x_i\beta}$ with $e^{0.5\sigma^2}$ (naive retransformation). The term $e^{0.5\sigma^2}$ can be easily estimated by $e^{0.5\hat{\sigma}^2}$, where $\hat{\sigma}^2$ is an unbiased estimator of the log-level regression model error. If the error term is not normally distributed, one has to know the specific distribution of the error term. Since the distribution is normally a priori unknown, I use the so-called smearing estimate, a nonparametric retransformation method developed by Duan (1983).

Duan's smearing estimate is obtained as follows: The estimated error ($\hat{\varepsilon}_i = \eta_i - x_i\hat{\beta}$) of equation (3) is used to provide a consistent estimate of a homoscedastic distribution robust retransformation factor $\hat{\phi} = \frac{1}{n}\sum_{i=1}^{n} e^{\hat{\varepsilon}_i}$. The smearing estimate is consistent and fairly efficient even if the error term is normally distributed (Duan, 1983, pp. 606-609). Despite the appealing advantages of Duan's smearing estimate, Mullahy (1998, pp. 254-260) points out that it is not sufficient to assume that the error term and the independent variables are *linearly* independent ($E(\varepsilon_i|x_i) = 0$). This assumption does not ensure that $E(\phi(\varepsilon_i)|x_i) = 0$ since ε_i and x_i could be uncorrelated but not independent such that $\phi(\varepsilon_i)$ and x_i are still correlated. If this is the case, the standard (homoskedastic) smearing retransformation factor is likely to be biased. Alternative solutions for the retransformation are, however, not feasible for the imputation of the SAVE survey. Calculating the smearing estimates for p distinct subgroups of the vector X, for example, is not a practical way for many variables since the subgroups of X had to be

---

[20] The inequality is based on Jensen's inequality for convex functions:
$$E(y_i|x_i) = e^{x_i\beta} E(e^{\varepsilon_i}) > e^{x_i\beta} e^{E(\varepsilon_i)} = e^{x_i\beta}.$$

defined for every variable in a different way depending on the correlation between $\phi(\varepsilon_i)$ and x_i. Similarly, the modified two-part model and the exponential conditional mean model proposed by Mullahy (1998, pp. 260-269) are not implementable. Both models, in fact, are nonlinear and are estimated by nonlinear least squares. Arbitrary starting values cannot be used since the algorithm might not converge. The adjustment of starting values is not appropriate for an imputation procedure since it is too time consuming and the complete imputation procedure should run from the beginning to the end without any stop caused by non-converging estimates. Furthermore, additional assumptions about the error term structure have to hold. If the log scale error is heavy-tailed, the suggested alternative models yield very imprecise estimates (Manning and Mullahy, 2001, pp. 462, 474-475). See Manning and Mullahy (2001) for further remarks on the evaluation of estimators.

In the following, Duan's smearing estimate, which calculates the expectation of exponentiated error term based on the average of the model's exponentiated residuals, will be applied. Given the assumption stated above, the smearing estimate provides a consistent estimate of $E(y_i|x_i)$ using least squared residuals. The magnitude of the bias from ignoring heteroscedasticity is unknown ex ante. Thus, it is important to take the structure of the error term into account in the imputation of all questions about euro amounts. It is not possible to additionally account for a correlation between $\phi(\varepsilon_i)$ and x_i since all the suggestions made in the literature are not practical from an imputation perspective. The next subsection explains the implementation of the smearing estimate in the SAVE imputation algorithm.

4.3.2   Implementation of Duan's smearing estimate in the SAVE imputation algorithm

After the estimation of a probit model for zero responses as described at the beginning of subsection 4.1, the procedure is summarized as follows:

- o   All values above the 99 percent percentile are excluded from the subsequent regression. This is done to reduce the influence of outliers on the estimated coefficients. It might seem arbitrary to a certain extent to exclude all the values above the 99 percent percentile, whereas a procedure for outlier detection based on other covariates should be preferred. But as Kennickell (1991, p. 18) states the restricted staff resources make it infeasible to explore this dimension systematically.

o The observed positive values of the dependent variable are transformed taking the natural logarithm.[21] An OLS regression on a maximum set of explanatory variables is performed (for more details on the conditioning variables see Schunk (2007, pp. 16-18)).

o For the observed as well as the missing observations, individual values are predicted based on Duan's smearing estimate.

o A randomly drawn error is added to the prediction since the imputation procedure has to reflect the uncertainty of the model, which predicts the missing values. The error term is drawn from the empirical distribution of the difference between the predicted and the observed values.[22] Finally, the error term is drawn in a way that the final outcome is within the range of the observed values.

This procedure is adapted to all questions about euro amounts for the years 2003-2008. The next subsection describes the evaluation procedure to compare the quality of the new and the old imputation method to estimate missing values for questions about euro amounts.

4.4 Evaluation of imputation methods

4.4.1 Evaluation measures

Ideally, an imputation procedure should lead to statistically valid inference, which means the efficient reproduction of the key outputs from statistical analyses on a fully observed dataset. I investigate alternative measures of performance since it is normally unknown which statistical analyses will be performed. In addition, missing values are unknown. Because the comparison of imputation methods is restricted to questions about euro amounts, I focus on evaluation methods for continuous variables. The basis of the evaluation consists of the following list of measures, which are common measures to judge the quality of estimates (Chambers, 2003, pp. 11-20; Rässler and Riphahn, 2006, pp. 227-228; Hu, Cohen, and Salvucci, 1998, pp. 311-313) and are not mutually exclusive. The list of measures is ordered by desirable properties for an imputation procedure (Chambers, 2003, pp. 11-12).[23]

---

[21] As described in Barceló (2006, p. 36), euro amounts are as well transformed by taking the logarithm in the EFF of 2002. In their paper there is no statement whether at all or how they deal with the retransformation problem.

[22] Usually, the random error is drawn from a normal error distribution. To draw the error term from the actual distribution is shortly discussed by Graham und Schafer (1999, p. 6). Normally, both applications should yield very similar results.

[23] An important other measure, which is often used to evaluate an imputation procedure, is the correlation between the imputed variable and other key variables. It is necessary to restrict the calculation of correlations to

(1) <u>Predictive accuracy:</u> The imputed value should be as close as possible to the true data value. Three measures are evaluated:

- mean absolute deviation:[24] $d_{absolute}^{mean}\left(\hat{Y}_i, Y_i^{obs}\right) = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{Y}_i - Y_i^{obs}\right|$

- median absolute deviation: $d_{absolute}^{median}\left(\hat{Y}_i, Y_i^{obs}\right) = \left|\hat{Y}_i - Y_i^{obs}\right|^{med}$

- square root of the mean square error (MSE): $d_{squared}\left(\hat{Y}_i, Y_i^{obs}\right) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i^{obs}\right)^2}$

The third measure attaches more importance to larger errors whereas all deviations from the observed values are equally weighted for the first two measures. Two relative measures are calculated additionally:

- mean relative deviation: $d_{relative}^{mean}\left(\hat{Y}_i, Y_i^{obs}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\hat{Y}_i - Y_i^{obs}\right|}{Y_i^{obs}}$

- median relative deviation: $d_{relative}^{mean}\left(\hat{Y}_i, Y_i^{obs}\right) = \left(\frac{\left|\hat{Y}_i - Y_i^{obs}\right|}{Y_i^{obs}}\right)^{med}$

(2) <u>Distributional accuracy:</u> The distribution of the imputed values should be as close as possible to the distribution of the observed values. The implemented measures compare the 25th, 50th, and 75th percentile of the distribution of imputed values to the true data values.

- 25th percentile bias: $d_{bias}^{p25}\left(\hat{Y}_i, Y_i^{obs}\right) = \left(\hat{Y}_i\right)^{p25} - \left(Y_i^{obs}\right)^{p25}$

- median bias: $d_{bias}^{median}\left(\hat{Y}_i, Y_i^{obs}\right) = \left(\hat{Y}_i\right)^{med} - \left(Y_i^{obs}\right)^{med}$

- 75th percentile bias: $d_{bias}^{p75}\left(\hat{Y}_i, Y_i^{obs}\right) = \left(\hat{Y}_i\right)^{p75} - \left(Y_i^{obs}\right)^{p75}$

(3) <u>Estimation accuracy:</u> The imputed values should reproduce the lower order moments of the distribution of observed values. The first and second (centered) moments are evaluated.

- mean bias: $d_{bias}^{mean}\left(\hat{Y}_i, Y_i^{obs}\right) = \frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i - \frac{1}{n}\sum_{i=1}^{n}Y_i^{obs}$

- standard deviation bias: $d_{bias}^{std.dev.}\left(\hat{Y}_i, Y_i^{obs}\right) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{\hat{Y}}\right)^2} - \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{obs} - \bar{Y}^{obs}\right)^2}$

---

a certain subset of variables using an extensive dataset like SAVE. However, the imputation procedure should preserve the complete correlation structure between all variables. An investigation would be a burdensome task. A much better way is to investigate the predictive accuracy. The closer the imputed value to the observed value, the closer is the correlation of the imputed dataset to the true correlation.

[24] This evaluation measure is often called absolute prediction error (APE).

For a better interpretation of the results, relative biases are calculated for the measures of distributional and estimation accuracy. The relative bias is defined as

$$d_{relative\_bias}\left(\hat{Y}_i, Y_i^{obs}\right) = \frac{f\left(\hat{Y}_i\right) - f\left(Y_i^{obs}\right)}{f\left(Y_i^{obs}\right)}$$, where $f(.)$ is a function which computes the mean, certain

percentiles (25th, 50th, and 75th percentile), or the standard deviation, respectively.

To shed more light on the mean bias, two additional measures are provided. The coverage is calculated, i.e., the number of 95% confidence intervals out of 1000 that contain the true mean. This means that the 95% confidence interval[25] of the mean based on certain simulation run must contain the true value based on all observations, which include the observed and the imputed values.[26] Moreover, the average width of the 95% confidence interval of the mean is calculated. The imputation method producing shorter confidence intervals is not automatically better. However, if an imputation method produces a higher or equal coverage and has shorter confidence intervals, this imputation method is preferred since it provides more concentrated point estimates around true values.

4.4.2 Evaluation procedure

Since the true values for the missing observations are unobserved, the following procedure is chosen to evaluate the imputation method, which tries to achieve two important goals: first, to ensure comparability to evaluation procedures in the literature (Bello, 1993, 1995; Hu, Cohen, and Salvucci, 1998, p. 310; Hu and Salvucci, 2001; Tseng, Wang, and Lee, 2003; Rässler and Riphahn, 2006, pp. 227-228; Wasito and Mirkin, 2006); and second, to provide an evaluation procedure that is as close as possible to data applications in the real world. The main difference to the previous literature is that the literature evaluates imputation procedures based on generated data, where the underlying data properties (mean, variance, and correlation of variables) are known. Samples are drawn from a well-defined universe, and a certain missing process, which varies in the missing mechanism (MCAR, MAR, NMAR) and the missing rates (e.g. 5%, 10%, 20%, 30%), is used. Since the imputation procedures here are applied to real data, the missing mechanism should reflect the missing mechanism in the data

---

[25] I assume that the variable of interest is normally distributed.
[26] This is different to the measures before, for which the reference category is always the observed values which have been deleted.

as close as possible.[27] To model this missing mechanism, the missing procedure is assumed to be MAR and a probit model is applied to estimate the probability of being a missing value. In addition, the missing rate is fixed at the observed missing rate. However, such a model is purely deterministic and allows splitting the sample only into one predictive and one test sample. To circumvent this problem, a stochastic process is included to determine the missing values and a sample with replacement is drawn from all observed values. The outline of the evaluation procedure is as follows:

(0) The missing mechanism is estimated using a probit model (0 = observed; 1 = missing) on a maximum set of explanatory variables out of the dataset.[28] Based on this model, the likelihood $p_i \in (0;1)$ for each observation of being a missing value is estimated. This assumes that the missing process is MAR. Appendix C provides several robustness tests if the MAR does not hold.

(1) The sample is restricted to all positive observed cases (N). I assume that the observations reflect the true and underlying data universe.

(2) A sample of the same size N is drawn from this universe with replacement (random process 1).[29] The sample is split into a prediction sample and a test sample according to the probability of being a missing value p_i (see step 0). Since the probability p_i is purely deterministic, an additional random process is introduced (random process 2). The observation is coded to be a missing value if $p_i > q_i$, where q_i is a random draw from a uniform distribution on the interval $[0 - k, 1 - k)$. k[30] is an adjustment variable to ensure that the size of the test sample relative to the predictive sample corresponds to the number of missing values relative to the number of observed values.[31] The

---

[27] See e.g. Bello (1995, pp. 54-55), Schafer et al. (1996), Graham und Schafer (1999), Jonsson and Wohlin (2004), as well as Giorgi et al. (2008) for applications of simulation studies using real data.

