IPSDS
INTERNATIONAL PROGRAM IN SURVEY & DATA SCIENCE

SPONSORED BY THE
Federal Ministry of Education and Research

AUFSTIEG DURCH BILDUNG >>
OFFENE HOCHSCHULEN

Multiple Imputation – Why and How

SURV 726
1 credit/2 ECTS
2017/2018 winter

Instructor(s)

Jörg Drechsler, Joerg.drechsler@iab.de

Short Course Description

This course will provide a detailed introduction to multiple imputation, a convenient strategy for dealing with (item) nonresponse in surveys. We will motivate the concept and illustrate why multiple imputation should generally be preferred over single imputation methods. The main focus of the course will be on strategies to generate (multiple) imputations and how to deal with common problems when applying the methods for large scale surveys. We will also discuss various options for assessing the quality of the imputations. All concepts will be demonstrated using software illustrations in R.

Course and Learning Objectives

By the end of the course, students will…

- understand why multiple imputation should be preferred over single imputation methods in most situations
- know about the two main approaches for multiple imputation
- be familiar with various imputation routines for different types of variables
- know how to implement these routines using R
- be able to deal with various problems that typically arise when imputing large scale surveys
- know about various strategies to assess the quality of the generated imputations

Prerequisites

Students should be familiar with generalized linear models and basic probability theory. We also expect that students know the basic concepts for dealing with nonresponse in surveys (the difference between item and unit nonresponse, formalizing the missing data mechanism, deterministic and stochastic approaches for imputation). For students unfamiliar with these concepts we highly recommend to enroll in the course “Nonresponse and Imputation” before participating in this course.

Some background knowledge in Bayesian statistics and Markov Chain Monte Carlo Methods (MCMC) is helpful but not mandatory. The statistical software R will be used for illustrations and for (some of) the homework assignments. Thus, basic knowledge of R is required to be able to complete the assignments.

Class Structure and Course Concept

This is an online course using a flipped classroom design. It covers the same material and content as an on-site course but runs differently. In this course, you are responsible for watching video recorded lectures and reading the required literature for each unit and then "attending" mandatory weekly one-hour online meetings where students have the chance to discuss the materials from a unit with the instructor. Just like in an on-site course, homework will be assigned and graded and there will be a final exam at the end of the course.

Although this is an online course where students have more freedom in when they engage with the course materials, students are expected to spend the same amount of time overall on all activities in the course – including preparatory activities (readings, studying), in-class-activities (watching videos, participating in online meetings), and follow-up activities (working on assignments and exams) – as in an on-site course. As a rule of thumb, for each credit offered by a course, students can expect to spend one hour per week on in-class activities and three hours per week on out-of-class activities over the span of a full 12-week term. This is a 1-credit course that runs for 4 weeks. Hence, the total average workload is about 12 hours per week.

Mandatory Weekly Online Meetings

Tuesday, 8 pm (CET)/2 pm (EST)
Meetings will be held online through Zoom. Follow the link to the meeting sessions on the course website on http://jpsmonline.umd.edu/. If video participation via Internet is not possible, arrangements can be made for students to dial in and join the meetings via telephone.

In preparation for the weekly online meetings, students are expected to watch the lecture videos and read the assigned literature before the start of the meeting. In addition, students are encouraged to post questions about the materials covered in the videos and readings of the week in the forum before the meetings (deadline for posting questions is Tuesday, 1 pm (CET)/7 am (EST)).

Students have the opportunity to use the Zoom meeting room set up for this course to connect with peers outside the scheduled weekly online meetings (e.g., for study groups). Students are encouraged to post the times that they will be using the room to the course website forum to avoid scheduling conflicts. Students are not required to use Zoom and can of course use other online meeting platforms such as Google Hangout or Skype.

Grading

Grading will be based on:

- 2 online quizzes (worth 20% total)
- 2 homework assignments (40% total)
- Participation in the weekly online meetings, engagement in discussions during the meetings and/or submission of questions via e-mail (10% of grade)
- A final online exam (30% of grade)

Dates of when assignment will be due are indicated in the syllabus. Late assignments will not be accepted without prior arrangement with the instructor.

