Survey Item Nonresponse and its Treatment

Susanne Rässler, IAB Nürnberg, Universität Erlangen-Nürnberg
and
Regina T. Riphahn, Universität Basel, IZA, DIW and CESifo
February 28, 2005

Abstract:

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 their effects using a simulation study. The paper concludes with recommendations for the applied researcher.

Keywords: Item nonresponse, imputation, weighting, survey data

JEL Code: C1, C81, C49

Correspondence to:
Regina T. Riphahn
WWZ Univ. of Basel
Postfach 517
CH - 4003 Basel, Switzerland
E-Mail: regina.riphahn@unibas.ch
Tel: 0041 - 61 - 267 3367
Fax: 0041 - 61 - 267 3351

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 1970s (e.g. Hartley and Hocking (1971), Rubin (1972, 1974), Little (1976), Kalton (1983), and Griliches (1986)).

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 weight 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 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 assess 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. Little and Rubin (2002) distinguish missing values with monotone and non-monotone patterns and discuss likelihood-based procedures derived from statistical models for the data generating and missing data mechanism.

An evaluation of the properties of the four approaches to solve the missing data problem 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 (see Little and Rubin 2002, p. 12) now distinguishes three cases: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR).

MCAR refers to data where the missing mechanism is 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 income data are said to be MCAR. Data are labeled MAR, if the missing mechanism is dependent on observed but not on unobserved variables. For example, special socio-economic groups or minorities are often disproportionately subject to missing values. If in such cases the missingness can be explained by the observed variables, the missing data are said to be MAR. 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.

The next section describes the prevalence, determinants, and effects of item non-

response using the German Socioeconomic Panel Survey (GSOEP) as an example. Section three 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 (GSOEP)

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) over all subsamples we obtain 0.4 percent item nonresponse for a subjective measure of 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 payments received and annuity payments made. In contrast, certain questions on social transfers such as child or welfare benefits received yielded nonresponse rates of below one percent.

Schräpler (2004, 2003) describes the development of item nonresponse behavior with respect to individual gross labor income. He compares the nonresponse rates of a given sample of respondents over the years and finds declining nonresponse rates which differ depending on the method of data collection and respondent characteristics. The item nonresponse behavior correlates with individual time in sample. Other studies confirm that individuals with a low propensity to continue responding to a panel survey are also less likely to disclose their income.

Since item nonresponse rates on financial questions are particularly high and because income measures are relevant for many empirical studies in the social sciences, the income variable is typically in the focus of research on determinants and effects of item nonresponse behavior.

2.2 Determinants and Effects of Item Nonresponse

The theoretical literature on item nonresponse behavior mainly applies two explanatory approaches, the cognitive and the rational choice model (see e.g. Schräpler 2004

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

for a discussion). 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 the issue must be 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). Schräpler (2003) provides a list of potential costs and benefits of survey participation. 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 data abuse and breach of privacy) as well as of the necessary effort to recall the facts desired by the questionnaire.

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. Similarly, Dillman et al. (2002) provide a classification of seven causes of item nonresponse (INR):

- 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 the respondents, even difficult answers may be given willingly. Also, interviewers' response to unanswered questions will affect 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 and should therefore not be treated identically.

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. This has been investigated for the German Socioeconomic Panel by Biewen (2001) and Frick and Grabka (2005). Biewen used the 1997 cross-section of the GSOEP and looked at whether three alternative methods of imputing missing income values differentially affect inequality measures for the earnings distribution. His results confirm that nonresponse behavior is only weakly correlated with observable characteristics and that particularly very low and very high income earners fail to report their earnings. Overall missing data and alternative imputation methods did not appear to affect the considered inequality measures here.

Frick and Grabka (2005) look at data from the 2000 and 2001 GSOEP waves and contrary to Biewen (2001) show clear effects of case-wise deletion on measures of income inequality and income mobility. Apparently item nonresponse effects can vary depending on whether gross monthly earnings are considered, as in Biewen's analysis, or whether equivalent post-government household incomes are investigated as in Frick and Grabka (2005). Given that item nonresponse may indeed bias the results of empirical analyses, 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]

- Complete case (CC) analysis considers only observations with completely recorded values for the variables of interest.
- Weighted complete case analysis in addition applies weights to compensate for bias due to missing information. This is a standard treatment for unit-nonresponse in surveys.
- Imputation entails single imputation such as hot-deck-methods and or multiple imputation (MI) as proposed by Rubin (1978, 1987).

[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).

- Model-based corrections procedures such as the expectation-maximization (EM) algorithm explicitly model both missingness and survey variables.

