Course Description

This goal of this course is to provide students with a basic knowledge of the potential implications of missing data for their data analyses as well as potential solutions. We will begin by discussing different types of mechanisms that can generate missing data. This will lay the groundwork for discussions of what types of missing data scenarios can be accommodated by each missing data method discussed subsequently. Finally, we will learn how to deal with missing data in Stata. More advanced techniques will be covered in Bayesian Statistics for the Social Sciences in the Spring.

Prerequisites

Any QMSS student is presumed to have sufficient background. Any non-QMSS students interested in taking this course should have sufficient background in regression modeling of discrete variables.

Grading

Grading will be based 25% on the bi-weekly assignments, 25% on the midterm exam, 15% class participation / final presentation, and 35% on the final exam.

Books

The main textbook for the class is Stef van Buuren’s *Flexible Imputation of Missing Data*, published by CRC Press in 2012. Columbia University affiliates can access this book for free in electronic form, or it can
be purchased from the usual booksellers (Google Books Link). Some other inexpensive books that are not required but are less intense are Allison, Paul (2002) Missing Data, Sage University Press (Google Books Link) and McKnight, Patrick E., McKnight, Katherine M., Sidani, Souraya, and Aurelio Jose Figueredo (2007) Missing Data: A Gentle Introduction, Guilford Press (Google Books Link).

Outline

The following outline describes the topics that will be covered along with anticipated associated readings. It corresponds roughly to the course weeks though we may end up adjusting time spent on each topic as we go. We generally suggest doing the readings after the Monday class but before the Wednesday class each week.

I. Tools

Week 1: Introduction (Wednesday only)

Week 2: Probability Theory


Week 3: Continue Probability Theory

Week 4: Endogenous Selection


II. Possible Solutions

Week 5: Missing Data Mechanisms

- van Buuren, sections 1.1 and 1.2
- See also Pearl, Judea and Mohan, Karthika. “Recoverability and testability of missing data: Introduction and summary of results”. Presentation at the Joint Statistical Meetings in Montreal, Summer 2013. Available from here
Week 6: Complete Case Analysis

- Galati, John C and Seaton, Katherine A. “MCAR is not necessary for the complete cases to constitute a simple random subsample of the target sample”, *Statistical Methods in Medical Research*, Online May 2013, forthcoming [Link](#)

Week 7: Midterm Week

- Monday: Go over homework and review semester
- Wednesday: Midterm exam

Week 8: Multiple Imputation and Why It Often Works

- van Buuren: Section 1.3, Section 1.4 and Chapters 2 – 4

Week 9: Missing Data with Stata

- Royston, “Multiple imputation of missing values: Further update of ice, with an emphasis on categorical variables”, *Stata Journal*, 9:3, 466 – 477, [Link](#)

Week 10: More on Multiple Imputation

- van Buuren, Chapters 5 – 6

Week 11: Multivariate Normal Imputation

- King, Honaker, Joseph, and Scheve (2000) [Link](#)
- Honaker, King, and Blackwell (2012) [Link](#)

Week 12: Post-imputation Diagnostics

- Abayomi, Gelman, and Levy (2008) [Link](#)

IV. Putting It All Together

Week 13: Presentations (more details later)

Week 14: (Monday only): Review for Final Exam