MTISD 2008 – Methods, Models and Information Technologies for Decision Support Systems
Università del Salento, Lecce, 18-20 September 2008

On the Imputation of Missing Data in Surveys with Likert-Type Scales

Maurizio Carpita, Marica Manisera
University of Brescia
carpita@eco.unibs.it - manisera@eco.unibs.it

Abstract: *The aim of this paper is two-fold: to propose the imputation procedure named ABBN for replacing missing data in likert-type scales and to compare its performance with some well-known imputation methods. ABBN is a hot-deck imputation procedure which modifies the Approximate Bayesian Bootstrap method by sampling the donor in the neighbourhood of the nonrespondent. The comparison among the imputation procedures is based on a simulation study with data on job satisfaction and procedural fairness scales coming from the recent survey of workers employed in the Italian social cooperatives (ICSI^{2007}). The effects of the imputation procedures on the respondents' score and on the quality of the scales are investigated.*

Keywords: missing data, single imputation, Likert-type scales, latent traits

1. Introduction

In the social and economic surveys mental properties of individuals are often the topic of interest. Examples are abilities such as skills in mathematics, but also subjective attitudes and perceptions as customer satisfaction, quality of work (job satisfaction) and aspects of quality of life. In order to obtain a person's position on these latent traits, i.e., complex concepts not directly observed, usually a test or scale is constructed. The scale consists of several items with ordinal responses (multiple-item or Likert-type scale), each item is referred to a different and partial aspect of the latent concept. The responses given by an individual to the set of scale items give his/her total score. An easy way to obtain a person's total score is to compute the unweighted sum of the item responses.
Once chosen the statistical model to measure latent trait and constructed the scale, the quality of measurement can be affected by nonresponses (Little & Rubin, 1987): for many reasons, individuals may decide not to respond to every item in the scale or they may unintentionally skip some items. In this paper the focus in on *item nonresponse*, occurring when the responses of an individual only to some items are missing (we do not consider *unit nonresponses*). Nonresponses threatens the quality of measurement; moreover, many standard statistical techniques cannot be used with incomplete data sets or, when they can be used, become more complicated.
When faced with missing data, the researcher can (a) ignore the missing data and omit subjects with missing data from the study; (b) use procedures based on weighting; (c) use model-based procedures; (d) impute the missing data. Ignoring the missing data and, for example, constructing the measure of the latent trait by summing over the nonmissing items leads to an underestimation of the individual's score. Omitting subjects with missing data from the study (listwise and pairwise deletions) results in loss of information and leads to serious biases due to the possible systematic differences between respondents and nonrespondents (depending on the amount of the missing data; Schafer, 1997). Examples of weighting and model-based procedures are the well-known EM and data augmentation algorithms. A last strategy to deal with missing data is to use imputation procedures; in this case, the missing values in the data are filled in with plausible values to create a completed data set that can be analysed with standard techniques, developed to deal with complete data sets. The danger in imputing missing values is related to the difference between responders and nonresponders, but it is a popular and widespread technique (Huisman, 2000).
When dealing with multiple-item scales, the missing data problem has well defined characteristics: firstly, because the items together measure one latent trait, there exist a relationship between them and the imputation technique should be able to reproduce in the completed data set the same

relationship. Secondly, the amount of missing data should be considered with regard to the entire data set: even with small percentages of missings per item, the issue can be problematic when looking at the entire data set. For example, the listwise procedure lead to a deletion of all subjects having one item with missing values, regardless of whether other items have complete data or not. Moreover, certain response categories often have to be rated as missings; these include any "don't know" categories and, in some cases, the mid-point "neutral" responses.
In the treatment of nonresponses, it is important to understand the nature of the missing data mechanism. It is *ignorable* when the differences between respondents and nonrespondents are not systematic (Missing Completely At Random or MCAR), or they are systematic but only depending on observed variables and the parameters of the missing data mechanism are not related to the parameters of the model for the data (Missing At Random or MAR). When the missing data mechanism is *nonignorable* (missing Not at Random or NR), a model for the missingness process should be used; in this case, covariate information can be useful and analysis under the MAR assumption can lead to reasonable results (Huisman, 2000).
Different imputation procedures have been proposed in the literature to face missingness in Likert-type scales (e.g., Huisman, 2000; van Ginkel *et al*., 2007); some methods are the random imputation, the item mean substitution, the person mean substitution, the corrected item mean substitution, the item correlation substitution and the two-way imputation.
Other imputation procedures are based on the so-called *hot-deck imputation* (Rubin, 1987), which matches to each nonrespondent another respondent (the donor) who resembles the nonrespondent on variables that are observed for both. The donor donates its observed scores to the missing scores of the nonrespondent. There are many hot-deck procedures, differing with respect to the definition of the donor case: for example, it can be the complete case closer to the incomplete case (*hot-deck nearest neighbour*), deterministically chosen on the basis of a distance function or randomly chosen from the set of the complete cases which are closer to the incomplete case.
In this paper, we propose a hot-deck imputation procedure, described in the following section.

2. Hot-deck imputation of missing item responses: the ABBN method

We propose an imputation procedure to face the problem of missing data in surveys with likert-type scales obtained by modifying the *Approximate Bayesian Bootstrap* (ABB; Rubin, 1987). ABB is a hot-deck procedure that imputes missing data by sampling from the complete data. Sampling is stratified by variables that predict whether the data are missing. Using a model-based measure, such as the propensity score obtained from a logistic regression, some strata are created (by categorising the fitted probabilities). Imputation is then performed within strata independently.
The modified ABB imputation method we propose can be called A*pproximate* B*ayesian* B*ootstrap* on N*eighbourhood* (ABBN) because it adds to ABB the selection of the donor case based on the neighbourhood of the nonrespondent (who has at least one missing value over the items). Subjects are stratified in G groups with equal size on the basis of the propensity score, obtained by fitting a logistic regression with response variable given by the status respondent/nonrespondent and predictors given by variables supposed to be related to the missingness (e.g., sex and age of the subject). Each group g, including (complete and incomplete) subjects with the same level of propensity score, can be represented by a block as in Figure 1. For each nonrespondent (pattern with missings), 25% of subjects having the most similar response pattern (the neighbours) are selected from the complete data of the same group; similarity is measured by the Gower's index. From these neighbours, a bootstrap sample (with replacement) is drawn and finally from this sample the donor case is drawn. It is evident that the donor case results from the resampled (ABB stage) set of the neighbour respondents (N stage) of the nonrespondent. The ABBN method requires to collect covariates that may predict missingness. Findings indicate that covariates related to missingness are gender, age, education, income, occupation (Huisman, 1998).

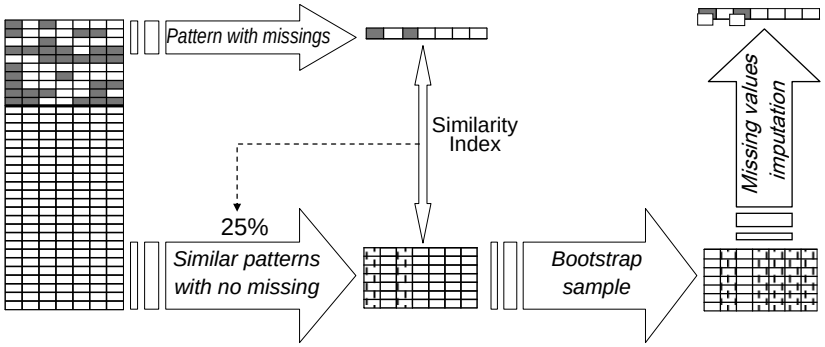

Figure 1: Diagram for the ABBN method (grey cells represent missing values)

Some authors suggest to replace the entire pattern of the nonrespondent with the donor data, in order to preserve the inter-relationships of the variables in the dataset (Allison, 2000). However, this lead to the deletion of some information when the nonrespondent did answer to at least one of the items. For this reason, in our proposal only the missing values of the nonrespondent are replaced, while its nonmissing values are kept into the data set. Obviously, the ABBN method can also be used in the context of multiple imputation.

3. Simulation design

To investigate the performance of the imputation techniques, we used an experiment with four factors : data, sample size, missing data mechanism, and proportion of missing values. With these factors, incomplete data sets are created and then imputed with the procedures described in sections 1 and 2.

To conduct the study, we considered real data rather than simulated ones, because they serve the purpose better than simulations based on data generated with some theoretical distributions (Huisman, 2000). The data used come from the survey on the Italian social cooperatives called ICSI^{2007} (Carpita, 2007), investigating the quality of work in the nonprofit sector. It involved 4,134 workers employed in 320 organizations. The two data sets used in this study refer to the scales of job satisfaction (JS) and procedural fairness (PF) studied by Carpita & Golia (2008). The two scales differ on length, number of response options, Cronbach's alpha, and average inter-item correlation. From the original JS and PF data sets, a fixed number (n=500,1000,2000) of complete cases are randomly drawn. Each of the six data matrices consists of responses of n subjects to k items $\mathbf{X}=[x_{vi}]$ (v=1,..,n and i=1,…,k) and covariates $\mathbf{Z}=[z_{vh}]$ with h=1,2 indicating gender and age. Missing values were generated in the data with four different mechanisms. A missing data indicator matrix M=[m_{vi}] is constructed, with m_{vi}=1 indicating a missing response of subject v on item i and m_{vi}=0 indicating an observed response of subject v on item i.