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. Especially in multivariate analysis, listwise deletion can reduce the available data considerably, so that they are no longer representative of the population of interest. In a dataset of 20 variables with a random five percent missing values for each variable, complete case analysis using all 20 variables will lead to an average loss of 64 percent of the sample.

Thus CC analysis can be wasteful, as informative data are dicarded 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 automatically be performed on different samples, depending on the missing data pattern, i.e. 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 (MCAR), 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 nonresponse in official statistics and survey research is weighting. Weighting is typically applied to correct for problems of unit nonresponse but also for different selection probabilities. In combination with complete case analysis procedures it can also be used to address item nonresponse problems, e.g. when using Horvitz-Thompson type estimators, see Little and Rubin (2002). A standard approach 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 n_j the number of respondents in this group. In general, the individual weight w_i of an observation i within a cell j is computed as ratio of the number of observations within a cell n_j multiplied by design weights d_j and the reciprocal

sampling fraction (N/n) to the population total of that cell (N_j):

$$w_i = \frac{n_j d_j N/n}{N_j} \tag{1}$$

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

$$\widehat{Y} = \frac{N}{n}\sum_{j=1}^{J}\frac{N_j}{n_j}\sum_{i=1}^{n_j} y_{ij} = \frac{N}{n}\sum_{j=1}^{J} N_j \overline{y}_j^{obs} = \frac{N}{n}\sum_{j=1}^{J}\left(\sum_{i=1}^{n_j} y_{ij} + (N_j - n_j)\overline{y}_j^{obs}\right). \tag{2}$$

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 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. In most cases, an iterated proportional fitting algorithm (IPF) is applied.

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 cumbersome (see Oh and Scheuren 1983), and (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 incorporate the data collector's knowledge and to use additional information not available to the analyst. Imputation of survey data and analysis of imputed data can be performed separately, which is an appealing feature. 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. While point estimates may be unbiased, confidence intervals will be too narrow, and p-values too low. 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

[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).

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; such approaches are discussed, e.g., by Lee et al. (2002).

Multiple imputation (MI), introduced by Rubin (1978) and discussed in detail in Rubin (1987), 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. Each of the imputed (and thus completed) datasets is first analyzed by standard methods; the results are then combined to produce estimates and confidence intervals that reflect the missing data uncertainty.

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 $(Y_{obs}, Y_{mis}^{(1)})$, $(Y_{obs}, Y_{mis}^{(2)})$, $\ldots$, $(Y_{obs}, Y_{mis}^{(m)})$. 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 paradigm that the MI point estimate for θ is simply the average

$$\widehat{\theta}_{MI} = \frac{1}{m} \sum_{t=1}^{m} \widehat{\theta}^{(t)}. \tag{3}$$

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

$$\widehat{var}(\widehat{\theta})_{between} = B = \frac{1}{m-1} \sum_{t=1}^{m} (\widehat{\theta}^{(t)} - \widehat{\theta}_{MI})^2, \tag{4}$$

and then the "within-imputation" variance

$$\widehat{var}(\widehat{\theta})_{within} = W = \frac{1}{m} \sum_{t=1}^{m} \widehat{var}(\widehat{\theta}^{(t)}) \,. \tag{5}$$

Finally, the estimated total variance is defined by

$$\widehat{var}(\widehat{\theta}_{MI}) = T = \widehat{var}(\widehat{\theta})_{within} + (1 + \frac{1}{m}) \widehat{var}(\widehat{\theta})_{between} = W + \frac{m+1}{m} B. \tag{6}$$

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

$$(\widehat{\theta}_{MI} - \theta)/\sqrt{T} \sim t_v \quad \text{and} \quad \widehat{\theta}_{MI} \pm t_{v,1-\alpha/2} \sqrt{T} \tag{7}$$

with degrees of freedom

$$v = (m-1)\left(1 + \frac{W}{(1+m^{-1})B}\right)^2 \tag{8}$$

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), Barnard and Rubin (1999), or Little and Rubin (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. In MCMC the desired distributions are achieved as stationary distributions of Markov chains which are based on the easier to compute complete-data distributions. There is a broad variety of available models. However, 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. However, multiple imputation explicitly formulates and evaluates these assumptions. A broad discussion of advantages and disadvantages of single and multiple imputation procedures is provided in several chapters of the book of Groves et al. (2002).