[28] Schafer et al. (1996) use a nonparametric hotdeck procedure to generate the patterns of nonresponse. Due to the very limited number of conditioning variables a hotdeck procedure was not implemented as data generating process.

[29] The drawback of drawing a sample with replacement might be the excessive duplications of units. The drawn sample might look relatively unrealistic since the variability of the true values is not reflected (Schafer et al., 1996, p. 30). The preferred way is to draw a random sample without replacement. Problematic is the sample size from which the distribution without replacement can be drawn. E.g. the maximum number of positive observed values is 1617 for annual saving; the minimum number of positive observed values is 352 for stock and real estate funds (SAVE 2003/04). To assure enough variability between the drawn samples, the sample size must be reduced by at least 50%. Especially for life insurance as well as stock and real estate funds the number of observations would be not sufficient any more for the subsequent deletion of values and the prediction based on the test sample.

[30] The adjustment by variable k is necessary since the number of missing values must not correspond to the observed fraction of missing values due to drawing a sample with replacement and due to the random draw of q_i. E.g. if the fraction of missing values is too low, q_i has to be reduced to allow more p_i to be above q_i.

[31] For variables, where the ownership must be imputed first, the relative size of the test sample corresponds to the fraction of missing values conditional on imputed ownership based on a deterministic ownership imputation (to hold the number of positive ownership constant).

procedure guarantees that observations with a higher likelihood of being a missing value are coded more often as a missing value. All the missing observations are part of the test sample and the non-missing values are part of the prediction sample. The prediction sample is used to estimate the parameters. These parameters are applied to the test sample and are used to predict the variable of interest (out-of-sample prediction).[32]

(3) The missing values of the test sample are imputed. Here the third random process comes into play, to add an error term which reflects the uncertainty in the imputation procedure (see subsection 4.3.2 for details).

(4) Calculate the evaluation measures based on the predicted and observed values for the test sample and store them.

(5) Since there are three random processes in play, the whole simulation process (splitting the sample (2), performing the imputation (3), and the calculation of the evaluation measures (4)) are repeated 1000 times.[33] The final evaluation measures are the average over the 1000 iterations.

To the best of my knowledge, this is the first simulation study about the evaluation of imputation procedures on a survey about household finances.[34] This is surprising since many large scale micro dataset about household finances impute their missing values.[35] Most closely related is the work of Kennickell (1997; 1998, pp. 10-14), who evaluates the multiple

---

[32] An out of sample prediction is necessary since a good within-sample fit does not necessarily ensure a good out-of-sample fit for the missing values.

[33] This is the number of iterations used in many of the evaluation studies named above.

[34] Schafer et al. (1996) use the Third National Health and Nutritional Examination Survey (NHANES III) and Graham und Schafer (1999) use the Adolescent Alcohol Prevention Trial (AAPT). Jonsson and Wohlin (2004) use a case study on architecture documentation in a large Swedish organization, and Giorgi et al. (2008) use French Cancer Registries. See Aittokallio (2009) for an actual summary of for evaluations of imputation algorithms related to biotechnology, such as gene expression microarrays, biomarker discovery, disease classification, or mass-spectrometry-based proteomics.

Datasets about household finances are different compared to other surveys. Normally each variable has its specific pattern of item-nonresponse. Depending on the kind of variable (nominal, ordinal, continuous), range limitation for continuous variables, the pattern of missingness, and the available set of covariates, imputation methods differ.

[35] Frick and Grabka (2007) compare different imputation methods of annual labor income between British Household Panel Study (BHPS), the German Socio-Economic Panel Study (SOEP), and the Survey of Household, Income and Labour Dynamics in Australia (HILDA). However, the analysis is restricted to a comparison between the observed and all values (including imputed values) without knowing the true values for the imputed values.

Nicoletti and Peracchi (2004) evaluate the imputation of income and poverty measures within the European Community Household Panel (ECHP). Their method to judge the quality of the imputation procedure is to quantify whether relevant information has been excluded from the imputation procedures.

Other surveys lack a detailed description of the imputation process, e.g. the Survey of Household Income and Wealth (SHIW) (Banca D'Italia, 2010, p. 39) or the BHPS (Taylor et al., 2010, pp. A5-22 – A5-24). Surveys like the EFF (Barceló, 2006; Bover, 2004) and the SCF (Kennickell, 1991, 1994, 1998) provide very detailed descriptions of their imputation procedure. A comparison between the imputed values and its nearest neighbors is used to evaluate and correct the imputation procedure of the EFF (Barceló, 2006, p. 40).

imputation procedure of the SCF by its ability to create an entirely simulated dataset that reduces disclosure risk completely when made available to everyone. He compares the distribution of the simulated data by multiple imputations to the original dataset, but the analyses are restricted to one specific imputation procedure and do not compare different ones.

4.5 Results of the evaluation of imputation methods

SAVE contains a wide range of monetary variables. The analysis is restricted to the following variables of interest to make the evaluation manageable: "annual saving", amount held in "checking accounts[36]", "stock and real estate funds[37]", "life insurance", and "home equity[38]". These variables can be considered as the most important ones, as they represent the most common categories of financial wealth, and home equity is the largest asset of households' total wealth, while annual saving is a key question of a questionnaire on "saving and old-age provision".[39] In addition, the analysis is done for the SAVE survey of 2003/2004, but applies also to the other years.[40] Table 3 provides some information about the simulations for each variable.

Table 3: Simulation details

variable	imputation procedure	obs > 0	missings in %	dublications maximum	mean	median
annual saving	old	1617	15%	5.77	2.00	2.00
	new	1617	15%	5.72	2.00	2.00
home equity	old	1344	7%	5.66	2.00	2.00
	new	1344	7%	5.66	2.00	2.00
saving deposits	old	1307	33%	5.59	2.00	2.00
	new	1307	33%	5.65	2.00	2.00
life insurance	old	501	46%	5.16	2.00	2.00
	new	501	46%	5.18	2.00	2.00
stock and real estate funds	old	352	41%	4.94	2.00	2.00
	new	352	41%	4.98	2.00	2.00

Source: own calculation based on SAVE 2003/04.

---

[36] This variable includes cash and saving deposits like saving accounts, fixed deposit accounts, and saving plans.

[37] This variable also includes mixed funds, reverse convertible notes, listed funds, and similar assets.

