Syllabus for Course 11: “Item Nonresponse and Multiple Imputation”

Instructors:
Dr. Florian Meinfelder
Email: florian.meinfelder@uni-bamberg.de
Homepage: www.uni-bamberg.de/

Prof. Dr. Susanne Rässler
Email: susanne.raessler@uni-bamberg.de
Homepage: www.uni-bamberg.de/

Date: August 21-25, 2017
Time: 11:00-13:00, 14:00-18:00
Course starts Monday afternoon at 14:00

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.

Prof. Dr. Susanne Rässler holds the Chair of Statistics and Econometrics at the Otto-Friedrich-University of Bamberg, Germany. She was a member of the German census board, and is speaker of the methods group of the National Educational Panel Study. Her research interest comprises methods for handling missing data in complex surveys, multiple imputation, Bayesian and computational statistics as well as matching techniques for causal inference.

Selected Publications:

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 (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 the investigation of their weaknesses is used
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 can be attended 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), Data Augmentation

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).
- Knowledge of R can be acquired in the course “Introduction to Data Analysis Using R” by Steinhauer/ Würbach in week 1.

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 insecure 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 afternoon and two hours of group work in the morning. Within the two hours of group work, participants can either work on tutorial sheets uploaded to ILIAS (participants receive access to the online learning platform on the 1st of July), or work on their research projects. As of day three we 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 will need 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 analyze incomplete survey data. With the availability of MI in Stata, SPSS or SAS, the popularity of Multiple Imputation (MI) has increased dramatically over the last couple of years. Simultaneously, nonresponse issues in surveys are no longer swept under the carpet 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 missing-data analysis. 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'), and the transfer should not pose a major problem (participants who are familiar with Stata only, can still use Stata to work on the exercises, but the instructors will not be able to help with occurring syntax errors).

Day-to-day Schedule and Literature:

<table>
<thead>
<tr>
<th>Day</th>
<th>Topic(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction to Missing-Data Terminology</td>
</tr>
<tr>
<td></td>
<td>▪ Missing-data mechanisms</td>
</tr>
<tr>
<td></td>
<td>▪ Missing-data patterns</td>
</tr>
<tr>
<td></td>
<td>Naive Missing-Data Handling</td>
</tr>
<tr>
<td></td>
<td>▪ Shortcomings</td>
</tr>
<tr>
<td></td>
<td>▪ Introduction of randomness</td>
</tr>
</tbody>
</table>
### 2 Introduction to Multiple Imputation (MI)
- Why MI?
- Basic concept of MI
- How to use Rubin’s Rules

**Suggested reading:**

### 3 Overview of MI algorithms
- Basic concept of data augmentation
- Joint modelling 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:**

### 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:**

**Suggested reading:**

### 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:**

**Suggested reading:**
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: