Electronic Journal of Applied Statistical Analysis
EJASA (2012), Electron. J. App. Stat. Anal., Vol. 5, Issue 3, 431 – 437
e-ISSN 2070-5948, DOI 10.1285/i20705948v5n3p431

MISSING DATA AND PARAMETERS ESTIMATES IN MULTIDIMENSIONAL ITEM RESPONSE MODELS

Federico Andreis, Pier Alda Ferrari*

[(1)]*Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, Italy*

Received 27 July 2012; Accepted 12 November 2012
Available online 16 November 2012

Abstract: *Statistical analyses of data based on surveys usually face the problem of missing data. However, some statistical methods require a complete data matrix to be applicable, hence the need to cope with such missingness. Literature on imputation abounds with contributions concerning quantitative responses, but seems to be poor with respect to the handling of categorical data. The present work aims at evaluating the impact of different imputation methods on multidimensional IRT models estimation for dichotomous data.*

Keywords: *Imputation methods, dichotomous response, simulation study, mcmc.*

1. Introduction

Missing data are widely present in surveys, and their treatment represents a relevant problem in statistical analyses. In fact, an improper handling of missing values may lead to incorrect inference and some statistical methods require the knowledge of a complete data matrix to be used. Several techniques exist to deal with such a problem when quantitative variables are involved [7], whereas few procedures are available for categorical data [3]. In the framework of multidimensional Item Response Theory (MIRT) models a researcher could choose among three alternatives: delete the incomplete units from the dataset and consider only those with complete observations (listwise deletion), ignore missingness and use for each unit only the observed values (passive approach), or consider all the units and variables imputing the missing values. Deletion might lead to a substantial loss of information in the presence of many missing values and to biased estimates. The passive approach, as is well known [7], leads to valid inference when simple models, such as unidimensional IRT models, are estimated, since sufficient statistics for the parameters of interest exist. When more complex models are considered, this

* E-mail: federico.andreis@unimi.it

approach might not be suitable and imputation seems in these cases to provide a preferable solution. Moreover, in some situations a complete data matrix could be required and imputation, thus, needed.
The aim of this paper is to compare some imputation algorithms for handling missing values in dichotomous responses. We will discuss the performance of imputation methods with respect to the problem of estimating item parameters of a MIRT model, focusing on the two-dimensional 2-Parameters Logistic model (M2PL) [10].
Analyses are carried out entirely using free statistical softwares, such as WinBUGS [8] and R [9], and the choice of imputation methods in comparison is also driven by the availability of such techniques in the aforementioned programming environments.

2. Methods

2.1 Missingness mechanisms and missing data handling methods

Three missingness mechanisms are usually considered in the literature: MCAR (also called *uniform non-response*), MAR (or *uniform non-response within classes*) and MNAR (or *non-ignorable missingness*). Data are Missing Completely At Random when the probability of having a missing value does depend on neither observed nor unobserved variables. Data are Missing At Random when, conditional to the observed data, the probability of an observation being missing does not depend on unobserved variables. When neither MCAR nor MAR holds, data are called Missing Not At Random. All these three mechanisms have been adopted in our simulation study.
As for what concerns the handling of missing data, we consider four methods already implemented in R packages or of very simple and straightforward coding: CD, FI, MF and MICE. **CD** (Complete Data) is the simplest and most renowned method. It consists in the deletion from the data matrix of the individuals (rows) with missing values. A simple logical constraint on the data matrix is sufficient to apply such method in R. **FI** (Forward Imputation) is a sequential model-based imputation method employing Non Linear Principal Component Analysis (NLPCA) and addressed to categorical data. This method is implemented in the R package *ForImp* and described in [3]. **MF** (Miss Forest) is a non-parametric missing value imputation method that makes use of random forests. This method is implemented in the R package *missForest* and described in [11]. **MICE** (Multivariate Imputation by Chained Equations) is a multiple imputation method that makes use of specification of conditional distributions for the variables in the dataset, is implemented in the R package *mice* and described in [2].