Technical Equipment Needs

The learning experience in this course will mainly rely on the online interaction between students and the instructor during the weekly online meetings. Therefore we encourage all students in this course to use a web camera and a headset. Decent quality headsets and web cams are available for less than $20 each. We ask students to refrain from using built-in web cams and speakers on their desktops or laptops. We know from our experience in previous online courses that this will reduce the quality of video and audio transmission and therefore will decrease the overall learning experience for all students in the course. In addition, we suggest that students use a wire connection (LAN), if available, when connecting to the online meetings. Wireless connections (WLAN) are usually less stable and might be dropped.

Long Course Description

Missing data are a common problem in surveys which can lead to biased results if the missingness is not taken into account at the analysis stage. Multiple imputation is widely accepted as the most convenient strategy for dealing with item nonresponse in a proper way and most statistical software packages offer routines to multiply impute missing values these days. However, when treating the imputation process as a black box relying on the default settings of the software the cure can be worse than the disease. The main aim of the course therefore will be to illustrate the usefulness (and limitations) of the approach and enable the students to come up with sensible imputation strategies when dealing with item nonresponse in large scale surveys.

The course will emphasize practical implementation and tricks for handling real data problems over detailed proofs regarding the underlying methodology, although we will provide some motivation for the analysis procedures for multiply imputed datasets and briefly touch on some of the methodological pitfalls of the approach.

The course will start by illustrating why the concept should generally be preferred over standard methods which impute missing values only once (single imputation). We will also present some intuition for "Rubin's combining rules" required to obtain valid inferences from the imputed data. In the next unit we will learn about the two main strategies for multiple imputation – joint modeling and sequential regression – and discuss the pros and cons of the two approaches. Based on these two approaches we will discuss the different modeling strategies for imputing continuous and (un)ordered categorical variables. We will also present some nonparametric alternatives. To make the imputation approach feasible in practice the course will also cover strategies for dealing with real data problems such as logical constraints between the variables or skip patterns that are common in most questionnaires. The final section will provide insights how to evaluate the quality of the imputed data.

Readings

Primary readings will be from the following volumes:

Carpenter, J. and Kenward, M. (2012). *Multiple imputation and its application*. New York: John Wiley & Sons.

Additional required and recommended readings will be made available on the course website:
jpsmonline.umd.edu

Interested students might find the following additional recommended books helpful in preparing for the course:

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley & Sons.

Rubin, D. B. (1987): *Multiple imputation for nonresponse in surveys*. John Wiley & Sons

Van Buuren, S. (2012). *Flexible imputation of missing data*. CRC press.

Academic Conduct

Clear definitions of the forms of academic misconduct, including cheating and plagiarism, as well as information about disciplinary sanctions for academic misconduct may be found at

http://www.graduate.umaryland.edu/policies/misconduct.html (University of Maryland) and https://www.uni-mannheim.de/1/english/research/Good%20Research%20Practice/141119-Satzung%20wiss%20FV%20Senat_en.pdf (University of Mannheim).

Knowledge of these rules is the responsibility of the student and ignorance of them does not excuse misconduct. The student is expected to be familiar with these guidelines before submitting any written work or taking any exams in this course. Lack of familiarity with these rules in no way constitutes an excuse for acts of misconduct. Charges of plagiarism and other forms of academic misconduct will be dealt with very seriously and may result in oral or written reprimands, a lower or failing grade on the assignment, a lower or failing grade for the course, suspension, and/or, in some cases, expulsion from the university.

Accommodations for Students with Disabilities

In order to receive services, students at the University of Maryland must contact the Disability Support Services (DSS) office to register in person for services. Please call the office to set up an appointment to register with a DSS counselor. Contact the DSS office at 301.314.7682; http://www.counseling.umd.edu/DSS/.

Students at the University of Mannheim should contact the Commissioner and Counsellor for Disabled Students and Students with Chronic Illnesses at http://www.uni-mannheim.de/studienbueros/english/counselling/disabled_persons_and_persons_with_chronic_illnesses/.

Course Evaluation

In an effort to improve the learning experience for students in our online courses, students will be invited to participate in an online course evaluation at the end of the course (in addition to the standard university evaluation survey). Participation is entirely voluntary and highly appreciated.