3.4 Model-based Procedures

Model-based procedures to adjust for nonresponse simultaneously model the distribution of the data Y and the response mechanism R. Selection models specify this joint distribution $f_{Y,R}(y, r; \theta, \xi)$ as

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

and have to formulate an explicit model for the distribution of the response missing-data mechanism $f_{R|Y}(r|y; \xi)$ where θ and ξ are the unknown parameters or in the Bayesian context are random variables as well. Keeping the notation simple, with missing data the likelihood of (9) is

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

Maximum-Likelihood estimates are found by maximizing this function with respect to θ and ξ. In the Bayesian context the posterior distribution is obtained by incorporating a prior distribution and performing the necessary integrations.

More often the observed-data likelihood, which is also called the likelihood ignoring the missing data mechanism, is considered:

$$L(\theta; y_{obs}) = \int f_{Y_{obs},Y_{mis}}(y_{obs}, y_{mis}; \theta) dy_{mis}. \tag{11}$$

Inferences about θ can be based on (11) rather than on the full likelihood (10) if the missing data mechanism is ignorable. Notice that ignorable Bayesian inference would add a prior distribution for θ. Rubin (1976) has shown that an ignorable missing data mechanism is given when two conditions hold. First, the parameters θ and ξ 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 θ have a couple of advantages. Usually the interest is in θ and not in the "nuisance" parameters ξ. Then the explicit modeling of the response mechanism can be cumbersome and easily misspecified. Also, often information for the joint estimation of θ and ξ is very limited. To sum up, estimates assuming MAR data turn out to be more robust in many senses, see Little and Rubin (2002).

However, in many missing data problems, even the observed-data likelihood (11) is a complicated function and explicit expressions for the ML estimate cannot be derived. In such situations, 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. Roughly speaking, 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 θ to achieve a completed likelihood function, and second (M-step) to recalculate the maximum likelihood (ML) estimate of θ given the observed values y_{obs} and the filled-in values of $Y_{mis} = y_{mis}$. Then 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 θ 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. For a detailed description of the EM algorithm and its properties the interested reader is referred to McLachlan and Krishnan (1997), Schafer (1997), Little and Rubin (2002), and the fundamental paper of Dempster et al. (1977).

3.5 Evidence from Comparison Studies

In this section we present a simple simulation study to illustrate the implications of alternative imputation procedures. We compare moments of a random variable (income) when applying different procedures to deal with its missing values: 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 complete case analysis (CC).

Assume that age (AGE) is normally distributed with mean 40 and standard deviation 10, and income (INC) is normally distributed with mean 1500 and standard deviation 300. Moreover, let the correlation between age and income be 0.8. So we let

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

A sample of n = 2000 is drawn from this universe. After being generated, the AGE variable is recoded into 6 categories, 1 <= 20 years, 2 = 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 is applied 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 width of the 95% confidence interval is reported and the usual correlation estimate between age (recoded) and income is given.

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.[4] 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 conditioned on. The single mean imputation within classes also leads to an overestimation of the correlation between recoded AGE and INC though

[4]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.

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:

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

No	Missing	Proc.	Cvg.	Mean(INC)	S.e. (INC)	CIwidth	Cor(AGE, INC)
1	None		0.96	1500.21	6.71	26.3	0.77
2	MCAR	CC	0.95	1500.14	8.01	31.44	0.77
3	MCAR	SI	0.82	1500.14	5.61	22.01	0.64
4	MCAR	SI CM	0.91	1500.20	6.28	24.63	0.82
5	MCAR	MI	0.95	1500.24	7.34	29.10	0.77
6	MAR	CC	0.04	1470.35	7.98	31.31	0.77
7	MAR	SI	0.01	1470.35	5.58	21.90	0.63
8	MAR	SI CM	0.88	1499.90	6.28	24.65	0.82
9	MAR	MI	0.93	1499.82	7.43	29.50	0.77
10	NMAR	CC	0.11	1474.29	7.99	31.34	0.77
11	NMAR	SI	0.03	1474.29	5.59	21.91	0.64
12	NMAR	SI CM	0.59	1489.33	6.26	24.56	0.82
13	NMAR	MI	0.71	1489.30	7.36	29.20	0.77

Table 1: Results of the simulation study

4 Conclusions and Recommendations

Item nonresponse is a common problem in empirical analyses. This is confirmed by the substantial incidence with which respondents refuse to provide information e.g. on financial variables. 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 an own 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 or whether the information is missing completely at random. Multiple imputation procedures appear to yield the best coverage of the true value by the estimated confidence intervals and the best reflection of existing correlation patterns in the data.

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. With increasing computational power, more and more multiple imputation techniques are also being implemented in available statistics software to create multiply-imputed datasets for further analyses.[5]

References

Barnard, J., Rubin, D.B. (1999), Small-Sample Degrees of Freedom with Multiple Imputation, Biometrika, 86, 948-955.

