
Statistical Modeling of Educational Data with Missingness Problems

Damazo T. Kadengye, Wim Van den Noortgate and Eva Ceulemans

Statistical analysis with missing data problems has been a major focus of research in the last few decades, especially in the field of medicine and in longitudinal studies. Missing response data are very common in clinical based longitudinal studies, for instance because of drop outs due to serious side effects, lack of recent contact information for follow-up, subjects being too sick to come for treatment, and so forth. In this kind of studies, investigators are typically interested in the “evolutions” of subjects, or in how subjects respond to treatment over time.

Our research interest is in the treatment of missing data in educational sciences, especially in the field of item-based web-learning, where data such as responses are logged when a person uses an electronic learning environment. In such an environment users normally make relatively small and independent exercises out of so many available items. Since the number of items is often relatively large and users can navigate freely through the items, datasets often include a large number of missing values. For such item-based data, the use of Item Response Theory (IRT) offers some new perspectives, for instance adapting automatically the learning environment to the user, by rendering items from which the difficulty match with the ability of the user. However, the high amount of missing data poses difficulties for analyzing the data, for instance using IRT to estimate the item difficulties and user abilities, or using cluster analysis techniques to find groups of similar persons or similar items.

Traditionally, missing data problems are approached by means of mean imputation, listwise deletion or pairwise deletion – also called available case analysis. Peugh and Enders (2004) carried out a methodological review of educational research journals and found that 96% of the 160 studies with missing data used these traditional methods. Nevertheless, statistical researchers do not recommend use of these traditional methods for parameter estimation in missing data situations, for instance since such methods tend to bias the estimates of means, variances and correlations (Baraldi & Enders, 2010; Little and Rubin, 2002, Wilkinson & Task Force on Statistical Inference, 1999).

When modeling data with missingness problems, attention needs to be paid to the mechanisms that generate this missingness. Little and Rubin (2002) have given an extensive explanation of the mechanisms that generate missing data which include data missing completely at random (MCAR), data missing at random (MAR) and data not missing at random (NMAR). Consequently, Multiple Imputation (MI) and Maximum Likelihood (ML)

Damazo T. Kadengye
Faculty of Psychology and Educational Sciences
Katholieke Universiteit Leuven Campus Kortrijk
E-mail: Trevor.Kadengye@kuleuven-kortrijk.be

Wim Van den Noortgate
Faculty of Psychology and Educational Sciences
Katholieke Universiteit Leuven Campus Kortrijk
E-mail: Wim.VandenNoortgate@kuleuven-kortrijk.be

Eva Ceulemans
Centre for Methodology of Educational Research
Katholieke Universiteit Leuven
E-mail: Eva.Ceulemans@ped.kuleuven.be

estimation have been described as “state of art” (Schafer and Graham, 2002) or “Modern” missing data techniques (Peugh and Enders, 2004, Baraldi and Enders, 2010).

IRT analyses often use ML estimation to estimate parameters of interest. Likelihood inference always gives unbiased estimates in complete data sets. But for situations where one is confronted with missing data either in the response, in the covariates or in both, direct likelihood methods like complete case analysis, available case analysis, and weighting methods will not give efficient estimates, especially when the missing mechanism is non-ignorable (see for example Molenberghs and Verbeke, 2005, Little and Rubin, 2002, Peugh and Enders, 2004, Baraldi and Enders, 2010). Multiple Imputation and the Expectation Maximization (EM) algorithm methods of handling missing data problems have been discussed extensively by Molenberghs and Verbeke (2005) who compared both approaches for different longitudinal examples. Multiple Imputation requires the missing mechanism to be MAR although Thijs *et al.* (2002) have used it in an NMAR setting. Finch (2008) has compared MI and EM under MAR and NMAR and has suggested that EM results in greater bias for parameter estimates. In the case of Finch, imputed values were rounded off to their nearest whole numbers and he strategically increased the probability of having missing values for subjects with low abilities. However, this may not always be the case. Also note that MI requires the multivariate normality assumption and there is potential bias when rounding values sampled from a multivariate normal distribution to suit binary response (Horton, Lipsitz & Parzen, 2003). Some researchers have however used EM algorithm to impute polytomous categorical data (Enders, 2004, Bernaards & Sitjsma, 1999) with no major challenges. It is however important to note here, just like De Boeck and Wilson (2004) noted, that no modeling approach, whether for MAR or for NMAR can fully compensate for the loss of information that occurs due to incompleteness of the data. The aim of taking into account the missingness mechanism is to be able to produce estimates that are more efficient.

A further complication in our datasets is that discrete repeated measurements of items are nested within groups of items there by inducing correlation among measurements. Clusters are known to exist in educational research studies. Such may include students being clustered in the same school or class. Also, items are known to be grouped under such themes like chapters, topics and courses. Clusters introduce hierarchies for persons as well as items in the data sets. As a result, simple IRT models are rendered ineffective to provide plausible parameter estimates because of the complex dependence structure brought about by the correlated nature of the data. Sometimes, clusters can be looked at as being randomly chosen from a population of clusters, requiring introducing random effects into our model (Agresti, 2002). In order to get efficient parameter estimates, these groupings need to be taken into account during statistical modeling since, say, individuals from the same school may not be independent.

