

SURV 726 Multiple Imputation

1 credit/2 ECTS

Prof. Jörg Drechsler
Video lecture by Prof. Jörg Drechsler

Fall 2023

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 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 prior to participating in mandatory weekly one-hour online meetings where students have the chance to discuss the materials from a unit with the instructor.

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, you can expect to spend approximately 3h/week on in-class-activities and 9 hours per week on out-of-class activities (preparing for class, readings, assignments, projects, studying for quizzes and exams). Therefore, the workload in all courses will be approximately 12h/week. This is a 1-credit/2-ECTS course that runs for 4 weeks. Please note that the actual workload will depend on your personal knowledge.

Mandatory Weekly Online Meetings

Wednesday, 2:00 PM EDT/8:00 PM CEST, starting November 8, 2023

Meetings will be held online through Zoom. Follow the link to the meeting sessions on the course website. 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 Wednesday, 7:00 AM EDT/1:00PM CEST).

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 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)

<i>A+</i>	<i>100 - 97</i>
<i>A</i>	<i>96 - 93</i>
<i>A-</i>	<i>92 - 90</i>
<i>B+</i>	<i>89 - 87</i>
<i>B</i>	<i>86 - 83</i>
<i>B-</i>	<i>82 - 80</i>
<i>Etc.</i>	

The grading scale is a base scale recommended by the UMD. Variations for grading on a scale are at the discretion of the instructor.

The final grade will be communicated under the assignment "Final Grade" in the Canvas course. Please note that the letter grade written in parentheses in Canvas is the correct final grade. The point-grade displayed alongside the letter grade is irrelevant and can be ignored.

Dates of when assignment will be due are indicated in the syllabus. There will be a grace period for late assignments (**not for quizzes**), but late assignments will be penalized according to the following rules:

1 day late:	10% off
2 days late:	25% off
3 days late:	50% off
4+ days late:	no credit

Technical Equipment Needs

The learning experience in this course will mainly rely on the online interaction between the students and the instructors 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

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

Required and Recommended Readings

Additional required and recommended readings will be made available on the course website:

<https://www.elms.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

<https://www.president.umd.edu/sites/president.umd.edu/files/documents/policies/III-100A.pdf> (University of Maryland)

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 Accessibility & Disability Service (ADS) office to register in person for services. Please call the office to set up an appointment to register with an ADS counselor. Contact the ADS office at 301.314.7682; <https://www.counseling.umd.edu/ads/>.

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.

Sessions

Week 1: MI Intro & MI Analysis

Video lecture: available Wednesday, November 1, 2023

Online meeting: Wednesday, November 8, 2023 2:00 pm EDT/8:00 PM CEST

Quiz 1 due: Wednesday, November 8, 2023 1:00PM EDT/7:00PM CEST

Required 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 Readings:

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

Week 2: MI for Continuous Variables

Video lecture: available Wednesday, November 8, 2023

Online meeting: Wednesday, November 15, 2023 2:00 pm EDT/8:00 PM CEST

Assignment 1 due: Wednesday, November 15, 2023 1:00 pm EDT/7:00 PM CEST

Required 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 Readings:

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

Week 3: MI for Categorical Variables and Nonparametric Alternatives

Video lecture: available Wednesday, November 15, 2023

Online meeting: Wednesday, November 22, 2023 2:00 pm EDT/8:00 PM

Assignment 2 due: Wednesday, November 22, 2023 1:00 pm EDT/7:00

PM CEST

Required 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 Readings:

Chapters 4.3 and 5.3 from Carpenter and Kenward

Week 4: Modeling Strategies and Quality Evaluations

Video lecture: available Wednesday, November 22, 2023

Online meeting: Wednesday, November 29, 2023 2:00 pm EDT/8:00 PM

Quiz 2 due: Wednesday, November 29, 2023 1:00 pm EDT/7:00 PM CEST

Required 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 Readings:

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.

Project/Homework/Final exam

A final open-book online exam

Due: Wednesday, December 6, 2023 3:00 pm EDT/9:00 PM CEST