8th GESIS Summer School in Survey Methodology
Cologne, August 2019

Syllabus for Course 07: "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	www.lifbi.de

Date: 12.-16. August 2019
Time: 09:00-13:00, 14:00-16:00
Course starts Monday morning at 09:00

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 a PhD student in Statistics at the University of Bamberg. She works as research assistant in the methods group of the Leibniz Institute for Educational Trajectories (LIfBi) and at the Department for Statistics and Econometrics of the University of Bamberg. In this context, she gives courses in Statistics on undergraduate and master's level. Her main research interests are the handling of ignorable or non-ignorable missing data and analytical inference in the context of complex survey data including hierarchical data structures and repeated measurements.

Selected Publications:

- 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
- Meinfelder, F. (2014). Multiple Imputation: an attempt to retell the evolutionary process. AStA Wirtschafts- und Sozialstatistisches Archiv, 8, 249–267.

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 half 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. However, at the latter stages of the course, frequent problems like regression with missing data will be addressed, and further typical situations will be covered by the lab session throughout the course, which is 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.

Therefore, by the end of the course, we hope that people will be able to use MI algorithms in general (irrespective of the software package) and to apply Rubin's combining rules correctly, as well as to explain to readers of their work how and why they used the method.

Keywords:

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

Course Prerequisites:

- profound understanding of sampling theory;
- an advanced understanding of the (generalized) linear model;
- familiarity with statistical distributions;
- basic knowledge of matrix algebra;
- solid skills in either R, SPSS, 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:

The course is structured around four hours of classroom instruction in the morning and a two-hour lab session in the afternoon. Within the lab session we will jointly work on problem sheets and tutorials uploaded to ILIAS. Alternatively, participants can work on their specific missing-data research projects. As of day three we additionally provide consulting time for research projects during group work hours. Please contact either of the instructors by July 31 for an appointment slot (if possible, send also the data set that is suffering from item nonresponse and/or a project description).

Software and Hardware Requirements:

Course participants are encouraged to bring a laptop computer with an installed web browser for performing the practical exercises for this course. R and RStudio can be downloaded and installed free of charge: www.rstudio.com/; www.r-project.org/. During exercises, participants will also have access to SPSS, Stata, and R in a PC lab.

Long Course Description:

This course introduces Multiple Imputation (Rubin 1987) as a general method to analyse incomplete survey data. With the availability of MI in Stata, SPSS or SAS, the popularity of Multiple Imputation (MI) has increased 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 **Naive Missing-Data Handling** ▪ Shortcomings ▪ Introduction of randomness
	Suggested reading: ▪ 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). ▪ Van Buuren, S. (2018), Flexible Imputation of Missing Data, CRC Press, Boca Raton. (ch. 1).
2	**Introduction to Multiple Imputation (MI)** ▪ Why MI? ▪ Basic concept of MI ▪ How to use Rubin's Rules

	Suggested reading: ▪ Raghunathan, T. (2016) Missing Data Analysis in Practice, CRC Press, Boca Raton. (ch. 4) ▪ Enders, C. (2010). Applied Missing-Data Analysis. Guilford Pubn, New York. (ch. 7+8). ▪ Van Buuren, S. (2018), Flexible Imputation of Missing Data, CRC Press, Boca Raton. (ch. 2). ▪ Meinfelder, F. (2014), Multiple Imputation – an attempt to retell the evolutionary process, AStA.
3	**Overview of MI algorithms** ▪ Basic concept of data augmentation ▪ Joint modeling and fully conditional specification ▪ Robust additions to the MI algorithm family: Predictive Mean Matching and Recursive Partitioning **Implementation of MI in R and SPSS** ▪ Tutorials on SPSS MI and some R packages
	Suggested reading: ▪ Van Buuren, S. (2018), Flexible Imputation of Missing Data, CRC Press, Boca Raton. (ch. 3+4).
4	**Empirical problems** ▪ Dealing with skips and implausible values ▪ Rounded and heaped data ▪ Passive imputation and logical consistency **Case studies with the R packages 'mice' and 'mi'** ▪ Graphical and analytical diagnostics ▪ Imputation model selection
	Compulsory reading: ▪ Van Buuren, S. and Groothuis-Oudshoorn, K. (2011) Multivariate Imputation by Chained Equations in R, JSS, 45, 3. (https://www.jstatsoft.org/article/view/v045i03). ▪ Su, Y-S., Gelman, A., Hill, J., Yajima, M. (2011) Multiple Imputation with Diagnostics (mi) in R Opening Windows into the Black Box, JSS, 45, 2. (https://www.jstatsoft.org/article/view/v045i02). **Suggested reading:** ▪ Van Buuren, S. (2018), Flexible Imputation of Missing Data, CRC Press, Boca Raton. (ch. 4).
5	**(Generalized Linear) Modeling with multiply imputed data** ▪ Missings in covariates and response variables ▪ Imputation of squares and interactions ▪ Multilevel modeling **Further Applications of MI** ▪ Data fusion and split questionnaire designs ▪ The Rubin Causal Model
	Compulsory reading: ▪ Von Hippel, P. (2009), 'How to Impute Interactions, Squares, and Other Transformed Variables', Sociological Methodology, 39, 1, 265–291. **Suggested reading:** ▪ Raghunathan, T. (2016) Missing Data Analysis in Practice, CRC Press, Boca Raton. (ch. 5+6) ▪ Van Buuren, S. (2018), Flexible Imputation of Missing Data, CRC Press, Boca Raton. (ch. 6+7)

Preparatory Reading:

None. The course requires advanced knowledge on sampling and probability theory as well as knowledge of standard statistical methods such as the generalized linear model.

Additional Recommended Literature:

- Raghunathan, T. (2016). Missing Data Analysis in Practice. Boca Raton: CRC Press.
- Enders, C.K. (2010). Applied Missing-Data Analysis. : New York: Guilford Pubn.
- Van Buuren, S. (2018). Flexible Imputation of Missing Data. 2nd edition. Boca Raton: CRC Press.
- Carpenter, J.R. & Kenward, M.G. (2014). Multiple Imputation and its Application. New York: Wiley.
- Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.
- Little, R.J.A. & D.B. Rubin (2002). Statistical Analysis with Missing Data. 2nd edition. New York: Wiley.