Allgemeines Statistisches Archiv 90, 217–232
https://doi.org/10.20378/irbo-55155


# Survey item nonresponse and its treatment

By Susanne Rässler and Regina T. Riphahn *

Summary: One of the most salient data problems empirical researchers face is the lack of informative responses in survey data. This contribution briefly surveys the literature on item nonresponse behavior and its determinants before it describes four approaches to address item nonresponse problems: Casewise deletion of observations, weighting, imputation, and model-based procedures. We describe the basic approaches, their strengths and weaknesses and illustrate some of their effects using a simulation study. The paper concludes with some recommendations for the applied researcher.

Keywords: Item nonresponse, imputation, weighting, survey data. JEL C1, C81, C49.

## 1. Introduction

Survey data can be imperfect in various ways. Sampling and noncoverage, unit nonresponse, interviewer error as well as the impact of survey design and administration can affect data quality. For the applied researcher item nonresponse, i. e., missing values among respondents' answers present a regular challenge. This problem receives increasing attention in the literature, where problems of statistical analysis with missing data have been discussed since the early 1970's (e. g., Hartley and Hocking, 1971; Rubin, 1972, 1974; or see Madow *et al.*, 1983).

Even though there exist numerous alternative approaches, most statistical software packages 'solve' the problem of item nonresponse by deleting all observations with incomplete data. This so-called 'complete case analysis' does not only neglect available information but may also yield biased estimates. In their eminent textbook Little and Rubin (1987, 2002) categorize the approaches to deal with missing data in four main groups. Besides complete case analysis there are weighting, imputation, and model-based procedures. Weighting approaches are typically applied to correct for unit nonresponse, i. e., the complete refusal of single respondents to provide information, which may lead to biased estimates as well. The basic idea is to increase the weights of respondents in some subsamples (e. g., among providers of complete data) in order to compensate for missing responses from respondents in other subsamples (e. g., incomplete data providers). Weighting procedures can consider population or sampling weights to align the observable sample with the relevant population.

In contrast, imputation techniques insert values for missing responses and generate an artificially completed dataset. A large number of alternative procedures are applied to choose the values by which missing values are replaced: hot deck imputations use values from other observations in

Received: 28.02.2005 / Revised: 23.08.2005

* We are grateful to an anonymous referee who provided helpful comments. Also we like to thank Donald B. Rubin for helpful comments and always motivating discussions as well as Ralf Münnich for inspiring discussions about raking procedures.

the sample, mean imputation fills missing variables using the mean of appropriate sub-samples, and regression imputation generates predicted values from regression models. Besides these single imputation methods, multiple imputation procedures impute more than one value for each missing value, in order to reflect the uncertainty of missingness and imputation.

Finally, model-based procedures rely on a specified model of the observed data. Inference is based on the likelihood or - in the Bayesian framework - on the posterior distribution under that model. In general, predictions of the missing data are generated based on the respondents' observed characteristics by taking advantage of correlation patterns measured for respondents without missing values. These value substitutions can occur at different levels of complexity.

An evaluation of the properties of the four approaches hinges on the assumptions regarding the nature of the missing values. The crucial role of this missing data mechanism was largely ignored until its concept was formalized by Rubin (1976). Modern statistical literature now distinguishes three cases: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR).

MCAR refers to missing mechanisms which are unrelated to the survey variables, missing or observed. If, for instance, the probability that income is reported is the same for all individuals, regardless of, e. g., their age or income itself, then the missing income data are said to be MCAR. Data are labeled MAR, if the missing mechanism is dependent on observed but not on unobserved variables. This is the case, e. g., if special socio-economic groups are disproportionately subject to missing values and the missingness can be explained by observed variables. Finally, data are termed NMAR, if the missingness depends on the values of the variables that are actually not observed. This might be the case for income reporting, where individuals with higher incomes tend to be less likely to respond, even conditional on their observed data.

The next section describes the prevalence, determinants, and effects of item nonresponse using the German Socioeconomic Panel Survey (GSOEP) as an example. Section 3 discusses the strengths and weaknesses of the alternative approaches to solve the item nonresponse problem. The paper concludes with recommendations for applied researchers.

## 2. Item Nonresponse in the German Socioeconomic Panel

2.1. Prevalence of Item Nonresponse in the GSOEP. The German Socioeconomic Panel is a household panel survey covering a broad range of issues. Its questionnaire has been administered annually since 1984. It now covers over 20,000 individual respondents. The extent of item nonresponse in the GSOEP varies considerably across items. Averaging across the available 19 annual panel waves (1984-2002) we obtain 0.4 percent item nonresponse for subjective health satisfaction, 0.5 percent for political party preference,

8.9 percent for gross monthly labor earnings, and 1.3 percent for the question on whether an individual has disability status.[1]

Riphahn and Serfling (2002, 2005) compared the item nonresponse rates across financial variables in the GSOEP cross-section of 1988. At the individual level item nonresponse rates varied between 2.6 percent e. g., for retirement benefits and 15.3 percent for income from self-employment. Among variables measured at the household level they observe more than 30 percent item nonresponse for questions about interest and annuity payments. In contrast, certain questions on social transfers such as child or welfare benefits yielded nonresponse rates of below one percent.