[38] Estimate of how much the own house or flat will sell for.

[39] Household net income is not included in this list since the questions is imputed differently due to the range questions if no exact answer could be given. Other variables, which are considered as less important, are saving goal, precautionary saving, minimum credit balance, additional categories of financial wealth, credits, business assets, and other assets.

[40] The imputation procedure is evaluated at the final imputation iteration (loop 5).

Based on the number of positively observed values, the drawn sample sizes range from 1617 (annual saving) to 352 (stock and real estate funds). The missing rate, which is based on the observed missing rate conditional on imputed ownership, is the lowest for home equity (with around 7%) and rises to 46% for life insurances. This shows that the variables chosen reflect a broad range of different sample sizes and missing rates, which is important since imputation procedures within an extensive dataset like SAVE should be able to cope with small and large predictive samples as well with low and high missing rates. As described in the previous subsection, a sample with replacement is drawn from the observed values. On average, the sample drawn contains each observation twice (mean as well as median). The average maximum number of duplicates within the drawn dataset increases with the sample size from below 5 identical observations to almost 6 identical observations. Subsequently, the results are presented separately for the evaluation measures of predictive, distributional, and estimation accuracy.

Predictive accuracy

The predicted value of an imputation procedure should be as close as possible to the true values. Table 4 shows different measures of predictive accuracy for the initial and newly implemented imputation procedure.

Table 4: Predictive accuracy

evaluation measure	annual saving			home equity			saving deposits		
	old	new	new/old	old	new	new/old	old	new	new/old
mean absolute dev.	5088	2914	57%	185936	137570	74%	22184	12466	56%
median absolute dev.	3340	1207	36%	142507	59498	42%	14897	4723	32%
mean relative dev.	2.98	1.20	40%	3.73	1.71	46%	43.76	21.10	48%
median relative dev.	0.90	0.43	48%	0.68	0.32	47%	1.91	0.66	34%
square root of MSE	8742	7323	84%	258103	347968	135%	35524	29516	83%

evaluation measure	life insurance			stock and real estate funds		
	old	new	new/old	old	new	new/old
mean absolute dev.	27645	25255	91%	69529	42426	61%
median absolute dev.	16022	9727	61%	36703	7394	20%
mean relative dev.	31.96	20.55	64%	111.39	46.23	42%
median relative dev.	0.92	0.66	73%	5.11	0.75	15%
square root of MSE	52865	49669	94%	124990	126702	101%

Source: own calculation based on SAVE 2003/04.

Taking the question about annual saving as an example, the mean absolute deviation is reduced from € 5,088 of the initial imputation procedure to € 2,914 for the new imputation method. This is a reduction of 43%. The relative reduction of 64% is even larger for the median absolute deviation. Similar results are obtained by comparing the performance of the

new and old imputation procedure with respect to the mean or median relative deviation. E.g. the median relative deviation is reduced from 90% to 43%, which is a reduction of 52%. The superior performance of the new imputation procedure can also be observed for the other four variables, although the magnitude of the improvement varies with the variable and the evaluation measure. If large outliers are more heavily weighted as done by the mean root of the mean square error (MSE), no clear predominance is observed. The square root of the MSE is smaller for the new imputation algorithm in case of annual saving, saving deposits, and life insurance. The performance is worse for home equity and no difference can be observed for stock and real estate funds.

Distributional accuracy

Table 5 displays the absolute and relative 25th, 50th, and 75th percentile bias. Focusing on annual saving again, the old imputation procedure overestimated the 25th percentile by € 1,018, the 50th percentile by € 2,304, and the 75th percentile by € 4,280. The new imputation procedure reduces the overestimation to € 257 for the 25th percentile, € 446 for the 50th percentile, and € 576 for the 75th percentile. Thus, the bias is reduced by 75%, 81%, and 87% respectively. Almost the same reduction is obtained for the measures of the relative bias. The predominance of the new imputation procedure is confirmed by comparing the performance of the old and the new imputation procedure for the remaining variables.

Table 5: Distributional accuracy

evaluation measure	annual saving			home equity			saving deposits		
	old	new	new/old	old	new	new/old	old	new	new/old
p25 bias	1018	257	25%	-2389	-1445	60%	7092	1677	24%
p25 relative bias	72%	19%	27%	-0.30%	-0.08%	26%	337%	81%	24%
median bias	2304	446	19%	46530	-802	-2%	13659	3775	28%
median relative bias	83%	16%	20%	24%	0%	-1%	246%	71%	29%
p75 bias	4280	576	13%	132290	5452	4%	19588	4883	25%
p75 relative bias	85%	12%	14%	50%	3%	5%	127%	36%	28%

evaluation measure	life insurance			stock and real estate funds		
	old	new	new/old	old	new	new/old
p25 bias	5274	1406	27%	18066	487	3%
p25 relative bias	96%	26%	27%	702%	20%	3%
median bias	11176	3875	35%	37438	1884	5%
median relative bias	92%	32%	34%	514%	25%	5%
p75 bias	20427	8086	40%	80923	7619	9%
p75 relative bias	75%	29%	39%	452%	36%	8%

Source: own calculation based on SAVE 2003/04.

Estimation accuracy

The mean bias of annual saving is reduced from € 1,528 for the initial imputation procedure to € -411 for the new imputation procedure, which reflects a reduction of 73% in absolute terms (table 6). The mean relative bias declines from 32% to -8%, which corresponds to a decrease of 75% in absolute terms. Equally strong reductions of the mean bias and the relative mean bias are observed for the remaining variables. The coverage rate of the 95% confidence interval of the mean of home equity is almost equal between the simulations runs of the old and the new imputation technique. The reason is the low missing rate of home equity of below 7%, since the coverage is calculated over the complete drawn sample. The larger the missing rate and the larger the reduction of the mean bias by the new imputation technique, the larger is the increase in the coverage rate from the old to the new imputation algorithm. Whereas the coverage rate increases only from 72% to 89% in the case of annual saving, the coverage rate increases from 16% to 83% in the case of saving deposits.[41] The confidence intervals have roughly equal length for annual saving, home equity, and saving deposits. The confidence intervals are smaller for life insurance as well as stock and real estate funds in the case of the new imputation method. This increases the efficiency of the estimates given the already increased coverage rates for these variables. No imputation procedure shows a better performance over all variables in preserving the standard deviation.

