GESIS Summer School in Survey Methodology 2023

Syllabus for course:
"Missing Data and Multiple Imputation"

Lecturers:	Florian Meinfelder	Angelina Hammon
Email:	florian.meinfelder@uni-bamberg.de	ahammon@diw.de

Date: 14 – 18 August 2023
Time: 09:00-12:00 | 13:00-16:00 CEST
Venue: Online via Zoom

About the Lecturers:

Florian Meinfelder is a senior lecturer at the Department for Statistics and Econometrics at the University of Bamberg, where he teaches, among others, statistical programming using R, Bayesian inference, and statistical analysis with missing data. He has mainly published on missing-data and empirical Bayes related topics. Prior to his academic appointment, he has supervised a team at GfK SE that focused on data integration and statistical matching projects.

Angelina Hammon is currently doing her PhD in Statistics. She holds a bachelor's degree in Sociology and did her master studies in Survey Statistics at the University of Bamberg. From February 2016 to September 2019, she was working as research associate in the methods group of the Leibniz Institute for Educational Trajectories (LIfBi). Since October 2019, she is research associate in the SOEP at the DIW Berlin. In addition, she works as lecturer at the Chair of Statistics and Econometrics of the University of Bamberg. Her research interests cover the appropriate handling of missing values, multiple imputation of non-ignorable missing data as well as the performance of valid inferences with complex survey data and non-probability samples.

Selected Publications:

- Hammon, A. (2022). Multiple imputation of ordinal missing not at random data, AStA Advances in Statistical Analysis.
- Meinfelder, F., & Schaller, J. (2022): Data Fusion for Joining Income and Consumtion Information using Different Donor-Recipient Distance Metrics, Journal of Official Statistics, 38 (2), pp.509-532.
- Hammon, A., Zinn, S. (2020). Multiple imputation of binary multilevel missing not at random data, Journal of the Royal Statistical Society: Series C (Applied Statistics), Vol. 69, Iss. 3, pp. 547-564
- Meinfelder, F. und Kluge, R. (Hrsg.) (2019). Bad Science: Die dunkle Seite der Statistik, Vahlen. ISBN-13: 978-3800660285
- Kamgar, S., Meinfelder, F., Münnich, R. & Navvabpour, H. (2018) Estimation within the new integrated system of household surveys in Germany. Statistical Papers, 1-27.

Course Description:

This course provides an introduction to the theory and application of Multiple Imputation (MI) (Rubin 1987) which has become a very popular way for handling missing data, because it allows for correct statistical inference in the presence of missing data. With the advent of MI algorithms implemented in statistical standard software (R, SAS, Stata, SPSS,…), the method has become more accessible to data analysts. For didactic purposes, we start by introducing some naive ways of handling missing data, and we use the examination of their weaknesses to create an understanding of the framework of Multiple Imputation. The first day of this course is of a somewhat theoretical nature, but we believe that a fundamental understanding of the MI principle helps to adapt to a wider range of

practical problems than focusing on a few select situations. We will subsequently shift to the more practical aspects of statistical analysis with missing data, and we will address frequent problems like regression with missing data. Further examples will be covered throughout the course, which are predominantly based on the statistical language R. We recommend basic R skills for this course, but it is possible to understand the course contents without prior knowledge in R, as the main MI algorithms are almost identical across all major software packages.

Keywords:

Missing Data, Item Nonresponse, Multiple Imputation, Missing at Random (MAR)

Course Prerequisites:

- General knowledge of data preparation and data analysis
- An advanced understanding of the (generalized) linear model
- Familiarity with statistical distributions
- Basic knowledge of matrix algebra helpful
- Solid skills in either R or Stata (recommended for exercises)

Target Group:

Participants will find the course useful if they:

- are survey methodologists working with incomplete data.
- are researchers who want to learn more about the analysis of incomplete data in general.
- are already aware of MI and its benefits, but feel uncomfortable about the available parameter settings in MI algorithms implemented in their preferred statistical software

Course and Learning Objectives:

By the end of the course participants will:

- be familiar with the theoretical implications of the MI framework and will be aware of the explicit and implicit assumptions (e.g. will be able to explain within an article why MAR was assumed, etc.).
- know when to use MI (and when not).
- be aware how to specify a "good" imputation model and how to use diagnostics.
- be familiar with the availability of the various MI algorithms.
- be able to not only replicate situations akin to the case studies covered in the course, but also know how to handle incomplete data in general.