Schräpler (2004) describes the development of item nonresponse behavior with respect to individual gross labor income. He compares the nonresponse rates of a sample of respondents over the years and finds declining nonresponse rates which differ depending on the method of data collection and respondent characteristics. Other studies confirm that individuals with a low propensity to continue responding to a panel survey are also less likely to disclose their income.

2.2. Determinants and Effects of Item Nonresponse. The theoretical literature on item nonresponse mainly applies two explanatory approaches, the cognitive and the rational choice model (see e. g., Schräpler 2004). Extending theoretical approaches from cognitive psychology to the interview situation, the cognitive model conceptualizes individual response behavior as a multi-stage process (Sudman *et al.*, 1996): after hearing a question it must be interpreted and understood. Next, the respondent gathers the relevant information, a stage which is affected by the complexity of the question. Finally, the information is translated to the answer format required by the questionnaire and possibly adjusted based on objectives such as self representation or social desirability.

In contrast, rational choice theory focuses only on this last stage, when respondents evaluate behavioral alternatives based on their expected costs and benefits (Esser, 1984). Benefits of responding consist of supporting a potentially appreciated cause, and of avoiding the negative effects of refusal such as breaking social norms generated by the interview situation or violating courtesy towards the interviewer. Key costs of answering a survey consist of the potential negative consequence of providing private information (e. g., from tax authorities or through breach of privacy) as well as of the necessary effort to recall the desired facts.

The hypotheses that can be derived from these theories regarding the determinants of item nonresponse behavior relate to the nature of the question (i. e., cognitive complexity and sensitivity), to the relationship between respondent and interviewer, to the interview situation, and finally to the characteristics of the respondent. Dillman *et al.* (2002) provide a classification of seven causes of item nonresponse (INR):

[1] We thank Oliver Serfling for generating these figures.

- Survey Mode: INR is higher in self-administered questionnaires than in face-to-face interviews.
- Interviewers: if the interviewer is able to develop a high level of rapport with respondents, difficult answers may be given willingly. Interviewers' response to unanswered questions affects nonresponse outcomes.
- Question Topic and Structure: certain contents such as finances, drug use, criminal and sexual behavior are notorious for INR. Also, open-ended or multiple-part questions, as well as those with complex branching structures produce more INR.
- Question Difficulty: cognitive difficulty of questions or coverage of long time horizons generate more INR.
- Institutional Policies: sensitive information e. g., sales or investment in business surveys have high INR rates. Offering a 'don't know' answer option also increases INR.
- Respondents' Attributes: in many surveys older and less educated people are less likely to respond.

Schräpler (2004), Frick and Grabka (2003), and Riphahn and Serfling (2005) estimated multivariate models of item nonresponse behavior controlling for relevant indicators. The studies differ in their empirical approach, the subsample taken from the GSOEP, the number of items considered, and in the key issues addressed.

Nevertheless some general findings can be summarized as follows: (i) there is significant heterogeneity in the processes determining item nonresponse behavior across items; (ii) the association between interviewer and respondent characteristics does not appear to be influential for item nonresponse behavior; (iii) item nonresponse rates are higher when the interviewer is female and when a new interviewer is assigned to respondents; (iv) item nonresponse on income is higher at low and high income levels; (v) face-to-face interviews yield lower nonresponse rates than self-reporting or computer assisted interviewing; (vi) item nonresponse and 'don't know' answers are determined by different mechanisms.

As item nonresponse behavior appears to affect financial variables most severely, analyses of income and wealth issues may be most subject to biases deriving from missing data. Given that item nonresponse may indeed bias the results of empirical analyses in general, correction methods need to be considered.

## 3. Dealing with Item Nonresponse

This section discusses four frequently applied methods for the analysis of data with missing values due to item nonresponse:[2]

[2] For a discussion of procedures to avoid item nonresponse in advance, such as interviewer training, questionnaire structure, or administration, see e. g., Groves *et al.* (2002).

3.1. COMPLETE CASE ANALYSIS. Software packages often handle incomplete data by deleting all cases with at least one missing item (listwise deletion or complete case analysis, CC). This practice is inefficient and often leads to substantially biased inferences. Listwise deletion can reduce the available data considerably, so that they are no longer representative of the population of interest.

Thus, CC analysis can be wasteful, as informative data are discarded when they belong to records that have missing values on other variables. As an alternative for univariate analyses often all values that are observed for a variable of interest are used independent of missing values on other variables (available case analysis, AC). A major disadvantage of AC analysis is that different analyses from a given dataset will be performed on different samples, depending on which observations have complete data for each analysis. This can lead to inconsistent estimates especially when comparisons are made using estimates from different subsamples. In general, basing inferences only on the complete cases implies the tacit assumption that the missing data are missing completely at random, which is typically not the case. The size of the resulting bias depends on the degree of violation of the MCAR assumption, the share of missing data, and the specifics of the analysis.