We generate missingness using the following four schemes.

1. MCAR: a fixed proportion p of entries is randomly deleted.
2. MAR: the probability of response of person v on item i depends on the covariates z_{vh}, that is

$$P(M_{vi}=0 \mid z_{v1}, z_{v2}) = \frac{\exp(\alpha_0 + \gamma_1 \cdot z_{v1} + \gamma_2 \cdot z_{v2})}{1+\exp(\alpha_0 + \gamma_1 \cdot z_{v1} + \gamma_2 \cdot z_{v2})} \quad (1)$$

3. NRX: the probability of response of person v on item i depends on the mean person score r_v and the mean item score $\bar{x}_{\bullet i}$, that is

$$P(M_{vi}=0 \mid r_v, \bar{x}_{\bullet i}) = \frac{\exp(\alpha_0 + \alpha_1 \cdot r_v + \delta \cdot \bar{x}_{\bullet i})}{1+\exp(\alpha_0 + \alpha_1 \cdot r_v + \delta \cdot \bar{x}_{\bullet i})} \quad (2)$$

4. NRXZ: the probability of response of person v on item i depends on *scale score* r_v, the mean item score $\bar{x}_{\bullet i}$, and two covariates z_{vh}, that is

$$P(M_{vi}=0 \mid r_v, \bar{x}_{\bullet i}, z_{1v}, z_{2v}) = \frac{\exp(\alpha_0 + \alpha_1 \cdot r_v + \delta \cdot \bar{x}_{\bullet i} + \gamma_1 \cdot z_{1v} + \gamma_2 \cdot z_{2v})}{1+\exp(\alpha_0 + \alpha_1 \cdot r_v + \delta \cdot \bar{x}_{\bullet i} + \gamma_1 \cdot z_{1v} + \gamma_2 \cdot z_{2v})} \quad (3)$$

The parameters in (1), (2) and (3) are fixed: α_1 and δ equal 1 (subjects with low r_v and for items with low $\bar{x}_{\bullet i}$ the probability of missing answer is large); γ_1 and γ_2 are fixed at −1 (women and older respondents are more likely to show missings). An observation is classified missing if $P(M_{vi}=0 \mid r_v, \bar{x}_{\bullet i}, z_{1v}, z_{2v}) \le u_{vi}$, where u_{vi} is a random number drawn from a uniform random variable over the interval [0,1].
Three different levels of p were used: 0.1, 0.2 and 0.3. They were achieved by choosing different values of parameter α_0 in (1), (2) and (3).

4. Comparison among imputation techniques

In order to evaluate the performance of the imputation techniques, the whole procedure is replicated 1000 times. Then, the imputed data are compared with the original complete data. Firstly, the impact of the imputation on the quality of the scale is judged. To evaluate whether the imputation procedures are able to preserve the relationships among the item, some item analysis indices as Cronbach's alpha are computed with reference to the original complete data and the imputed data. Secondly, the effect of the imputation procedures on the respondents' score is investigated, by summarizing discrepancies between the scores after imputation and the scores computed on the original complete data.
The study may show to what extent the proportion of missing data, the missing data mechanism, the characteristics of the scale, and the sample size affect the performance of any imputation procedure, included the one proposed in this paper. Moreover, the study may allow to select the best method(s) of imputation with respect to the quality of measurement. The selected method(s) should be recommended when the goal of the study is to measure a latent trait starting from multiple item forming a scale and respondents have no answered all the items.

Bibliography

Allison P. (2000), Multiple imputation for missing data: a cautionary tale, Sociological Methods and Research, 28: 301-309.
Carpita M. (2007), L'Indagine sulle Cooperative Sociali Italiane 2007 (ICSI 2007): organizzazione della ricerca e caratteristiche del campione, Impresa Sociale, 3: 33-52.
Carpita M., Golia S. (2008), Measuring the subjective quality of work in the social enterprise, Paper presented at the Colloquio Scientifico Annuale sull'Impresa Sociale, II Ed., Bari, May 23-24.
Huisman M. (1998), Missing data in behavioral sciences research: investigation of a collection of data sets, Kwantitatieve Methoden, 57: 69-93.
Huisman M. (2000), Imputation of missing item responses: Some simple techniques, Quality & Quantity, 34: 331-351.
Little R.J.A. & Rubin D.B. (1987), Statistical analysis with missing data, Wiley, New York.
Rubin D.B. (1987), Multiple imputation for nonresponse in surveys, Wiley, New York.
Schafer J.L. (1997), Analysis of incomplete multivariate data, Chapman & Hall, London.
van Ginkel J.R., van der Ark L.A., Sijtsma K. (2007), Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results, Multivariate Behavioral Research, 42: 387-414.

