Master Research Report: Multilevel Multivariate Imputation by Chained Equations through Bayesian Additive Regression Trees

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Heleen Brüggen

Word count: ...
Candidate Journal: Computational Statistics & Data Analysis
FETC Case Number: 23-1778

Supervisors:
MSc. T. Volker — Utrecht University
Dr. G. Vink — Utrecht University
MSc. H. Oberman — Utrecht University

1 Introduction

1.1 Introducing missing data, multiple imputation & multilevel data structure

Incomplete data is a common challange in many fields of research. A common approach for dealing with incomplete data is to remove all missing values from the data. However, this could possibly lead to biased results if the data is not Missing Completely At Random (MCAR) [van Buuren, 2018, Kang, 2013, Enders, 2017, Austin et al., 2021]. MCAR is one of the missing data mechanisms described by Rubin [Rubin, 1976]. Where MCAR means the cause of the missing data are unrelated to the data, Missing At Random (MAR) that it is related to observed data and Missing Not At Random (MNAR) that it is related to unobserved data [van Buuren, 2018, Rubin, 1976]. Furthermore, other approaches to dealing with incomplete data include: pairwise deletion, mean imputation and regression imputation, which also yield biased results [van Buuren, 2018].

Multiple imputation (MI) is considered a valid method for dealing with incomplete data [Mistler and Enders, 2017, van Buuren, 2018, Enders, 2017, Burgette and Reiter, 2010, Austin et al., 2021, Audigier et al., 2018, Van Buuren, 2007, Grund et al., 2021, Hughes et al., 2014]. MI imputes each missing value more than once, thereby considering necessary variation associated with the missingness problem. The multiply imputed data sets are analyzed, and the corresponding inferences are pooled according to Rubin's rules [van Buuren, 2018, Austin et al., 2021, Rubin, 1987]. Generally, multiple imputation operates under two frameworks: joint modeling and fully conditional specification [Mistler and Enders, 2017, van Buuren, 2018, Enders et al., 2018a, Enders et al., 2018b, Hughes et al., 2014]. Joint modeling (JM) employs a multivariate data distribution and regression model to impute missing values. Fully conditional specification (FCS), or chained equations, iteratively imputes one variable with missing values at a time through conditional univariate distributions [Mistler and Enders, 2017, van Buuren, 2018, Enders et al., 2018a, Enders et al., 2018b, Hughes et al., 2014].

1.2 Literature review (difficulty of imputing multilevel data)

1.3 Revelence of research

1.4 Research question

1.5 Hypotheses

2 Method

3 Results

References

[Audigier et al., 2018] Audigier, V., White, I. R., Jolani, S., Debray, T. P. A., Quartagno, M., Carpenter, J., Van Buuren, S., and Resche-Rigon, M. (2018). Multiple Imputation for Multilevel Data with Continuous and Binary Variables. *Statistical Science*, 33(2).

[Austin et al., 2021] Austin, P. C., White, I. R., Lee, D. S., and Van Buuren, S. (2021). Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Canadian Journal of Cardiology*, 37(9):1322–1331.

[Burgette and Reiter, 2010] Burgette, L. F. and Reiter, J. P. (2010). Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology*, 172(9):1070–1076.

[Enders, 2017] Enders, C. K. (2017). Multiple imputation as a flexible tool for missing data handling in clinical research. *Behaviour Research and Therapy*, 98:4–18.

[Enders et al., 2018a] Enders, C. K., Hayes, T., and Du, H. (2018a). A Comparison of Multilevel Imputation Schemes for Random Coefficient Models: Fully Conditional Specification and Joint Model Imputation with Random Covariance Matrices. *Multivariate Behavioral Research*, 53(5):695–713.

[Enders et al., 2018b] Enders, C. K., Keller, B. T., and Levy, R. (2018b). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods*, 23(2):298–317.

[Grund et al., 2021] Grund, S., Lüdtke, O., and Robitzsch, A. (2021). Multiple imputation of missing data in multilevel models with the R package mdmb: A flexible sequential modeling approach. *Behavior Research Methods*, 53(6):2631–2649.

[Hughes et al., 2014] Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., and Sterne, J. A. (2014). Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, 14(1):28.

[Kang, 2013] Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5):402.

[Mistler and Enders, 2017] Mistler, S. A. and Enders, C. K. (2017). A Comparison of Joint Model and Fully Conditional Specification Imputation for Multilevel Missing Data. *Journal of Educational and Behavioral Statistics*, 42(4):432–466.

[Rubin, 1976] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

[Rubin, 1987] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

[Van Buuren, 2007] Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3):219–242.

[van Buuren, 2018] van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics Series. CRC Press, Taylor & Francis Group, Boca Raton London New York, second edition edition.