Deutsches Institut für
Wirtschaftsforschung

DIW BERLIN

Data Documentation

41

Martin Spiess

Compensating for Missing Data in the SOEP

Berlin, Mai 2009

IMPRESSUM

DIW Berlin
Deutsches Institut für Wirtschaftsforschung
Mohrenstr. 58
10117 Berlin
Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
www.diw.de

ISSN 1861-1532

DIW BERLIN

Data Documentation 41

Martin Spiess*

Compensating for Missing Data in the SOEP

Berlin, Mai 2009

* Prof. Dr. Martin Spiess, Universität Hamburg, Fachbereich Psychologie, martin.spiess@uni-hamburg.de.

Inhaltsverzeichnis

1 Missing Data

Almost all surveys that are based on voluntarily participation are affected by nonresponse. Whether or on should compensate for the missing values when analyzing an uncompletely observed data set, depends on the assumed mechanism that led to the missing data (e.g., Little and Rubin, 2002). If the missing data are missing completely at random (MCAR) or at random (MAR) then compensation is possible based on information available from the observed part of the data set and some additional weak assumptions. On the other hand, if the missing data are not missing at random (NMAR), then compensation is usually only possible under external information and strong assumptions. Hence, methods to compensate for missing values, like those adopted in the SOEP, are usually based on the MAR assumption.

Generally, a distinction is made between unit and item nonresponse. The former indicates the situation where units (e.g., households or individuals) are not observed at all, whereas item nonresponse refers to the situation where units that are otherwise willing to respond are not completely observed, i.e. do not answer all survey questions. In panel data sets, a specific type of unit-nonresponse is attrition which denotes the situation where units observed at least once drop off the survey in a later wave. Traditionally, researchers deal differently with both kinds of nonresponse: weighting is a technique usually adopted to compensate for unit nonresponse and attrition, whereas some imputation strategy is often chosen to compensate for item nonresponse. However, looking more closely at the distinction between unit and item nonresponse reveals that it is rather artificial: unit nonresponse is simply an extreme form of item nonresponse. On the other hand, up to now there is no unifying approach available to satisfactorily deal with both problems simultaneously.

2 Compensating for Unit Nonresponse in the SOEP

To compensate for unit nonresponse, the SOEP supports weighting strategies by delivering various weights together with the SOEP data. Weighting as a strategy to compensate for known sampling probabilities is standard in design-based statistics (e.g. Horvitz and Thompson, 1952; Särndal, Swensson and Wretman, 1992). With unit nonresponse, an additional selection stage, from the gross sample to the observed sample, is introduced where the “selection” probabilities are, however, unknown and must be estimated. Weighting in this context

is standard under the assumption that missingness depends on observed variables only and response probabilities can consistently be estimated, although the fact that the weights are in part based on estimates, is usually ignored. In model-based approaches, weighting as a means to compensate for differing sampling and response probabilities has only been dealt with for approximately 10 to 15 years. Up to then the problem was largely ignored.

The work of Robins and colleagues (e.g. Robins, Rotnitzky and Zhao, 1994, 1995) and by Wooldridge (2002a, 2002b, 2004) goes far beyond what has been discussed in the design-based literature. For example, in the papers of Robins and colleagues, strategies to compensate for first wave nonresponse, attrition and missing items in the context of semi-parametric estimation of panel data models are developed. Wooldridge (2002a, 2002b, 2004) discusses weighting as a strategy to compensate for different selection probabilities as well as unit nonresponse in the context of extremum estimators for cross sectional and a certain class of panel models.

Important results from this line of thinking and research imply that as many information as possible should be incorporated into the models to estimate response probabilities ("kitchen sink" approach). In fact, it can be shown that including many variables does not increase the (asymptotic) variance of the resulting estimators of a wide class of estimators of interest. Furthermore, ignoring the fact that weights are based on estimated probabilities, leads to an, usually only minor, overestimation of the standard errors and thus to conservative inferences, which is seen as being less problematic than anti-conservative inferences. Unfortunately, with standard software, it is not possible at present to use information that allows to compensate for the uncertainty in the estimated weights, even if this information were delivered with a data set. On the other hand, weighted estimation leads to larger standard errors than unweighted estimation (if both strategies are valid), weighted estimators based on estimated weights nevertheless have smaller variances than estimators based on known weights. An interesting and important discussion on when to weight (or not to weight) in model-based approaches, can be found in Wooldridge (2004).

The strategy adopted for the SOEP in 1984—almost 25 years ago (cf. Galler, 1987)—to use as much information as possible to estimate response probabilities and to base the weights on a sequence of estimated response probabilities is in line with this literature. Further, by providing various weights (design weights, the inverse of estimated attrition probabilities, denoted as "staying factors", and cross sectional weights; cf. Rendtel, 1995; Spiess, 2005) researchers

may derive their own weights according to their assumptions and needs to account for different sampling probabilities, observation probabilities possibly adjusted for various sampling probabilities as well as different versions of longitudinal weights to estimate panel data models. Although most standard software is not yet able to deal with, e.g., time varying weights, such weights are already available with the SOEP.

