15 Survey Item Nonresponse and its Treatment *

Susanne Rässler[1] and Regina T. Riphahn[2]

[1] Kompetenzzentrum für Empirische Methoden, IAB Institut für Arbeitsmarkt- und Berufsforschung
susanne.raessler@iab.de
[2] Lehrstuhl für Statistik und empirische Wirtschaftsforschung, Universität Erlangen-Nürnberg
regina.riphahn@wiso.uni-erlangen.de

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.

15.1 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

*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.

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 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.