gesis Leibniz Institute for the Social Sciences

10th GESIS Summer School in Survey Methodology

[2nd Virtual GESIS Summer School]

28 July – 20 August 2021

Syllabus for Course 5: Statistical Analysis of Incomplete Data

Instructors:	Dr. Florian Meinfelder	Angelina Hammon, M.Sc
E-mail:	florian.meinfelder@uni-bamberg.de	angelina.hammon@lifbi.de
Homepage:	www.uni-bamberg.de/stat-oek	www.lifbi.de

Date:	09-13 August 2021
Time:	09:00-12:00 + 13:00-16:00
Time zone:	CEST/CEDT, course starts Monday at 09:00 am
Venue:	Online via Zoom

About the Instructors:

Dr. 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 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. In addition, she works as research assistant 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., 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.
- Meinfelder, F. & Schnapp, T. (2015). BaBooN: Bayesian Bootstrap Predictive Mean Matching - Multiple and Single Imputation for Discrete Data. R package, online available here: CRAN.R-project.org/package

Short 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, SPSS, or Stata (recommended for exercises).

Target Group:

Participants will find the course useful if:

- 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 intersperse teaching with many breaks for doodling/ trying out ideas on your own. In order to prevent fatigue (on yours and our side) uninterrupted teaching (lecture style) will be no longer than about an hour, before we 'break out'. In total, the pure teaching part will amount to max. 4 hours a day, but the total course time per day will be longer due to the lab/ break-out parts.

We will also prepare some smaller videos with fundamental stuff that you can download and watch anytime. Course notes and other material (videos, R Markdown documents,…) will be made available via ILIAS.

Software and Hardware Requirements:

We recommend that course participants download and install Zoom (www.zoom.us) as well as R and RStudio from www.r-project.org/ and www.rstudio.com/ (note that you should install R first and the RStudio editor subsequently) on the computer / notebook they are planning to use for the online course. If you feel comfortable enough around R, feel free to already download and install the *VIM* and the *mic*e package.

We also recommend that you use multi-display if your OS supports this, so that you can use R/RStudio on your PC/ Notebook, while the Zoom functionalities (zoom controls, tiles, chat,…) are displayed on a different screen (if available). We make annotations to the course slides during lectures, so we recommend either you have a printout version of the course notes prepared or you are using a touch-screen and software to annotate pdf files.

Long Course Description:

This course introduces Multiple Imputation (Rubin 1987) as a general method to analyze incomplete survey data. With the availability of MI in Stata, SPSS or SAS, the popularity of Multiple Imputation (MI) has in-creased over the last couple of years. Simultaneously, nonresponse issues in surveys are no longer swept under the rug in scientific publications, and the awareness of missing-data issues has increased in general. Although Multiple Imputation is based on a Bayesian framework, the inferences based on multiply imputed data sets are “classical frequentist.” Since MI is implemented in statistical standard software, the course will discuss examples for available routines (mainly in R and SPSS, but Stata features almost identical algorithms).

Participants are encouraged to suggest/share data sets in the run-up to the course which can be used for demonstration and exercises.

This one-week course can be loosely categorized into three parts: The first part of the course introduces the notation and assumptions used in statistical analysis with missing data. We will examine the distinction between missing-data patterns (e.g. non-monotone, missing b design), and missing-data mechanisms, such as ‘missing at random’ (MAR), which has gained some prominence among survey researchers. The drawbacks of ‘standard’ solutions, such as listwise deletion and mean or regression imputation are discussed, and small simulation studies are used to demonstrate their shortcomings. The first part is rounded off by looking at the mechanism of the MI framework, and why it allows for correct statistical inference in the presence of missing data if the underlying assumptions hold. The second part of the course gives an overview of available options of MI algorithms (like mice or AMELIA), discussing their strengths and weaknesses as well as their applicability to specific data situations. The third part, eventually, focuses on practical applications and exercises using different data scenarios. All stages of the MI and analysis process are reviewed using various types of diagnostics. We will further address particular empirical problems, such as avoidance of implausible values, skips (filter questions), or imputation of heaped data (e.g. arbitrarily rounded income). Additionally, we will introduce suggestions for regression analysis with partially missing covariates and/or missing response variables, and we will discuss the current status of MI research on advanced topics such as multilevel analysis of incomplete data. One advanced topic, which has been covered by recent literature and which we are not going to address into detail, is inference under non-ignorable missing-data mechanisms.

R provides a large number of packages that contain powerful MI algorithms, but we will guide through some examples using SPSS' MI algorithm as well. Stata will not be explicitly used within this course, but its MI algorithm is very close to the SPSS implementation (both are based on Stef van Buuren's 'mice'). The transfer should not pose a major problem (participants who are familiar with Stata only, can still use Stata to work on exercise problems, but the instructors will not be able to help with occurring syntax errors).

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
	<u>Compulsory reading:</u> ▪ 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/ <u>Suggested reading:</u> ▪ 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
	<u>Compulsory reading:</u> ▪ 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** ▪ Missings in covariates and response variables ▪ Imputation of squares and interactions ▪ Multilevel modelling **Further Applications of MI** ▪ Data fusion and split questionnaire designs ▪ The Rubin Causal Model
	Compulsory reading: ▪ Van Buuren, S. (2018), Flexible Imputation of Missing Data, CRC Press, Boca Raton. (ch. 6+7) Suggested reading: ▪ 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