3.2. WEIGHTING. The most common procedure to correct for (unit) nonresponse in official statistics and survey research is weighting. In general, weighting is applied to address problems of nonresponse and to adjust the sample when unequal probabilities of selection have been used. Therefore, two types of weights for a unit $i$, the nonresponse or poststratification weights $g_i$ and the inverse-probability or design weights $d_i = 1/\pi_i$, should be distinguished (Gelman and Carlin, 2002). The former are typically used to correct for differences between sample and population and have to be estimated. The latter are usually known in advance, and are needed to generate unbiased estimates for the population target quantity under repeated sampling given a specific sampling design.

There is common agreement that for estimating population totals, means, and ratios, weighted averages are appropriate. An example are Horvitz-Thompson type estimators which are, e. g., for a population total given by

$$\widehat{Y} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} d_i y_i \,.$$

In combination with complete case analysis weights may also be used to address nonresponse problems. If the probabilities of response for each responding unit were known, then

$$P(\text{selection and response}) = P(\text{selection}) P(\text{response}|\text{selection})$$

(Little and Rubin, 2002) and the individual weights $w_i$ for a unit $i$ are given by $w_i = d_i g_i$. In practice, the response probability is unknown and a standard approach, e. g., is to form adjustment cells based on background variables measured for respondents and nonrespondents. The nonresponse weight for individuals in an adjustment cell is then the inverse of the response rate in that cell.

For illustration, let the sample be divided into $J$ homogeneous cells or groups with respect to the assumed response generating process. Let $n_j$ denote the expected or planned sample size in group or cell $j$, $j = 1, 2, \ldots, J$, e. g., among young working women, and $m_j$ the number of respondents in this group. The individual weight $w_i$ of an observation $i$ within a cell $j$ is computed as $d_i g_i = d_i \frac{n_j}{m_j}$.

If only sample counts are used in the weighting procedure, weighting can be interpreted as a single conditional mean imputation. To illustrate this, consider the so-called weighting-class estimator (Oh and Scheuren, 1983) which is given by

$$\widehat{Y} = \frac{N}{n} \sum_{j=1}^{J} \frac{n_j}{m_j} \sum_{i=1}^{m_j} y_{ij} = \frac{N}{n} \sum_{j=1}^{J} n_j \overline{y}_j^{obs} = \frac{N}{n} \sum_{j=1}^{J} \left( \sum_{i=1}^{m_j} y_{ij} + (n_j - m_j) \overline{y}_j^{obs} \right),$$

where $N/n$ is the sampling fraction. This weighting-class estimator is identical to the estimate derived by single conditional mean imputation. Thus, naive estimates of standard errors and confidence intervals will be biased downwards as it is typically the case with single imputation. The derivation of an unbiased variance estimator is cumbersome.[3]

In practice, the population totals of the cells, one wants to adjust for, are often unknown, but the marginals of different weighting variables are known for the population. In this situation, a set of weighting vectors can be estimated, which satisfies the constraints given by the population margins: this procedure is termed raking. It applies iterated proportional fitting (IPF) to obtain weighted sample counts that match the population on the set of margins. Approaches that make use of auxiliary information comprise regression and ratio estimates; for these and extensions see Deville and Särndal (1992) and Deville *et al.* (1993). To sum up, calibration and raking procedures which include the generalized regression (GREG) estimator and iterative proportional fitting are widely used in the case of unit nonresponse. If, e. g., only a population quantity such as the total is to be estimated, they may also be used in the presence of item nonresponse.

While weighting methods are often relatively easy to implement, they face three major disadvantages: (i) especially in the presence of outliers weighted estimates can have high variances, (ii) variance estimation for weighted estimates can be computationally expensive, if, e. g., linearization or jackknife methods have to be used (see Gelman and Carlin, 2002), and

[3] Notice that often additional information is available and instead of weighting a multiple imputation procedure (see Section 3.5) can be applied successfully, see Rässler and Schnell (2004).

(iii) weighting methods typically do not model the joint distribution of the data as is done by multiple imputation or model-based approaches.

3.3. IMPUTATION TECHNIQUES. Imputation techniques fill in one or more plausible values for each missing datum so that one or more completed datasets are created (i. e., single vs. multiple imputation). Often it is easier to first impute missing values and to then use standard complete-data methods of analysis than to develop statistical techniques that allow the analysis of incomplete data directly. Imputation allows to use information not available to the analyst. Imputation of survey data can be performed separately from the analysis, which is appealing. The application of standard methods on data with singly imputed values will result in underestimated standard errors, if the uncertainty of the imputation procedure is ignored. Due to its operational convenience, single imputation has long been used, especially by statistical offices. Among the key challenges for single imputation is to preserve the covariance structures in the data and at the same time to appropriately reflect the uncertainty due to the imputation process. Usually this means that for every point estimate based on singly imputed data its frequency valid variance estimate has to be derived separately; see Lee *et al.* (2002).