Table 6: Estimation accuracy

evaluation measure	annual saving			home equity			saving deposits		
	old	new	new/old	old	new	new/old	old	new	new/old
mean bias	1538	-411	-27%	41070	-17387	-42%	10646	2133	20%
mean relativ bias	32%	-8%	-25%	18%	-6%	-32%	72%	17%	24%
coverage rate	72%	89%		91%	91%		16%	83%	
width CI	734	736	100%	25822	26530	103%	4064	4016	99%
std. dev. bias	-3438	-3926	114%	-9395	-7092	75%	-4879	-6123	125%
rel. std. dev. bias	-29%	-43%	146%	4%	10%	258%	-12%	-11%	91%

evaluation measure	life insurance			stock and real estate funds		
	old	new	new/old	old	new	new/old
mean bias	13303	9459	71%	59221	8152	14%
mean relativ bias	58%	40%	69%	316%	36%	11%
coverage rate	20%	44%		17%	69%	
width CI	8334	7605	91%	25315	19556	77%
std. dev. bias	17289	9946	58%	49002	-23833	-49%
rel. std. dev. bias	59%	35%	60%	163%	6%	4%

Source: own calculation based on SAVE 2003/04.

---

[41] The 95% coverage rate of the drawn samples without the generation of missing values is close to 95% for annual saving, home equity, saving deposits, and life insurance. Only in the case of stock and real estate funds, the coverage drops to 85%, which might be based on the small sample size. This observation was also made by Graham and Schafer (1999, pp. 23-24), who report severe undercoverage if the corresponding samples are drawn with replacement as done here.

In summary, the simulations demonstrate that the new imputation procedure performs better than the initial one. Despite no better performance can be observed with respect to the square root of the mean square error and the standard deviation bias, the new imputation procedure based on Duan's smearing estimate clearly dominates the initial imputation procedure with regard to all three domains of predictive, distributional, and estimation accuracy. The mean or median (relative) deviation is minimized, the (relative) bias of the 25th, 50th, and 75th percentile and the mean (relative) bias is strongly reduced, the coverage rates of the 95% confidence interval of the mean increased and the lengths of the confidence intervals are at least partly reduced. I show that Duan's smearing estimate is able to deliver reasonable imputed values, is easy to apply, and allows large sets of conditioning variables to preserve the correlation structure of the dataset.

5. Conclusion

The German SAVE survey suffers from the problem of item-nonresponse as do other surveys about household finances. Monetary variables such as income, wealth, and saving are key components of questionnaires about household finances and have normally relatively high missing rates. A "Markov Chain Monte Carlo Multiple Imputation Procedure" was implemented to reduce the bias and efficiency loss caused by missing values (Schunk, 2007, 2008). The goal of the imputation procedure is to construct the missing data structure. However, it is difficult to evaluate the performance of an imputation procedure since the true values which are missing are unknown. This paper documents the evaluation of different imputation procedures to impute monetary variables based on the SAVE survey. To *"illuminate the unknown"*, a pattern of missingness is imposed on all positively observed values and the remaining sample is used to predict the missing values. Since the true values are now known for the missings, evaluation measures are applied to compare the ability of each imputation algorithm to replicate the missing data. Two imputation algorithms are compared. The initially implemented imputation procedure imputes monetary variables by an OLS regression, where a random draw is added until all observations are within the observed range. This imputation procedure shifts the whole distribution to the right since only positive values are allowed. To overcome this problem the logarithm of the dependent variable is taken and the predicted values are retransformed to the original scale by Duan's smearing

estimate. Duan's smearing estimate clearly dominates the initial imputation procedure with regard to all three domains of predictive, distributional, and estimation accuracy.

Now, all the datasets from 2003 to 2008 are based on the same new imputation procedure, which allows a consistent treatment of all waves using panel estimation techniques.[42] A challenging and work intensive improvement would be a multiple panel imputation. This would not only allow for increasing the accuracy of the estimations but also for preserving the correlation structure over the years.

---

[42] The five multiple imputed SAVE datasets are always delivered with an indicator datasets. Before the logical panel imputation was done, each variable in the indicator dataset flagged with "1" implied a missing value and a variable flagged with "0" an observed value. After the logical panel imputation was done, the flag-dataset was updated: "0" indicates an observed value, "1" implies a stochastically imputed missing value and "2" a logically imputed value using the panel structure. This procedure allows the researcher identifying the missing values and the imputation procedure used.

6. Literature

Aittokallio, T. (2009): "Dealing with missing values in large-scale studies: microarray data imputation and beyond," Briefings in Bioinformatics, 11(2), 253-264.

Banca D'Italia (eds.) (2010): "Supplements to the Statistical Bulletin - Sample Surveys: Household Income and Wealth in 2008," New series: Volume XX- Number 8 - 10 February 2010.

Barceló, C. (2006): "Imputation of the 2002 wave of the Spanish Survey of Household Finances (EFF)," Occasional Paper No. 0603, Bank of Spain.

Bello, A. L. (1993): "Choosing among imputation techniques for incomplete multivariate data: a simulation study," Communications in Statistics - Theory and Methods, 22(3), 853-877.

Bello, A. L. (1995): "Imputation techniques in regression analysis: Looking closely at their implementation," Computational Statistics & Data Analysis, 20, 45-57.

Börsch-Supan, A., M. Coppola, L. Essig, A. Eymann, and D. Schunk (2008): "The German SAVE study. Design and Results," Mea Studies 06, MEA-Mannheim Research Institute for the Economics of Aging, University of Mannheim.

Bover, O. (2004): "The Spanish Survey of Household Finances (EFF): Description and Methods of the 2002 Wave," Documentos Ocasionales N.° 0409.

Cameron, A. C., and P. K. Trivedi (2005): "Microeconometrics. Methods and Applications," Cambridge [et al.]: Cambridge University Press.

Cameron, A. C., and P. K. Trivedi (2009): "Microeconometrics using Stata," College Station, Texas: Stata Press.

Chambers, R. (2003): "Evaluation Criteria for Statistical Editing and Imputation," Working Paper for the Euredit Project on the Development and Evaluation of New Methods for Editing and Imputation, University of Southampton, Southampton, UK.

Duan, N. (1983): "Smearing Estimate: A Nonparametric Retransformation Method," Journal of the American Statistical Association, 78(383), 605-610.

Essig, L., and J. Winter (2003): "Item nonresponse to financial questions in household surveys: An experimental study of interviewer and mode effects," Mea-Discussion-Paper 39-2003, MEA-Mannheim Research Institute for the Economics of Aging, University of Mannheim.

Frick, J. R., and M. M. Grabka (2007): "Item non-response and Imputation of Annual Labor Income in Panel Surveys from a Cross-National Perspective," DIW Discussion Paper 736.

Giorgi, R., A. Belot, J. Gaudart, G. Launoy and the French Network of Cancer Registries FRANCIM (2008): "The performance of multiple imputation for missing covariate data within the context of regression relative survival analysis," Statist. Med., 27, 6310–6331.

Graham, J. W., S. M. Hofer, and D. P. MacKinnon (1996): "Maximizing the usefullness of data obtained with planned missing value patterns: An application of maximum likelihood procedures," Multivariate Behavioral Research, 31, 197-218.

Graham, J. W., and J. L. Schafer (1999): "On the performance of multiple imputation for multivariate data with small sample size," in: R. Hoyle (eds): "Statistical Strategies for Small Sample Research," 1-29, Thousand Oaks, CA: Sage.

Hu, M., S. M. Salvucci, and M. P. Cohen (1998): "Evaluation of Some Popular Imputation Algorithms," in Proceedings of the Survey Research Methods Section, American Statistical Association, 308-313.

Hu, M., and S. Salvucci (2001): "A Study of Imputation Algorithms," U.S. Department of Education, National Center for Education Statistics, Working Paper No. 2001–17, by Project Officer, Ralph Lee. Washington, DC.

Jonsson, P., and C. Wohlin (2004): "An Evaluation of k-Nearest Neighbour Imputation Using Likert Data," Proceedings of the Software Metrics, 10th International Symposium, 108-118, September 11-17, 2004.

Kennickell, A. B. (1991): "Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation," Proceedings of the Section on Survey Research Methods, American Statistical Association. Atlanta, Georgia.

Kennickell, A. B. (1994): "Multiple Imputation of the 1983 and 1989 waves of the SCF," presented at the 1994 Annual Meetings of the American Statistical Association, Toronto, Ont.

Kennickell, A. B. (1997): "Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances," SCF Working Paper.

Kennickell, A. B. (1998): "Multiple Imputation in the Survey of Consumer Finances," Proceedings of the 1998 Joint Statistical Meetings, Dallas TX.

Lee, H., E. Rancourt, C. E. Sarndal (2002): "Variance estimation from survey data under single imputation," in: R. M. Groves, D.A. Dillman, J. L. Eltinge, and R. J. A. Little (eds): "Survey Nonresponse," 315-328. Wiley, New York.

Little, R. J. A., and T. E. Raghunathan (1997): "Should Imputation of Missing Data Condition on All Observed Variables?" American Statistical Association Proceedings of the Section on Survey Research Methods, 617-622.

Little, R. J. A., and D. B. Rubin (1987, 2002): "Statistical Analysis with Missing Data," 1. and 2. ed., J. Wiley & Sons, New York.

Manning, W. G. (1998): "The logged dependent variable, heteroscedasticity, and the retransformation problem," Journal of Health Economics, 17, 283-295.

Manning, W. G., and J. Mullahy (2001): "Estimating log models: to transform or not to transform?" Journal of Health Economics, 20, 461-494.

Mullahy, J. (1998): "Much ado about two: reconsidering retransformation and the two-part model in health econometrics," Journal of Health Economics, 17, 247-281.

Nicoletti, C., and F. Peracchi (2004): "The effects on income imputation on micro analyses: Evidence from the ECHP," Working Papers of the Institute for Social and Economic Research, paper 2004-19, Colchester: University of Essex.

Rässler, S., and R. Riphahn (2006): "Survey item nonresponse and its treatment," Allgemeines Statistisches Archiv, 90, 217-232.

Rick, A. (2010): "The Saving Behavior of German Families: Heterogeneity in the Effect of Children on Annual Saving, Saving Motives, and the Regularity of Saving," Mea Studies 11, MEA-Mannheim Research Institute for the Economics of Aging, University of Mannheim.

Rubin, D. B. (1976): "Inference and missing data. Biometrika," 63, 581-592.

Rubin, D. B. (1978): "Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse," Proceedings of the Survey Research Methods Sections of the American Statistical Association, 20-40.

Rubin, D. B. (1987, 2004): "Multiple Imputation for Nonresponse in Surveys," 1. and 2. ed., J. Wiley & Sons, New York.

Rubin, D. B. (1996): "Multiple imputation after 18+ years (with discussion)," Journal of the American Statistical Association, 91, 473-489.

Särndal, C.-E. (1992): "Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used," Survey Methodology, 18, 241-252.

Schafer, J.L., T.M. Ezzatti-Rice, W. Johnson, M. Khare, R.J.A. Little, and D.B. Rubin (1996): "The NHANES III multiple imputation project," Proceedings of the Survey Research Methods Section of the American Statistical Association, 28-27.

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

Schunk, D. (2007): "A Markov Chain Monte Carlo Multiple Imputation Procedure for Dealing with Item Nonresponse in the German SAVE Survey," Mea-Discussion-Paper 121-2007, MEA-Mannheim Research Institute for the Economics of Aging, University of Mannheim.

Schunk, D. (2008): "A Markov chain Monte Carlo algorithm for multiple imputation in large surveys," Advances in Statistical Analysis, 92(1), 101-114.

Taylor, M. F., J. Brice, N. Buck, and E. Prentice-Lane (eds.) (2010): "British Household Panel Survey User Manual – Volume A – Introduction, Technical Report and Appendices," UK Data Archive Study Number 5151 - British Household Panel Survey: Waves 1-18, 1991-2009.

Tseng, S., K. Wang, and C. Lee (2003): "A pre-processing method to deal with missing values by integrating clustering and regression techniques," Applied Artificial Intelligence, 17 (5/6), 535-544.

Wasito, I., and B. Mirkin (2006): "Nearest neighbours in least-squares data imputation algorithms with different missing patterns," Computational Statistics & Data Analysis, 50(4), 926-949.

Wooldridge, J. M. (2003): "Introductory econometrics – A modern approach," 2nd ed., Mason, Ohio u.a.: Thomson.

Ziegelmeyer, M. (2009a): "Documentation of the logical imputation using the panel structure of the 2003-2008 German SAVE Survey." MEA Discussion Paper 173-09, MEA Mannheim.

Ziegelmeyer, M. (2009b): "Analysis of the Precautionary Saving Motive Based on a Subjective Measure (SAVE 2005-2007)," Mea Studies 07, MEA-Mannheim Research Institute for the Economics of Aging, University of Mannheim.

Appendix A: Rubin's Rules for inference based on repeated imputations

The calculation methods to combine the results of *m* repeated imputations are presented by Rubin (1987, 2004). For each single imputed dataset j (j = 1, 2, …, *m*), estimates and standard errors must be stored, where $\hat{Q}_j$ is a scalar point estimate of interest (e.g. a mean or a regression coefficient) and $\hat{U}_j$ is the corresponding standard error of $\hat{Q}_j$.

The overall scalar point estimate is the average of the individual estimates:

$$\overline{Q} = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}_j.$$

The overall standard error S consists of two parts. First, one must calculate the within-imputation variance:

$$\overline{U} = \frac{1}{m}\sum_{j=1}^{m}\hat{U}_j.$$

Second, the between-imputation variance is calculated as follows:

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{Q}_j - \overline{Q}\right)^2.$$

Finally, the overall standard error is:

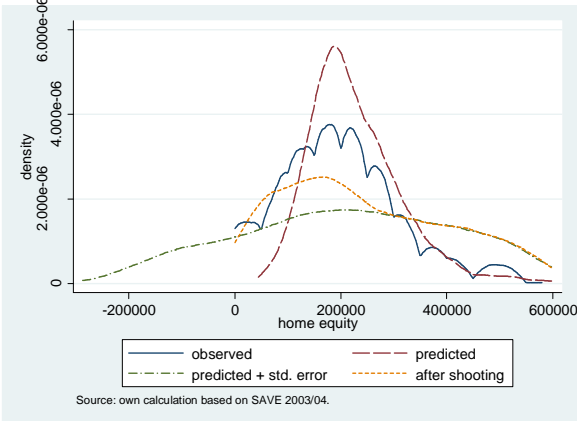$$S = \sqrt{\overline{U} + \left(1 + \frac{1}{m}\right)B}.$$

To test the null hypothesis $Q = 0$, the ratio $t = \overline{Q}/S$ must be compared a Student's t-distribution with *df* degrees of freedom:

$$df = (m-1)\left(1 + \frac{m}{(m+1)}\frac{\overline{U}}{B}\right)^2.$$
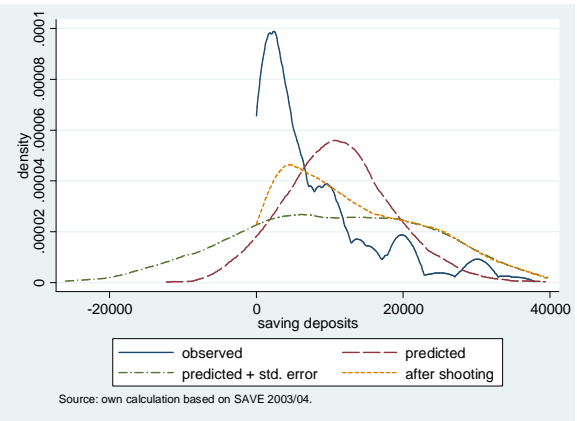
See Schafer (1997, chapter 4) for a review of additional methods for combining the results from multiply imputed data.

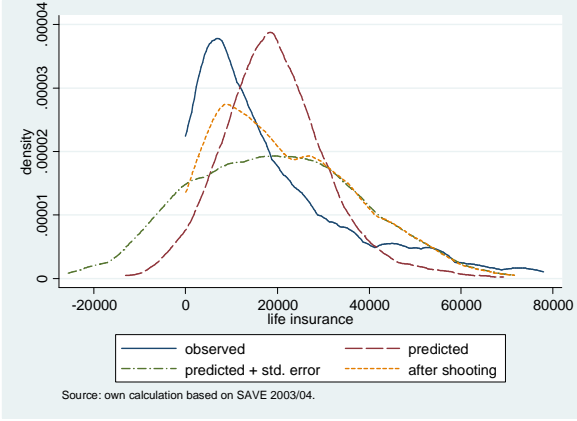# Appendix B: Effect of the initial imputation procedure based on an in sample prediction
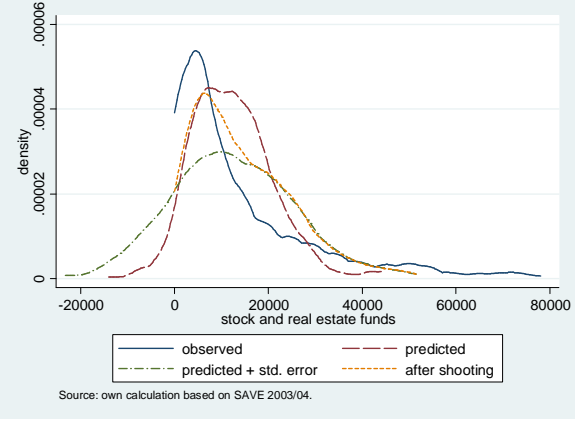
Home equity

Saving deposits

Life insurance

Stock and real estate funds

Appendix C: Not Missing At Random (NMAR)

The MAR assumption is normally not testable. However, it is likely that the MAR does not hold, e.g., more wealthy household are more likely to have a fully booked time schedule. This implies that they have less time to answer the questionnaire, which in turn increases the number of missing values. If the time schedule of an individual is unobserved, the MAR assumption is violated and higher amounts of asset categories are more likely to be missing. The simulation study is changed in the follow way to allow the missing generating process to be NMAR. The missing generating process should rely not only on observed variables but also on unobserved variables, which is modeled in a way that the missing process relies partly on the variable of interest itself. This additional investigation should add additional credibility to the predominance of the new imputation procedure although the MAR assumption is not satisfied.

The difference between the new evaluation procedure and the procedure defined in subsection 4.4.2 is that q_i of step (2) is defined differently. In subsection 4.4.2 q_i is defined as a random draw from a uniform distribution on the interval $[0 - k, 1 - k)$, where k is an adjustment variable to ensure that the size of the test sample relative to the predictive sample corresponds to the number of missing values relative to the number of observed values.

Now q_i is constructed as follows:

- All observations are sorted from small to large values according to the size of the variable of interest, where the smallest values gets number 1 and the largest value gets number N (N = sample size). This new variable ranging from 1 to N is standardized (mean = 0, variance = 1) and called z_i.[43]

- Different not-missing-at-random mechanisms are introduced to test various forms of missingness (Hu and Salvucci, 2001, p. 48):

  o Large values more likely missing: $q_i = rv_i e^{(-(z_i - z_{\min}))} + k$

  o Low values more likely missing: $q_i = rv_i \left(1 - e^{(-(z_i - z_{\min}))}\right) + k$

  o Tail values more likely missing: $q_i = rv_i e^{(-|z_i|)} + k$

  o Center values more likely missing: $q_i = rv_i \left(1 - e^{(-|z_i|)}\right) + k$

---

[43] The reason why the variable of interest is not standardized directly is the skewed distribution of monetary variables. Since the later introduced functions of q_i should work at each end of the distribution in a similar way, the distribution should be symmetrical around the median.

where rv_i is a random draw from a uniform distribution on the interval [0,1), z_{min} is the minimum value of the standardized variable z_i, and k is defined as stated above.

Table 7 shows the pattern of missingness for annual saving[44] generated by functions of q_i over the first to tenth decile. If large values are more likely to be missing, the missing rate in the first decile is around 1-2% and in the tenth decile around 30-32%. If small values are more likely to be missing, the missing rate ranges from around 34-35% in the first decile to around 11-12% in the tenth decile. If tail values are more likely to be missing, the missing rate in the first decile is around 16%, in the fifth decile around 7%, and in the tenth decile around 31-34%. Finally, if center values are more likely to be missing, the missing rate varies from around 4% in the first decile, to around 42-43% in the fifth decile, and around 8-10% in the tenth decile.

Table 7: NMAR – Simulation details

missings in %	MAR	Large values …		Small values …		Tail values …		Center values …	
		old	new	old	new	old	new	old	new
1st decile	12%	2%	1%	34%	35%	16%	16%	4%	4%
2nd decile	14%	3%	3%	19%	19%	15%	15%	7%	6%
2rd decile	13%	5%	5%	13%	13%	10%	10%	8%	8%
4th decile	16%	11%	10%	14%	14%	10%	10%	19%	19%
5th decile	16%	12%	13%	12%	13%	7%	7%	43%	42%
6th decile	17%	19%	19%	13%	13%	10%	9%	33%	31%
7th decile	16%	23%	24%	12%	12%	13%	13%	16%	16%
8th decile	16%	23%	24%	11%	11%	17%	17%	11%	10%
9th decile	16%	24%	26%	11%	11%	24%	25%	9%	9%
10th decile	16%	32%	30%	12%	11%	34%	31%	10%	8%

The header "... more likely missing" spans the NMAR columns.

Source: own calculations based on SAVE 2003/04. Due to drawing the sample with replacement and the draws of the random variable rv_i, the patterns of missingness are not completely equal between the old and new imputation procedure.

The results of the 1000 simulation runs are shown in table 8. The new imputation procedure obtains better results with respect to predictive accuracy over all imposed patterns of not missing at random. With respect to the distributional and estimation accuracy, the performance of the new imputation procedure becomes worse if an increased number of large values are more likely to be missing (large values or tail values more likely to be missing). The better performance of the old imputation procedure is based on the strong upward bias of

---

[44] Since no additional insights are expected from an extended investigation of several variables, the analysis is restricted to annual saving.

the distribution of predicted values. This upward bias corrects for the inability of the imputation model to account for the unobserved fact that more large missing values are unobserved. It becomes clear that modeling a missing process, which is not missing at random, is always arbitrary to a certain extent. The imposed missing structure based on an unobserved process dominates clearly the missing process based on observed characteristics (see table 7 for how clearly the original missing at random pattern is changed). If less large missing values are deleted by the unobserved process, the new imputation procedure shows its superiority also in the domains of distributional and estimation accuracy.

Table 8: NMAR – Evaluation measures

evaluation measure	Missing at random			Large values ...			Small values ... ... more likely missing			Tail values ...			Center values ...		
	old	new	new/old	old	new	new/old	old	new	new/old	old	new	new/old	old	new	new/old
Predictive accuracy:															
mean absolute dev.	5088	2914	57%	5943	4766	80%	5018	2785	55%	5904	4570	77%	5025	2766	55%
mean relative dev.	2.98	1.20	40%	1.11	0.65	59%	4.93	2.45	50%	2.45	1.29	53%	2.10	0.96	46%
median absolute dev.	0.90	0.43	48%	0.67	0.51	76%	1.51	0.68	45%	0.77	0.55	72%	0.97	0.54	56%
std. dev. relative dev.	12.63	5.72	45%	1.94	0.93	48%	10.32	5.80	56%	5.77	3.13	54%	3.71	1.74	47%
median absolute dev.	3340	1207	36%	3386	2305	68%	3527	1509	43%	3440	2019	59%	3301	1568	48%
standard deviation	8742	7323	84%	12217	10975	90%	8055	6299	78%	12106	11415	94%	7941	5695	72%
Distributional accuracy															
p25 bias	1018	257	25%	-369	-751	204%	1673	958	57%	1086	737	68%	734	34	5%
p25 relative bias	72%	19%	27%	-12%	-25%	207%	216%	129%	60%	90%	61%	68%	36%	2%	5%
median bias	2304	446	19%	262	-1199	-457%	3064	1277	42%	787	-278	-35%	2931	1133	39%
median relative bias	83%	16%	20%	6%	-24%	-434%	159%	67%	42%	22%	-5%	-25%	112%	44%	39%
p75 bias	4280	576	13%	371	-2492	-671%	4946	1465	30%	87	-2405	-2779%	6092	2302	38%
p75 relative bias	85%	12%	14%	6%	-28%	-462%	124%	39%	31%	3%	-27%	-999%	159%	61%	39%
Estimation accuracy:															
mean bias	1538	-411	-27%	-1420	-2758	194%	2562	640	25%	-1109	-2143	193%	2748	781	28%
mean relativ bias	32%	-8%	-25%	-17%	-37%	217%	69%	20%	29%	-14%	-31%	219%	66%	21%	32%
coverage mean	72%	89%		61%	37%		48%	93%		67%	50%		46%	92%	
coverage mean true	92%	94%		94%	94%		94%	94%		94%	94%		94%	94%	
width CI	734	736	100%	612	670	109%	740	756	102%	600	649	108%	755	774	102%
std. dev. bias	-3438	-3926	114%	-7440	-7298	98%	-2621	-3055	117%	-8011	-8111	101%	-1776	-1659	93%
rel. std. dev. bias	-29%	-43%	146%	-51%	-63%	124%	-22%	-36%	167%	-54%	-65%	119%	-8%	-15%	178%

Source: own calculations based on SAVE 2003/04.