Biewen, Martin (2001), Item non-sponse and inequality measurement: Evidence from the German earnings distribution, Allgemeines Statistisches Archiv 85(4), 409-425.

Dempster, A.P., N.M. Laird, and D.B Rubin (1977), Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm (with discussion), Journal of the Royal Statistical Society B, 39, 1–38.

Dillman, Don A., J.L. Eltinge, R.M. Groves, and Roderick J.A. Little (2004), Survey Nonresponse in Design, Data Collection, and Analysis, in: Groves, R.M.

[5]For a more detailed description see Rässler et al. (2003).

et al. (eds.), Survey Nonresponse, Wiley Series in Probability and Statistics, 3-26, New York et al.

Esser, Hartmut (1984), Determinanten des Interviewer- und Befragtenverhaltens: Probleme der theoretischen Erklärung und empirischen Untersuchung von Interviewereffekten, in: K. Mayer, P. Schmidt (eds.), Allgemeine Bevölkerungsumfrage der Sozialwissenschaften, 26-71, Frankfurt.

Frick, Joachim R. and Markus M. Grabka (2003), Missing Income Data in the German SOEP: Incidence, Imputation and its Impact on the Income distribution, DIW Discussion Papers 376, DIW Berlin.

Frick, Joachim R. and Markus M. Grabka (2005), Item-non-response on income questions in panel surveys: incidence, imputation and the impact on inequality and mobility, forthcoming: Allgemeines Statistisches Archiv 89(1).

Griliches, Zvi (1986), Economic Data Issues, in: Z. Griliches and M.D. Intriligator (eds.), Handbook of Econometrics-Volume III, Elsevier Science Publishers, North Holland, 1465-1514.

Groves, R.M., Dillman, D.A., Eltinge, J.L., and R.J.A. Little, (eds.) (2002) Survey Nonresponse, Wiley, New York.

Hartley, H.O. and R.R. Hocking (1971), The Analysis of Incomplete Data, Biometrics 27, 783-808.

Kalton, Graham (1983), Compensating for missing survey data, Research Report Series, Institute for Social Research, University of Michigan, Ann Arbor.

Lee, H., Rancourt, E., and Särndal C.E. (2002), Variance Estimation from Survey Data under Single Imputation, Survey Nonresponse (eds. Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A.), 315-328, Wiley, New York.

Little, R.J.A. (1976), Inference about means from incomplete multivariate data, Biometrika 63, 593-604.

Little, R.J.A. and Donald B. Rubin (1987, 2002), Statistical analysis with missing data, 1. / 2. edition, John Wiley & Sons Inc., Hoboken, NJ.

McLachlan, G.J. and T. Krishnan (1997), The EM Algorithm and Extensions. Wiley, New York.

Oh, J.L. and F. Scheuren (1983), Weighting Adjustment for Unit Nonresponse, in: W.G. Madow, I. Olkin, D.B. Rubin (eds.), Incomplete Data in Sample Surveys, 2,143-184. Academic Press, New York.

Rässler, S., Rubin, D.B., and N. Schenker (2003). Imputation, Encyclopedia of Social Science Research Methods, Bryman, A., Lewis-Beck, M., Liao, T.F. (eds.), Sage, 477-482.

Rässler, S. and R. Schnell (2004). Multiple Imputation for Unit-Nonresponse versus Weighting including a comparison with a Nonresponse Follow-Up Study, Diskussionspapier 65/2004, Nürnberg.

Riphahn, Regina T. and Oliver Serfling (2002), Item non-response on income and wealth questions, IZA Discussion Paper No. 573, IZA Bonn.

Riphahn, Regina T. and Oliver Serfling (2005), Item non-response on income and wealth questions, forthcoming: Empirical Economics.

Rubin, Donald, 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 (Applied Statistics) 21, 136-141.

Rubin, Donald, B. (1974), Characterizing the estimation of parameters in incomplete-data problems, Journal of the American Statistical Association 69, 467-474.

Rubin, Donald, 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), Multiple Imputation for Nonresponse in Surveys, Wiley, New York.

Rubin, D.B., and N. Schenker (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, Jörg-Peter (2003), Gross income non-response in the German Socio-Economic Panel - Refusal or Don't Know? Schmollers Jahrbuch 123, 109-124.

Schräpler, Jörg-Peter (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(1), 118-156.

Sudman, Seymour, Norman M. Bradburn, and Norbert Schwarz (1996), Thinking about answers. The application of cognitive processes to survey methodology, Jossey Bass Publishers, San Francisco.