Organizational Structure of the Course:

We aim to include many smaller breaks so that lecture-style teaching will be no longer than about an hour at a time. Besides the pure teaching part, there will also be several virtual lab sessions per day, so that you have the opportunity to directly implement and practice the covered material. In addition, there will be room for individual consultations on the treatment of missing data in your own projects. Course notes and other material (videos, R Markdown documents,…) will be made available via the e-learning platform ILIAS.

Software and Hardware Requirements:

Course participants will need a computer or laptop with R (https://cran.r-project.org/) and RStudio installed (https://www.rstudio.com/). Both programs are free and open source. We recommend using the Zoom desktop client for the best online teaching experience in Zoom.

Recommended Literature to look at in advance:

- Van Buuren, S. (2018), Flexible Imputation of Missing Data, CRC Press, Boca Raton. Book online available at https://stefvanbuuren.name/fimd/
- Van Buuren, S. and Groothuis-Oudshoorn, K. (2011) Multivariate Imputation by Chained Equations in R, JSS, 45, 3.
- Raghunathan, T. (2016) Missing Data Analysis in Practice, CRC Press, Boca Raton.

Day-to-day Schedule and Literature:

Day	Topic(s)
1	**Introduction to Missing-Data Terminology** ○ Missing-data mechanisms ○ Missing-data patterns **Introduction to Multiple Imputation (MI)** ○ Why MI? ○ Basic concept of MI ○ How to use Rubin's Rules
	Compulsory reading (have to be read before the session): ▪ R Markdown Document on Naïve Missing Data Handling ▪ Van Buuren, S. (2018), Flexible Imputation of Missing Data, CRC Press, Boca Raton. (ch. 1+2). Book online available at https://stefvanbuuren.name/fimd/ Suggested reading (suggested, yet do not have to be read before the session): ▪ Raghunathan, T. (2016) Missing Data Analysis in Practice, CRC Press, Boca Raton. (ch. 1+3) ▪ Enders, C. (2010). Applied Missing-Data Analysis. Guilford Pubn, New York. (ch. 1-3).
2	**Implementation of MI in R** ○ Sequential Regression and Joint Modeling ○ Introduction to the mice package ○ Overview of similarities and differences for the MI implementations in Stata and SPSS
	Compulsory reading: ▪ Van Buuren, S. (2018), Flexible Imputation of Missing Data, CRC Press, Boca Raton. (ch. 4) ▪ Van Buuren, S. and Groothuis-Oudshoorn, K. (2011) Multivariate Imputation by Chained Equations in R, JSS, 45, 3. (link).
3	**Digging deeper into MI** ○ Imputation methods ○ Analysis of multiply imputed data
	Compulsory reading: ▪ Van Buuren, S. (2018), Flexible Imputation of Missing Data, CRC Press, Boca Raton. (ch. 3+5)
4	**Empirical problems** ○ Dealing with skips and implausible values ○ Rounded and heaped data ○ Passive imputation and logical consistency
	Compulsory reading: ▪ Van Buuren, S. (2018), Flexible Imputation of Missing Data, CRC Press, Boca Raton. (ch. 6)

	▪ Raghunathan, T. (2016) Missing Data Analysis in Practice, CRC Press, Boca Raton. (ch. 5)
5	**Generalized Linear) Modelling with multiply imputed data** o Missings in covariates and response variables o Imputation of squares and interactions o Multilevel modelling **Further Applications of MI** o Data fusion and split questionnaire designs o The Rubin Causal Model
	<u>Compulsory reading:</u> ▪ Van Buuren, S. (2018), Flexible Imputation of Missing Data, CRC Press, Boca Raton. (ch. 6+7) <u>Suggested reading:</u> ▪ Raghunathan, T. (2016) Missing Data Analysis in Practice, CRC Press, Boca Raton. (ch. 5+6) ▪ Von Hippel, P. (2009), 'How to Impute Interactions, Squares, and Other Transformed Variables', Sociological Methodology, 39, 1, 265–291. ▪ Bartlett, J.W., Seaman, S.R., White, I.R., Carpenter, J.R. (2015) Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model, Stat Methods Med Res. 24(4):462-87