2.2 Model and comparison among methods

MIRT models, developed in the area of ability assessment and psychometrics and then extended to the evaluation of attitudes (e.g., of customers towards products and services [6]), express the probability of the response of a person to an item in a questionnaire as depending on persons and items parameters, whose interpretation depends on the context. We consider here one of the simplest MIRT models, the two-dimensional M2PL, characterized by three items and two persons parameters. This model can be expressed as:

$$logit\left[P\left(X_{ij}=1\right)\right]=\sum_{l=1}^{2}a_{jl}\theta_{il}+d_j$$

where, in the context of Customer Satisfaction, X_{ij} is the response of the $i-$th individual to the $j-$th item ($X_{ij}=0$ if the customer is unsatisfied and $X_{ij}=1$ if satisfied) and, following the interpretation in [1], d_j is the quality parameter while a_{jl} and θ_{il} are, respectively, the item relevance and the personal satisfaction parameters on the $l-$th latent dimension (l= 1,2). The larger the item relevance, personal satisfaction, or item quality, the higher the probability of a $X_{ij}=1$ (satisfied) response. Our focus is on the items parameters d_j, a_{j1} and a_{j2}.
In order to compare the different treatments of missing data, the aforementioned model is fitted on a matrix of full data and parameters' values are stored, to serve as 'population parameters'. On the basis of these initial values some of the items are chosen, presenting various combinations of quality and relevance levels, so to be representative of 'typical' items a researcher might encounter [4]. In the simulation study these items are populated with missing values under different scenarios. The performance of different methods of missing data handling under different experimental conditions is then evaluated through comparison of the absolute bias in parameters' estimates.

3. Simulation Study

For the analysis we refer to real data from a customer satisfaction survey, presented in [6]. We have selected 10 items (*q50*, *q51*, *q55*, *q56*, *q18*, *q19*, *q22*, *q23*, *q32*, *q36* hereby labelled from 1 to 10 respectively) and, according to our aim, dichotomized, the responses (originally on a 1-5 Likert scale) to 0 (1 to 3) or 1 (4 or 5). From these data a complete subset was taken to serve as population matrix, for a total of 113 individuals and 10 items with no missing values. Starting from this matrix of dichotomous data, the two-dimensional M2PL model was fitted through MCMC techniques implemented in WinBUGS *14*, called from R *2.15* through the package *R2WinBUGS*. Suitable prior distributions and conditions were chosen to ensure model identifiability. Independent standard normal distributions were assumed as priors for the 113 person parameters and the 10 quality parameters, while independent log-normal distributions (mean-log = 0, sd-log = $\log^{-1} 2$) were adopted for the 2x10 relevance parameters, one of them being fixed to 1, to fulfil identification requirements [5]. A single chain was run for 3000 iterations, after a burn-in period of 1500, albeit showing stationarity (investigated also through the use of multiple chains) after few hundreds of runs. An over-relaxed form of MCMC sampling available in WinBUGS was employed to improve convergence.
Four of the ten items, characterized by various combinations of parameters' values, were then chosen: item 3 ($d=1.24, a_1=0.32, a_2=1.63$), item 5 ($d=0.73, a_1=4.07, a_2=0.56$), item 6 ($d=2.11, a_1=1.79, a_2=0.72$) and item 10 ($d=-0.88, a_1=0.50, a_2=0.54$).
The selected items (columns) were populated with missing values, varying both percentage (5%, 10% and 30%) and missingness mechanism (MCAR, MAR, MNAR). Under each of the nine possible combinations of different proportion of missingness and missing mechanism, the dataset was handled with the aforementioned methods (CF, FI, MF and MICE) in order to obtain a complete data matrix, on which the M2PL's parameters were estimated. These values were

then compared to the population ones. A total of 100 replications for each scenario was carried out.
In order to generate the missing data according to the three different mechanisms, we proceeded as follows at each simulative iteration:
MCAR: a uniform sample of fixed proportion was drawn from the four columns, and corresponding values deleted. The R function *prodNA* provides an efficient way to do that.
MAR: each individual was assigned to one of 3 classes (A, B, C) according to his/her score (0-1, 2-3 and 4-6, respectively) on all the items except the 4 chosen to contain the missing values. The probability of a missing response for a person was assigned on target items according to their class as follows: $p_A = 2p_B = 4p_C$. Inclusion probabilities were calculated, and suitable procedures implemented in the *sampling* package in R, e.g. *UPsampford* or *UPtille,* and used to draw a sample of fixed size from the four columns. Sampled observations were then deleted.
MNAR: a probability of missing response was assigned to each person (row) based on the population matrix response on the 4 target items; individuals who originally answered 0 on an item have a probability of generating a missing value 3 times higher than those that answered 1. Also in this case appropriate inclusion probabilities were calculated, and suitable sampling functions used to obtain a sample of fixed size from the four columns. Sampled observations were then deleted.
Problems were met during the simulation study, due to the numerical stability of estimation algorithms, specific features of the considered imputation methods and the efficiency of the sampling method used to generate the missing values. In more detail, we solved the following problems:

- Numerical stability of WinBUGS estimation. Extreme values generated along the chain sometimes may yield extremely close-to-zero probability values, whose logarithm is in turn approximated by the program to minus infinity. This led to invalid computations that stopped the estimation process. To fix this, we imposed restrictions on the prior distributions.
- FI does not work if only one level (0 or 1) and missing values are observed in a row. We solved the problem through a stochastic pre-imputation for those subjects with 0 or maximum score after the missing value generation step.
- MICE sometimes left one or more of the columns not imputed under the MNAR scenarios, yielding a matrix that still contained missing values. The package manual reckons that, in such a situation, one should model ad-hoc the missing mechanism each time, but for a simulation study this is not a feasible option. As a workaround, we filled in the leftover missing values stochastically (as for the FI problem) after the imputation.
- The *UPsampford* function from the *sampling* package, initially employed, implements a sampling scheme that is too slow when the missing proportion is high; we then switched to a different sampling scheme, specifically the one implemented in the *UPtille* function in the same package.

All computations except for the estimation were carried out using R environment exclusively. The code written for these analyses is available upon request to the corresponding author.

4. Results and Discussion

The four methods for handling missing data we discussed, all implemented in the R environment and suitable for dichotomous responses in the framework of MIRT models, have been compared and their performances evaluated with regards to bias of two-dimensional M2PL model parameters' estimates under different experimental conditions.
For the sake of brevity, only the main results are now presented and a part of the plots that have been produced are shown as an example.
We first discuss the estimates and their biases concerning parameter *d* under all scenarios; the corresponding boxplots are shown in **Figure 1**. All the imputation methods we considered lead to similar results in terms of recovery of the 'true' values. We observe that application of these procedures eventually resulted in consistently overestimating the population *d* under every combination of missingness mechanism and percentage, except for the 30% one; the same is true for CD. In this latter case, the estimates seem to be pulled towards under-estimation (particularly evident under the MNAR approach) and this is probably connected to the chosen pattern of missingness. MF almost always leads to slightly worse over-estimations under MAR, whereas seems to behave better under MNAR.

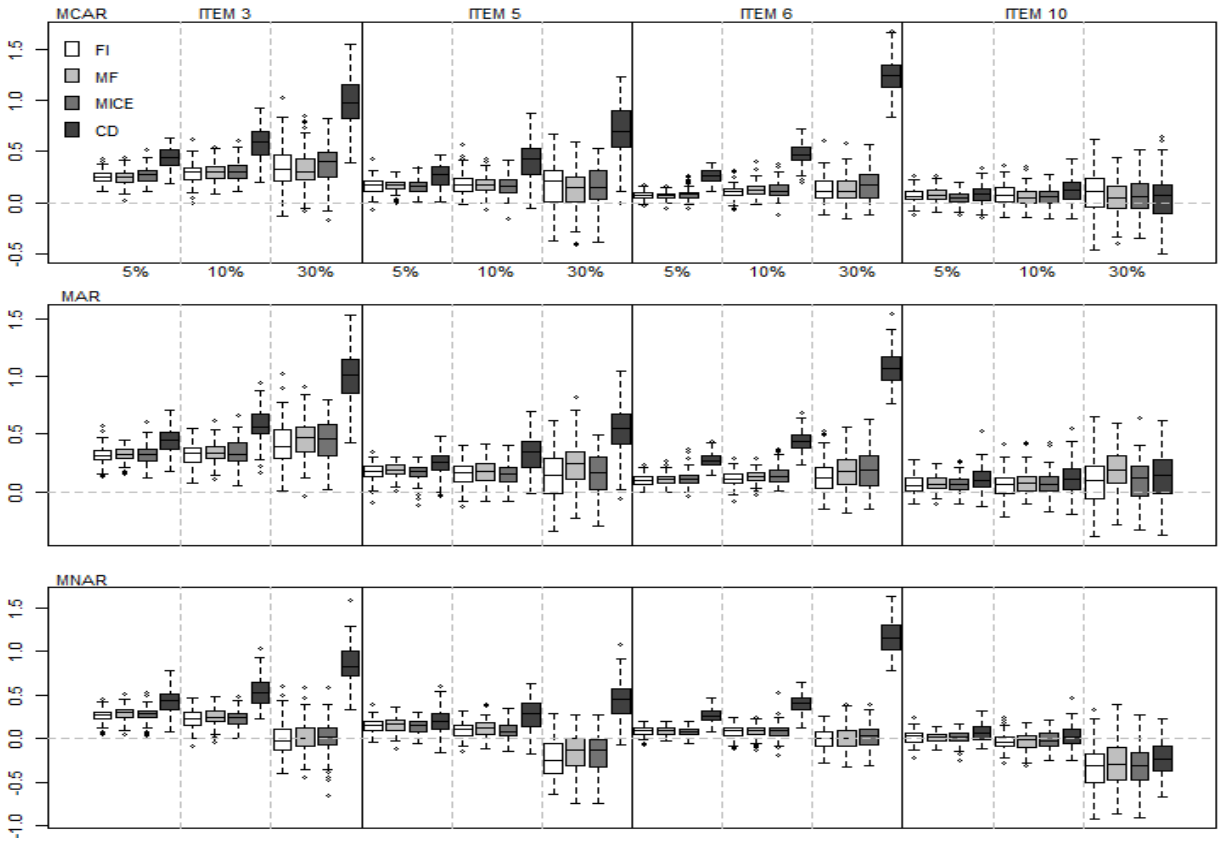

Figure 1. Bias boxplots for *d* under the three mechanisms, varying missingness percentage and item.

As for what concerns a_1 and a_2, whose boxplots are omitted for brevity, all the imputation methods consistently lead to under-estimates of the true values with one item-dependent

exception on item 10, where using MF always leads to over-estimation. The CD technique leads for the relevance parameters to satisfactory results in almost all situations, yielding moderate variability in bias distribution and an absolute bias comparable with those provided by the three imputation methods for every item.
On the base of these simulations, the bias definitely seems to be item-dependent with regards to both versus and magnitude, we thus suspect it could be to some extent forecast on the basis of item characteristics. The complete data technique performs the worst as for what concerns the *d* parameter; in situation of heavy missingness, though, it out-performs the imputation methods we considered as for what concerns a_1 and a_2. It could be that such parameters are particularly sensitive to variability in responses and, being the fact that the variable is dichotomous, in the presence of a large number of missing values adding 'guessed' values can inflate variability too much, for estimates to be reliable.
Overall, FI and MF would seem to be the techniques-of-choice in the presence of heavy missingness, in terms of absolute bias in recovering the true *d* parameter value, whereas MF and MICE should be preferred were the interest only on a_1 and a_2. We point out, however, that MICE implements the possibility to model the missingness mechanism, and this could bring some benefits, provided the model choice is correct.
These are the findings of a first analysis on a real dataset, and as such, should be read with caution. Further and more extensive simulation studies have to be carried out before being able to give more general indications, also from a practical perspective. Specifically, the role and the magnitude of items parameters in the population, which seem to be relevant factors, should be deeply investigated, e.g. considering the possibility of simulating the original data matrix by fixing the values of *d*, a_1 and a_2 to explore a larger set of combinations and scenarios and offer a wider analysis of the performance of these missing data treatment methods.

Acknowledgements

The authors acknowledge financial support from the European Social Fund Grant (FSE), Regione Lombardia.

References

[1]. Andreis, F. and Ferrari, P.A. (2012). Multidimensional extensions of IRT models and their application to customer satisfaction evaluation, in *MMLV2012*, Napoli, 17-19 May 2012, 5-8.
[2]. van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M. and Rubin, D.B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049-1064.
[3]. Ferrari, P.A., Annoni, P., Barbiero, A. and Manzi, G. (2011). An imputation method for categorical variables with application to nonlinear principal component analysis. *Computational Statistics and Data Analysis*, 55, 2410–2420.
[4]. Finch, H. (2008). Estimation of Item Response Theory Parameters in the Presence of Missing Data. *Journal of Educational Measurement*, 45: 225–245.

[5]. Jackman, S. (2001). Multidimensional analysis of roll call data via Bayesian simulation: identification, estimation, inference and model checking. *Political Analysis*, 9(3): 227-241.
[6]. Kennett, R.S. and Salini, S. (2012). *Modern analysis of customer surveys: with applications using R*. New York: John Wiley.
[7]. Little, R.J.A and Rubin, D.B. (1987). *Statistical analysis with missing data.* New York: John Wiley.
[8]. Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325-337.
[9]. R Development Core Team (2011). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
[10]. Reckase, M.D. (2009). *Multidimensional Item Response Theory*. Springer, New York.
[11]. Stekhoven, D.J. and Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, vol. 28, no. 1, 112–118.