Multiple imputation (MI), introduced by Rubin (1978) and discussed in detail in Rubin (1987, 2004), retains the advantages of imputation while allowing the data analyst to make valid assessments of uncertainty. Multiple imputation reflects uncertainty in the imputation of the missing values through wider confidence intervals and larger $p$-values than under single imputation. MI is a Monte Carlo technique that replaces the missing values by $m > 1$ simulated versions, generated according to a probability distribution which indicates how likely the true values are given the observed data. Typically $m$ is small, e. g., $m = 5$, although with increasing computational power $m$ can be 10 or 20. In general, this depends on the amount of missingness and on the distribution of the parameters to be estimated.

To illustrate this, let $Y_{obs}$ denote the observed components of any uni- or multivariate variable $Y$, and $Y_{mis}$ its missing components. Then, $m$ values are imputed for each missing datum according to some distributional assumptions creating $m > 1$ independent simulated imputations $(Y_{obs}, Y_{mis}^{(1)})$, $(Y_{obs}, Y_{mis}^{(2)})$, ..., $(Y_{obs}, Y_{mis}^{(m)})$. Standard complete-case analysis can be performed for each of the $m$ imputed datasets, enabling us to calculate the imputed data estimate $\widehat{\theta}^{(t)} = \widehat{\theta}(Y_{obs}, Y_{mis}^{(t)})$ along with its estimated variance $\widehat{var}(\widehat{\theta}^{(t)}) = \widehat{var}(\widehat{\theta}(Y_{obs}, Y_{mis}^{(t)}))$, $t = 1, 2, \ldots, m$. The complete-case estimates are combined according to the MI rule that the MI point estimate for $\theta$ is simply the average

$$\widehat{\theta}_{MI} = \frac{1}{m} \sum_{t=1}^{m} \widehat{\theta}^{(t)}. \tag{1}$$

To obtain a standard error $\sqrt{\widehat{var}(\widehat{\theta}_{MI})}$ for the MI estimate $\widehat{\theta}_{MI}$, we first calculate the 'between-imputation' variance

$$\widehat{var}(\widehat{\theta})_{between} = B = \frac{1}{m-1}\sum_{t=1}^{m}(\widehat{\theta}^{(t)} - \widehat{\theta}_{MI})^2, \tag{2}$$

and then the 'within-imputation' variance

$$\widehat{var}(\widehat{\theta})_{within} = W = \frac{1}{m}\sum_{t=1}^{m}\widehat{var}(\widehat{\theta}^{(t)})\,. \tag{3}$$

Finally, the estimated total variance is defined by

$$\widehat{var}(\widehat{\theta}_{MI}) = T = \widehat{var}(\widehat{\theta})_{within} + (1+\frac{1}{m})\widehat{var}(\widehat{\theta})_{between} = W + \frac{m+1}{m}B\,. \tag{4}$$

For large sample sizes, tests and two-sided $(1-\alpha)100\%$ interval estimates can be based on Student's $t$-distribution

$$(\widehat{\theta}_{MI} - \theta)/\sqrt{T} \sim t_v \quad \text{and} \quad \widehat{\theta}_{MI} \pm t_{v,1-\alpha/2}\sqrt{T} \tag{5}$$

with degrees of freedom

$$v = (m-1)\left(1 + \frac{W}{(1+m^{-1})B}\right)^2. \tag{6}$$

MI is in general applicable when the complete-data estimates are asymptotically normal (e. g., ML estimates) or $t$ distributed; see Rubin and Schenker (1986), Rubin (1996), Barnard and Rubin (1999), or Little and Rubin (1987, 2002).

The theoretical motivation for multiple imputation is Bayesian, although the resulting multiple imputation inference is usually also valid from a frequentist viewpoint. Basically, MI requires independent random draws from the posterior predictive distribution of the missing data given the observed data. Usually this is performed by a two-step procedure. First, we take random draws of the parameters according to their observed-data posterior distribution. Second, we perform random draws of the missing data according to their conditional predictive distribution. This is done $m$ times. If only one variable has missing values, such a specification is rather straightforward and univariate (Bayesian) regression models may be applied. When the data have a multivariate structure and different missing data patterns, the observed-data posteriors are often not standard distributions from which random numbers can easily be generated. However, with increasing computational power simpler methods have been developed to enable multiple imputation based on Markov Chain Monte Carlo (MCMC) techniques. Common concerns with multiple imputation address the model-based assumptions and the complexity of the Bayesian posterior predictions. Clearly, there is no assumption-free imputation method but multiple imputation explicitly formulates and evaluates these assumptions. For a broad discussion of advantages and disadvantages of imputation procedures see Groves *et al.* (2002, Chapter 22 and 23).