3 Compensation for Missing Items in the SOEP

A theoretically sound approach that became applicable through corresponding software with increasing computing power within the last years to compensate for missing items is the method of multiple imputation (e.g. Rubin, 1987, 1996). However, up to now, available techniques and statistical software does neither allow the (proper) imputation of complex surveys nor does it allow the substitution of weights by imputations (but see Spiess 2006 for the imputation of missing items and dropouts in a longitudinal analysis). First experiences with imputations are gathered in the SOEP by generating predictions for missing wealth and household income values.

As for the weighting strategy, of course, not any imputation procedure allows the generation of imputations that lead to valid inferences in the design-based or model-based analyses of interest. If the imputation procedure is proper in the sense of Rubin (1987, 1996), then the inferences of interest based on a multiply imputed data set should be valid as well, if the analysis method applied to the complete data set would lead to valid inferences. According to Rubin, a multiple imputation procedure tends to be “proper” if the imputations are (independent) draws from the corresponding predictive posterior distribution of all variables with missing values (for details see Rubin, 1987, 1996). However, in complex data sets with different types of variables (continuous, binary, truncated, ordered categorical etc.) this is complicated and may even not be practical. And in fact, available software does not allow to draw imputations from such distributions. Furthermore, although necessary in complex data sets, most of the available software packages do not allow to generate imputations under restrictions, e.g. on the range of the variables to be imputed or under other logical constraints. There does exist software (e.g. IVEware, MICE) that allows researchers to generate imputations based on univariate marginal distributions for some simple data structures (e.g. assuming that

the data are not clustered) and univariate parametric models. Adopting such an approach it may happen that a common posterior distribution for all variables with missing values does not exist. Although a few results are available that imply that this might have only negligible consequences for the inferences of substantive interest, there is still need for further research.

Other problems with software currently available are that the imputation models usually are not able to adequately deal with clustered data structures, e.g. individuals within households, within geographical units etc., different types of variables and restrictions in a minimal restrictive way (e.g. semi- or nonparametric models) at the same time. Further, the imputation models adopted are usually parametric models, and although multiple imputation can be 'self-correcting' in the sense of multiple imputations (at least) being 'confidence proper' (Rubin, 1996, 2003), there still is lack on research with respect to the consequences of misspecified imputation models.

4 Discussion and Prospective

To further improve the weighting procedure as well as to be able to generate proper multiple imputations for a complex data set like the SOEP, and thus to support users to draw valid inferences even in the presence of missing data for a wide range of situations, future projects involve the implementation of fast and stable estimation procedures for (preferably) very flexible models with arbitrary variables to be imputed. Further, since theoretical and empirical results in the statistical literature imply that as much information as possible should be used to generate weights as well as imputations, the estimation procedures must be augmented, e.g., by additional techniques to prevent the estimation of the large models to abort due to high multicollinearity. Further, much more research is needed with respect to the consequences of misspecified models to generate weights and imputations.

However, the basic decision which was taken a quarter of a century ago to generate weights for the SOEP based on a sequence of detailed attrition analyses is again justified by the latest model-based research on weighting.

5 References

Galler, Heinz P. (1987)
Zur Längsschnittgewichtung des Sozio-ökonomischen Panels. In: Hans-Jürgen Krupp and Ute Hanefeld (eds.); *Analysen 1987.* Frankfurt/M. - New York, S. 295-317.

Horvitz, D.G. and Thompson, D.J. (1952)
A Generalization of Sampling without Replacement from a Finite Universe. *Journal of the American Statistical Association,* 47, 663-685.

Little, R.J.A and Rubin, D.B. (2002)
Statistical Analysis with Missing Data (2nd ed.), New York.

Rendtel, Ulrich (1995)
Lebenslagen im Wandel: Panelausfälle und Panelrepräsentativität. Frankfurt.

Robins, James M., Rotnitzky, Andrea G. and Zhao, Lue P. (1994)
Estimation of Regression Coefficients when some Regressors are not always observed. *Journal of the American Statistical Association,* 89, 846-866.

Robins, James M., Rotnitzky, Andrea G. und Zhao, Lue P. (1995)
Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data.
Journal of the American Statistical Association 90, 106-121.

Rubin, Donald B. (1987)
Multiple Imputation for Nonresponse in Surveys. New York.

Rubin, Donald B. (1996)
Multiple Imputation After 18+ Years. *Journal of the American Statistical Association,* 91, 473-489.

Rubin, Donald B. (2003)
Discussion on Multiple Imputation. *International Statistical Review* 71, 619-625.

Särndal, Carl-E., Swensson, Bengt and Wretman, Jan (1992)
Model Assisted Survey Sampling. New York.

Spiess, Martin (2005)
Derivation of design weights: The case of the German Socio-Economic Panel (SOEP). *DIW Berlin Data Documentation 8.*

Spiess, Martin (2006)

Estimation of a Two-Equation Panel Model With Mixed Continuous and Ordered Categorical Outcomes and Missing Data. *Journal of the Royal Statistical Society: Series C* (Applied Statistics), 55, 525-538.

Wooldridge, Jeffrey M. (2002a)
Econometric Analysis of Cross Section and Panel Data. Cambridge, Massachusetts.

Wooldridge, Jeffrey M. (2002b)
Inverse Probability Weighted M-Estimators for Sample Selection, Attrition and Stratification. *Portugese Economic Journal* 1, 117-139.

Wooldridge, Jeffrey M. (2004)
Inverse probability weighted estimation for general missing data problems. CeMMAP Working Papers, CWP05/04.