Another important feature that needs to be taken into account during analysis is the type of the outcome such that instruments that exploit the nature of the data can be applied (Molenberghs and Verbeke in De Boeck and Wilson, 2004). Generalized Linear Models (GLMs) have for instance been proposed to handle data with categorical responses (Agresti, 2002) and extensions of the same to situations with missingness problems have been made in statistical literature.

Ibrahim, Chen and Lipsitz (2001) have looked at missing responses in generalized linear mixed models when the missing data mechanism for the outcome variable is non-ignorable but assumed that all covariates are observed, as well as, missing covariates in generalized linear models assuming that the response vector was completely observed (Ibrahim, Lipsitz & Chen, 1999). Stubbendick and Ibrahim (2003) employed maximum likelihood methods to non-ignorable missingness of both the response and covariates in normal random effects models and later modified the method to suit discrete longitudinal data (Stubbendick and Ibrahim 2006). Wu and Wu (2007) have also employed generalized linear mixed models for data with informative dropouts and missing covariates. In all these cases, parameter estimation was done via an extension of the EM algorithm although Finch (2008) notes that EM approaches rely heavily on the assumption of multivariate normality, which does not apply to dichotomous item responses.

We shall evaluate the methods highlighted in the previous paragraphs in the case of highly nested educational dichotomous response data with a substantial part of missingness. We shall first search the literature of the current methods being employed to handle missing data in educational item response data sets. This will enable

us to get some ideas and suggestions which will be the driving motivation of our research. Secondly, we shall evaluate the performance of traditional methods versus modern methods for educational item response data structures where longitudinality is not an issue and with high levels of missingness as well as being highly nested. We intend to propose adaptations of those methods that may break down in our situation in order to fit educational data settings. How to overcome such challenges like random effects being highly dimensional in IRT models will be taken into account. We intend to generate artificial hierarchical item response datasets and to apply and compare the different methods highlighted above; the methods will also be illustrated with a real life data set.

The second area of concern for us is the presence of latent clusters in item-based web-learning. Possible hidden clusters are likely to exist in any data set which dictate, for instance, the behavior of learners towards a certain group of items. Several clustering methods have been proposed in the literature and these can be grouped under three major categories i.e. hierarchical clustering methods, non-hierarchical clustering methods and clustering based on statistical methods (Everitt, 1993). The goal of clustering is to reduce the amount of data by categorizing or grouping similar data items together. Chang and Yang (2009) have for example applied a K-means Clustering method to cluster learner's ability in web based learning. Performance of such clustering methods when dealing with missing data in item-based web-learning environments will be another area of interest in our research. With these insights in mind, we believe that the successful completion of our research will set a trend of analyzing item-based data from web-based learning environments.

The aim of this proposed presentation is to discuss our research plans. We do not intend to present or discuss any results at this stage.

References

1. Agresti, A. (2002). *Categorical Data Analysis*. New Jersey: John Wiley and Sons.
2. Baraldi, A.N., and Enders, C.K. (2010). An introduction to modern missing data analyses. *Journal of School of Psychology*, **48**, 5 – 37.
3. Chang, W-C., and Yang, H-C. (2009). Applying IRT to estimate learning ability and K-means clustering in web based learning. *Journal of Software*, **4**(2), 167 – 174.
4. De Boeck, P., and Wilson, M. (2004). *Explanatory Item Response Models. A Generalized Linear and Nonlinear Approach*. New York: Springer-Verlag.
5. Everitt, B.S. (1993). *Cluster Analysis*. New York: John Wiley and Sons.
6. Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, **45**(3), 225 – 245.
7. Horton, N.J., Lipsitz, S.R., and Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician*, **57**(4), 229 – 232.
8. Ibrahim, J.G., Chen, M-H., and Lipsitz, S.R. (2001). Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, **88**(2), 551 – 564.
9. Ibrahim, J.G., Lipsitz, S.R., and Chen, M-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Royal Statistical Society*, **61**(1), 173 – 190.
10. Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New Jersey: John Wiley and Sons.
11. Molenberghs, G., and Verbeke, G. (2005). *Models for Discrete Longitudinal data*. New York: Springer Science and Business Media.

12. Peugh, J.L., and Enders, C.K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, **74**(4), 525 – 556.
13. Schafer, J.L., and Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, **7**(2), 147 – 177.
14. Stubbendick, A.L., and Ibrahim, J.G. (2003). Maximum Likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics*, **59**, 1140 – 1150.
15. Stubbendick, A.L., and Ibrahim, J.G. (2006). Likelihood-based inference with nonignorable missing responses and covariates in models for discrete longitudinal data. *Statistica Sinica*, **16**, 1143 – 1167.
16. Wu, K., and Wu, L. (2007). Generalized linear mixed models with informative dropouts and missing covariates. *Metrika*, **66**, 1 – 18.