3.4. MODEL-BASED PROCEDURES. Model-based procedures to adjust for nonresponse simultaneously have to model the distribution of the data $Y$ and the response mechanism $R$. Without any further assumptions regarding the response mechanism, the joint distribution $f_{Y,R}(y,r;\theta,\xi)$ has to be modelled. In so-called nonignorable nonresponse models this is done in two slightly differing ways. On the one hand, selection models as considered by Heckman (1976), specify $f_{Y,R}(y,r;\theta,\xi)$ as

$$f_{Y,R}(y,r;\theta,\xi) = f_Y(y;\theta) f_{R|Y}(r|y;\xi) \tag{7}$$

and have to formulate an explicit model for the distribution of the response missing-data mechanism $f_{R|Y}(r|y;\xi)$ where $\theta$ and $\xi$ are the unknown parameters or in the Bayesian context are random variables as well. Keeping the notation simple, with missing data the likelihood of (7) is

$$L(\theta,\xi;y,r) = \int f_{Y_{obs},Y_{mis}}(y_{obs},y_{mis};\theta) f_{R|Y_{obs},Y_{mis}}(r|y_{obs},y_{mis};\xi) dy_{mis} . \tag{8}$$

On the other hand, pattern-mixture models as discussed by Glynn et al. (1986) factor the joint distribution in a different way:

$$f_{Y,R}(y,r;\theta,\xi) = f_{Y|R}(y|r;\theta) f_R(r;\xi) , \tag{9}$$

where the distribution of $Y$ is conditioned on the missing data pattern $R$. Therefore, the resulting marginal distribution of $Y$ will be a mixture of distributions.

Under the MCAR assumption expressions (7) and (9) are equivalent. If distributional assumptions are added and the data are not MCAR, these specifications can lead to different models. Maximum-likelihood estimates are found by maximizing the likelihood functions with respect to $\theta$ and $\xi$. In the Bayesian context the posterior distribution is obtained by incorporating a prior distribution and performing the necessary integrations.

In general, either way has its merits and demerits. Specification models usually require the existence of identifying restrictions, are very sensitive to model misspecification, and the results are often claimed to be unstable. Pattern-mixture models are often under-identified and also require identifying restrictions. Typically, pattern-mixture models are suggested to be used for sensitivity analyses, see, e. g., Little (1993).

Since the assumption of MAR cannot be contradicted by the observed data, more often the observed-data likelihood, which is also called the likelihood ignoring the missing data mechanism, is considered:

$$L(\theta;y_{obs}) = \int f_{Y_{obs},Y_{mis}}(y_{obs},y_{mis};\theta) dy_{mis} . \tag{10}$$

Inferences about $\theta$ can be based on (10) rather than on the full likelihood (8) if the missing data mechanism is ignorable. Notice that ignorable Bayesian inference would add a prior distribution for $\theta$. Rubin (1976) has shown that an ignorable missing data mechanism is given when two conditions hold.

First, the parameters $\theta$ and $\xi$ have to be distinct, i.e., they are not functionally related or - in the Bayesian framework - are a priori independent. Second, the missing data are MAR.

Ignorable ML methods focussing on the estimation of $\theta$ have a couple of advantages. Usually the interest is in $\theta$ and not in $\xi$. Then the explicit modeling of the response mechanism can be cumbersome and easily misspecified. Also, often information for the joint estimation of $\theta$ and $\xi$ is limited. Thus, estimates assuming MAR data turn out to be more robust in many cases.

However, in many missing data problems, even the observed-data likelihood (10) is complicated and explicit expressions for the ML estimate cannot be derived. Here, the Expectation-Maximization (EM) algorithm is a broadly applicable approach to the iterative computation of maximum likelihood estimates. On each iteration of the EM algorithm there are two steps, called the expectation or E-step and the maximization or M-step. The basic idea of the EM algorithm is first (E-step) to fill in the missing data $Y_{mis}$ by their conditional expectation given the observed data and an initial estimate of the parameter $\theta$ to achieve a completed likelihood function, and second (M-step) to recalculate the maximum likelihood (ML) estimate of $\theta$ given the observed values $y_{obs}$ and the filled-in values of $Y_{mis} = y_{mis}$. The E-step and M-step are iterated until convergence of the estimates is achieved.

More precisely, it is the log likelihood $\ln L(\theta; y)$ of the complete-data problem that is manipulated in the E-step. As it is based partly on unobserved data, it is replaced by its conditional expectation

$$E(\ln L(\theta; Y)|y_{obs}; \theta^{(t)})$$

given the observed data $y_{obs}$ and a current fit $\theta^{(t)}$ for the unknown parameters. Thus the E-step consists of calculating this conditional expectation $E(\ln L(\theta; Y)|y_{obs}; \theta^{(t)})$. The simpler M-step computation can now be applied to this completed data and a new actual value $\theta^{(t+1)}$ for the ML estimate is computed therefrom. Now let $\theta^{(t+1)}$ be the value of $\theta$ that maximizes $E(\ln L(\theta; Y)|y_{obs}; \theta^{(t)})$. Dempster *et al.* (1977) have shown that $\theta^{(t+1)}$ then also maximizes the observed-data likelihood $L(\theta; y_{obs})$ in the sense that the observed-data likelihood of $\theta^{(t+1)}$ is at least as high as that of $\theta^{(t)}$, i.e., $L(\theta^{(t+1)}; y_{obs}) \geq L(\theta^{(t)}; y_{obs})$.

Starting from some suitable initial parameter values $\theta^{(0)}$, the E- and the M-steps are repeated until convergence, for instance, until $|\theta^{(t+1)} - \theta^{(t)}| \leq \epsilon$ holds for some fixed $\epsilon > 0$. Not all the problems are well-behaved, however, and sometimes the EM does not converge to a unique global maximum.[4]

3.5. EVIDENCE FROM A COMPARISON STUDY. In this section we present a simple simulation study to illustrate the implications of alternative imputa-

[4] For a detailed description of the EM algorithm and its properties see McLachlan and Krishnan (1997), Schafer (1997), Little and Rubin (2002), and the fundamental paper of Dempster *et al.* (1977).

tion procedures. We compare moments of a random variable (income) when applying multiple imputation (MI), simple single mean imputation (SI), single mean imputation within classes (also known as conditional mean imputation and here equivalent to a weighting procedure as shown in Section 3.2) (SI CM), and complete case analysis (CC).

Assume that a randomly drawn variable which we label age (AGE) is normally distributed with mean 40 and standard deviation 10, and another randomly drawn variable labelled income (INC) is normally distributed with mean 1500 and standard deviation 300. Because real income variables do not generally follow a normal distribution, often their log transformation log(INC) is used to achieve approximate normality. Let the correlation between age and income be 0.8, then[5]

$$(AGE, INC) \sim N\left(\begin{pmatrix} 40 \\ 1500 \end{pmatrix}, \begin{pmatrix} 10^2 & 0.8 \cdot 3000 \\ 0.8 \cdot 3000 & 300^2 \end{pmatrix}\right).$$

A sample of n = 2000 is drawn from this universe. After being generated, the AGE variable is recoded into 6 categories, 1 $\leq$ 20 years, 2 = over 20 - 30 years, ..., 6 > 60 years. First, the complete cases are analyzed, the mean income estimate, its standard error (s.e.), and the 95% confidence interval are calculated. Then different missingness mechanisms (MCAR, MAR, NMAR) are applied on income. Under MAR, income is missing with higher probability when age is higher, under NMAR, the probability that income is missing is higher the higher income is itself.

After discarding 30% of the income data, first the complete cases are analyzed, then a simple mean imputation is performed, and, finally, a proper multiple imputation procedure with $m = 5$ is used according to Rubin (1987, p. 167). The whole simulation process of creating the data, applying the missingness, performing the imputations, and analyzing the sample is repeated 1000 times. The coverage (cvg.) is counted, i.e., the number of confidence intervals out of 1000 that cover the true mean value. The average bias, the standard errors, and the usual correlation estimates between age (recoded) and income are given in Table 1.

The results in Table 1 show how precision is reduced when only the complete cases are used under MCAR, and how biased the complete case estimate (CC) gets when the missingness is MAR or NMAR.[6] The table also shows how biased a simple mean imputation is and how this bias is corrected when conditional means are imputed instead of the overall mean (cf. the means in Rows 7 and 8 and 11 and 12). However, this conditional mean imputation requires that the missingness depends on the variable

[5] For robustness checks this study was also run with lower correlation values. However, that did not change the main message. Notice that the lower the correlation the less efficient are the procedures under NMAR.

[6] For the precision compare the standard errors in Row 1 to those of the CC analyses in Rows 2, 6, and 10. For bias compare the means in Rows 2, 6 and 10.

conditioned on. The single mean imputation within classes also leads to an overestimation of the correlation between recoded AGE and INC though the simple single imputation underestimates it (see the last column of Table 1). Moreover, with single imputation the standard errors are always too small to get the nominal coverage.

Even if the missingness is MCAR, a simple mean imputation affects standard errors and correlations. Under MAR and even under NMAR, multiple imputation yields results much closer to the true values. Particularly in a NMAR scenario MI borrows strength from the correlation between age and income. Standard errors, correlation, and the nominal coverage are well reproduced by MI. Notice that confidence intervals under MI can be even narrower than confidence intervals based on complete case analysis (CC). This is especially true if the imputed sample is substantially larger than the complete case sample. Therefore, typically, the following comparisons hold for most surveys and most estimates of standard errors:

$$\text{s.e.(SI)} < \text{s.e.(truth)} < \text{s.e.(MI)} < \text{s.e.(CC)}.$$

More elaborate comparisons by simulation studies are provided, e. g., by Schafer (1997), Raghunathan and Rubin (1998), or Münnich and Rässler (2005). The latter are comparing especially GREG and Horvitz-Thompson estimators using nonresponse corrections as well as MI procedures.