Class Schedule

Please note that assignments and dates are subject to change. Information (e.g., articles and assignments) posted to the course website supersedes the information noted here.

Unit 1: MI Intro & MI Analysis

Online meeting (Jörg Drechsler): Tuesday January 9, 2018, 8 pm (CET)/2 pm (EST)

Online quiz 1: Tuesday, January 9, 11:59 p.m. (CET)/5:59 p.m. (EST)

Video lecture (Jörg Drechsler): Tuesday January 2, 2018

Readings:

- Chapters 2.1 – 2.4 and Chapter 2.6 from Carpenter and Kenward
- Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research* 8, 3-15.

Recommended

- Rässler, S., Rubin, D.B., Zell, E.R (2007). Incomplete data in epidemology and medical statistics. In: Rao CR, Miller J, Rao DC (eds) *Handbook of Statistics* 27, Elsevier, pp 569-601.

Unit 2: MI for Continuous Variables

Online meeting (Jörg Drechsler): Tuesday January 16, 2018, 8 pm (CET)/2 pm (EST)

Homework assignment 1: Tuesday, January 16, 11:59 p.m. (CET)/5:59 p.m. (EST)

Video lecture (Jörg Drechsler): Tuesday January 9, 2018

Readings:

- Chapter 3 from Carpenter and Kenward
- Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology* 27, 85-96.

Recommended:

- Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* 45.

Unit 3: MI for Categorical Variables and Nonparametric Alternatives

Online meeting (Jörg Drechsler): Tuesday January 23 2018, 8 pm (CET)/2 pm (EST)

Homework assignment 2: Tuesday, January 23, 11:59 p.m. (CET)/5:59 p.m. (EST)

Video lecture (Jörg Drechsler): Tuesday January 16, 2018

Readings:

- Chapters 4.1 – 4.2, 4.4 – 4.9, 5.1 - 5.2, 5.4 – 5.8 from Carpenter and Kenward
- Kropko, J., Goodrich, B., Gelman, A., & Hill, J. (2014). Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. *Political Analysis* 22, 497-519.

Recommended:

- Chapters 4.3 and 5.3 from Carpenter and Kenward

Unit 4 Modeling Strategies and Quality Evaluations

Online meeting (Jörg Drechsler): Tuesday January 30 2018, 8 pm (CET)/2 pm (EST)

Online Quiz 2: Tuesday, January 30, 11:59 p.m. (CET)/5:59 p.m. (EST)

Video lecture (Jörg Drechsler): Tuesday January 23 2018

Readings:

- Chapters 8.3 and 8.5 from Carpenter and Kenward
- Abayomi, K., Gelman, A., and Levy, M. (2008). Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society, Series C* 57, 273-291.

Recommended:

- Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 9, 538–558.
- Little R.J.A. and Raghunathan, T.E. (1997). Should imputation of missing data condition on all observed variables? *Proceedings of the Survey Research Methods Section, American Statistical Association* 1997, 617-622.

Final Exam
Due: Tuesday, February 6, 11:59 p.m. (CET), 5:59 p.m. (EST)

Note: Student access to the course website will be revoked two weeks after the final exam.

	Unit 1	**Unit 2**	**Unit 3**	**Unit 4**
Video available	Tuesday January 02, 2018	Tuesday January 09, 2018	Tuesday January 16, 2018	Tuesday January 23, 2018
Online meeting	Tuesday January 09 2018, 8 pm (CET)/2 pm (EST)	Tuesday January 16 2018, 8 pm (CET)/2 pm (EST)	Tuesday January 23 2018, 8 pm (CET)/2 pm (EST)	Tuesday January 30 2018, 8 pm (CET)/2 pm (EST)
Online quiz due	Tuesday, January 9, 11:59 p.m. (CET)/5:59 p.m. (EST)			Tuesday, January 30, 11:59 p.m. (CET)/5:59 p.m. (EST)
Homework due		Tuesday, January 16, 11:59 p.m. (CET)/5:59 p.m. (EST)	Tuesday, January 23, 11:59 p.m. (CET)/5:59 p.m. (EST)	
Final exam due				Tuesday, February 6, 11:59 p.m. (CET), 5:59 p.m. (EST)