| No | Missing | Proc. | Mean(INC) | Bias(INC) | S.e.(INC) | Cvg. | Cor(AGE, INC) |
|---|---|---|---|---|---|---|---|
| 1 | None | | 1500.21 | 0.21 | 6.71 | 0.96 | 0.77 |
| 2 | MCAR | CC | 1500.14 | 0.14 | 8.01 | 0.95 | 0.77 |
| 3 | MCAR | SI | 1500.14 | 0.14 | 5.61 | 0.82 | 0.64 |
| 4 | MCAR | SI CM | 1500.20 | 0.20 | 6.28 | 0.91 | 0.82 |
| 5 | MCAR | MI | 1500.24 | 0.24 | 7.34 | 0.95 | 0.77 |
| 6 | MAR | CC | 1470.35 | -29.65 | 7.98 | 0.04 | 0.77 |
| 7 | MAR | SI | 1470.35 | -29.65 | 5.58 | 0.01 | 0.63 |
| 8 | MAR | SI CM | 1499.90 | -0.10 | 6.28 | 0.88 | 0.82 |
| 9 | MAR | MI | 1499.82 | -0.18 | 7.43 | 0.93 | 0.77 |
| 10 | NMAR | CC | 1474.29 | -25.71 | 7.99 | 0.11 | 0.77 |
| 11 | NMAR | SI | 1474.29 | -25.71 | 5.59 | 0.03 | 0.64 |
| 12 | NMAR | SI CM | 1489.33 | -10.66 | 6.26 | 0.59 | 0.82 |
| 13 | NMAR | MI | 1489.30 | -10.70 | 7.36 | 0.71 | 0.77 |

TABLE 1. Results of the simulation study.

## 4. CONCLUSIONS AND RECOMMENDATIONS

Item nonresponse is a common problem in empirical analyses. Research on the determinants of nonresponse behavior yields a catalogue of relevant

factors. The evidence on German data confirms that data collection methods and respondent characteristics affect nonresponse behavior. Extant studies also confirm that different ways of dealing with item nonresponse may affect the results of empirical analyses.

We discuss the strengths and weaknesses of four commonly used approaches to deal with item nonresponse and provide a simulation study. This simulation yields that the most commonly used approach, which considers only observations without missing values, can lead to substantial biases in the estimates. The performance of single imputation procedures depends on whether there are patterns in the missingness of the data and on whether the information is missing (completely) at random. Multiple imputation procedures appear to yield the best coverage of the true value and the best reflection of existing correlation patterns.

Casewise deletion can only be an appropriate procedure if the missing data are missing completely at random. In all other cases it involves biased estimates and other procedures are preferable. Weighting is a first step to correct for nonresponse and disproportionalities. The literature suggests that multiple imputation under MAR often is quite robust against violations of the MAR assumption. Only when NMAR is a serious concern and the share of missing information is substantial it seems necessary to jointly model the data and the missingness using model-based procedures. Since missing values cannot be observed, there is no direct evidence in the data to test a MAR assumption. Therefore, it seems useful to consider alternative models and to explore the sensitivity of resulting inferences. We conclude that a multiple imputation procedure seems to be the best alternative at hand to account for missingness and to exploit all available information. In particular it generates the only format with correct standard errors allowing valid inference from standard complete case analysis.

It is recommendable that empirical researchers step beyond standard complete or available case analysis and investigate the robustness of findings by applying alternative procedures. This is aided by the fact that various single imputation techniques, such as mean imputation, conditional mean imputation, or regression imputation, are now available in commercial statistical software packages. Free programs and routines comprise the stand-alone Windows program NORM or the S-PLUS / R libraries NORM, CAT, MIX, PAN, and MICE which are all basically data augmentation algorithms. NORM uses a normal model for continuous data, CAT a log-linear model for categorical data. MIX relies on a general location model for mixed categorical and continuous data. PAN is created for panel data applying a linear mixed-effects model. Moreover, there are the free SAS-callable application IVEware as well as a STATA packet MVIS which are, like MICE, based on the very flexible sequential regression approach. The SAS procedures PROC MI with PROC MIANALYZE provide a parametric and a nonparametric regression imputation approach, as well as the multivariate normal model. Finally, there is the free Windows or Gauss version AMELIA. With increasing computational power, more and more multiple

imputation techniques are now implemented in available statistics software to create multiply-imputed datasets for further analyses.[7]

## REFERENCES

BARNARD, J., RUBIN, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* **86** 948–955.

DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39** 1–38.

DEVILLE, J. C., SÄRNDAL, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87** 376–382.

DEVILLE, J. C., SÄRNDAL, C. E., SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* **88** 1013–1020.

DILLMAN, D. A., ELTINGE, J. L. , GROVES, R. M., LITTLE, R. J. A. (2002). Survey nonresponse in design, data collection, and analysis. In *Survey Nonresponse* (R. M. Groves, D. A. Dillman, J. L. Eltinge, R. J. A. Little, eds.), 3–26. Wiley, New York.

ESSER, H. (1984). Determinanten des Interviewer- und Befragtenverhaltens: Probleme der theoretischen Erklärung und empirischen Untersuchung von Intervieweffekten. In *Allgemeine Bevölkerungsumfrage der Sozialwissenschaften* (K. Mayer, P. Schmidt, eds.), 26-71. Campus, Frankfurt.

FRICK, J. R., GRABKA, M. M. (2003). Missing income data in the German SOEP: Incidence, imputation and its impact on the income distribution. DIW Discussion Papers 376, DIW Berlin.

GELMAN, A., CARLIN, J. B. (2002). Poststratification and weighting adjustment. In *Survey Nonresponse* (R. M. Groves, D. A. Dillman, J. L. Eltinge, R. J. A. Little, eds.), 289–302. Wiley, New York.

GLYNN, R., LAIRD, N. M., RUBIN, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In *Drawing Inferences from Self-Selected Samples* (H. Wainer, ed.), 119–146. Springer, New York.

GROVES, R. M., DILLMAN, D. A., ELTINGE, J. L., LITTLE, R. J. A. (2002). *Survey Nonresponse*. Wiley, New York.

HARTLEY, H. O., HOCKING, R. R. (1971). The analysis of incomplete data. *Biometrics* **27** 783–808.

HECKMAN, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* **5** 475–492.

HORTON, N. J., LIPSITZ, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *American Statistician* **55** 244–254.

[7] For links and further details see www.multiple-imputation.com, Horton and Lipsitz (2001), or Rässler *et al.* (2003).

LEE, H., RANCOURT, E., SÄRNDAL, C. E. (2002). Variance estimation from survey data under single imputation. In *Survey Nonresponse* (R. M. Groves, D. A. Dillman, J. L. Eltinge, R. J. A. Little, eds.), 315–328. Wiley, New York.

LITTLE, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88** 125-134.

LITTLE, R. J. A., RUBIN, D. B. (1987, 2002). *Statistical analysis with missing data.* 1. and 2. ed., Wiley, Hoboken, New Jersey.

MADOW, W. G. , OLKIN, I., RUBIN, D. B. (1983). *Incomplete Data in Sample Surveys.* Academic Press, New York.

MCLACHLAN, G. J., KRISHNAN, T. (1997). *The EM Algorithm and Extensions.* Wiley, New York.

MÜNNICH, R., RÄSSLER, S. (2005). PRIMA: A new multiple imputation procedure for binary variables. *Journal of Official Statistics* (to appear).

OH, J. L., SCHEUREN, F. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys 2* (W. G. Madow, I. Olkin, D. B. Rubin, eds.), 143–184. Academic Press, New York.

RAGHUNATHAN, T. E., RUBIN, D. B. (1998). Roles for Bayesian Techniques in Survey Sampling. *Proceedings of the Silver Jubilee Meeting of the Statistical Society of Canada* 51–55.

RÄSSLER, S., RUBIN, D. B., SCHENKER, N. (2003). Imputation. In *Encyclopedia of Social Science Research Methods* (A. Bryman, M. Lewis-Beck, T. F. Liao, eds.), 477–482. Sage, Thousand Oaks.

RÄSSLER, S., SCHNELL, R. (2004). Multiple imputation for unit nonresponse versus weighting including a comparison with a nonresponse follow-up study. Diskussionspapier der Lehrstühle für Statistik 65/2004, Nürnberg.

RIPHAHN, R. T., SERFLING, O. (2002). Item non-response on income and wealth questions. IZA Discussion Paper No. 573, IZA Bonn.

RIPHAHN, R. T., SERFLING, O. (2005). Item non-response on income and wealth questions. *Empirical Economics* (to appear).

RUBIN, D. B. (1972). A non-iterative algorithm for least squares estimation of missing values in any analysis of variance design. *The Journal of the Royal Statistical Society, Series C* **21** 136–141.

RUBIN, D. B. (1974). Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association* **69** 467–474.

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., Wiley, Hoboken, New Jersey.

RUBIN, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association* **91** 473–489.

RUBIN, D. B., SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* **81** 366–374.

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman and Hall, London.

SCHRÄPLER, J. P. (2004). Respondent behavior in panel studies. A case study for income nonresponse by means of the Germany Socio-Economic Panel (SOEP). *Sociological Methods and Research* **33** 118–156.

SUDMAN, S., BRADBURN, N. M., SCHWARZ, N. (1996). *Thinking about Answers. The Application of Cognitive Processes to Survey Methodology.* Jossey Bass Publishers, San Francisco.

Susanne Rässler
Kompetenzzentrum für Empirische Methoden
IAB Institut für Arbeitsmarkt- und Berufsforschung
Regensburger Str. 104
90478 Nürnberg
susanne.raessler@iab.de

Regina T. Riphahn
Lehrstuhl für Statistik und empirische Wirtschaftsforschung
Universität Erlangen-Nürnberg
Lange Gasse 20
90403 Nürnberg
regina.riphahn@wiso.uni-erlangen.de