Missing Data
Spring 2017

Course Time and Location:
Friday 9:30 am – 1 pm
TBA
Instructor: Dobrin Marchev
Office Hours: TBA

Course Description and Prerequisites:
The goal of this course is to provide students with a basic knowledge of the potential implications of missing data on 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. Simple missing data fixes (for example listwise deletion) will be described next as well as the problems they can create in terms of bias and loss of efficiency. We next explore some slightly more complicated fixes (for instance various types of single imputation) and the assumptions required for valid inference for each. The course will end with at least three weeks of focus on multiple imputation including discussions of the general framework, different models and algorithms and the basic theory. More detailed focus will be spent on implementation of the `mi` package in `R`. If time allows the course may finish with a discussion of missing data mechanisms that are not missing at random (NMAR) or more details for the Bayesian methods in multiple imputation.

The prerequisite is at least three semesters of quantitative methods beyond introductory statistics (for instance RESCH-GE.2003 and 2004 or the equivalent as approved by the instructor). It is particularly important that students should be comfortable with the following concepts before the course begins: Binomial probability model, logistic regression, transformations, and use of regression diagnostics to identify lack of model fit. Computer code will be presented in the `R` statistical software language (freely available) and `R` will be required for the multiple imputation work. While students are not required to have prior experience with `R` they will be asked to complete an online tutorial on the basics of `R` before the class begins (more information to be provided through emails and the website of the course).

Assignments. Each week students will perform data analyses that correspond to that week's readings and lecture. These will be performed both on a common dataset and on students' own data. We will discuss the results as a class and all *students will be expected to be able to contribute to this discussion by explaining how they approached parts of the assignment*. The weekly analyses will be used towards a final project that will be turned in at the end of the semester.

Grading. Grading will be based primarily (85%) on one project comprising an amalgamation of all the weekly assignments. Short in-class quizzes will count as 15% of the semester grade.

Reading materials

Required reading materials

- van Buuren, S. (2012) *Flexible Imputation of Missing Data*, Chapman & Hall

Recommended reading materials

- Venables, W. N. (2009) *An Introduction to R*
- Allison, Paul (2002) *Missing Data*, Sage University Press.
- McKnight, Patrick E., McKnight, Katherine M., Sidani, Souraya, and Aurelio Jose Figueredo (2007) *Missing Data: A Gentle Introduction*, Guilford Press.
- Little, Roderick J. A., Rubin, Donald B. (2002) *Statistical Analysis with Missing Data*, Wiley-Interscience

The other required and recommended readings will either be available through e-journals via the library or will be posted on the course website.

Other online resources that might be of interest:

A repository of information on multiple imputation:
http://www.multiple-imputation.com/

A repository of R documentation/tutorials:
http://cran.r-project.org/

Outline of course topics and readings:

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. Readings highlighted with an * are recommended, not required. All readings not freely available on the web or through the library's ejournals system will be posted on Classes under Resources/Readings.

Topics and assigned readings:

0) Introduction to R. Please complete the following on your own *before the first class.*

Please complete the following tutorials:

Verzani, *simpleR*, p. 94 (installing R, external packages), pp. 1-35, pp. 41-46, 77-89, 94-100 (this document is available at cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf)

http://tryr.codeschool.com/levels/1/challenges/1
You may also find the resources at DataCamp helpful, e.g.:
https://www.datacamp.com/courses/free-introduction-to-r

1) **Getting comfortable with R.**
   Missing Data Mechanisms. How are missing data generated and why should we care?
   Complete case analyses.
   van Buuren, pp. 3-9, 25-35, 48, 63-65
   *Allison, pp. 1-6
   *McKnight et al., Chapters 2 and 3

2) **Simple missing data fixes: available case, LVCF, mean imputation, dummy variable**
   van Buuren, pp. 8-23
   *Allison, pp. 6-11
   *McKnight et al., Chapters 5, Chapter 7 (pp. 150-151), Chapter 9 (pp. 173-190)

3) **More complicated missing data fixes: weighting, hotdecking, regression imputation**
   van Buuren, pp. 8-23, 53-56
   *McKnight et al., Chapter 8 (pp. 170-172), Chapter 9 (pp. 182-195)
   *Allison, pp. 11-27

4) **Stochastic regression imputation (regression imputation with noise)**
   Conceptual overview of multiple imputation
   van Buuren, pp. 25-43, 53-56
   *McKnight et al., Chapter 10
   *Allison, pp. 27-50

5) **Multiple imputation in practice**
   Software in R, simple analyses, and diagnostics)
   van Buuren, pp 34-51 (Remember though that we will not be using `mice` to perform imputations in `R`)

6) **Multiple imputation in practice**
   More complicated models and considerations, more advanced diagnostics
   van Buuren, pp 53-82
   Abayomi, Gelman, and Levy paper on multiple imputation diagnostics

7) **More advanced Bayesian imputation and other missing data methods**
   Little and Rubin, Chapter 10
   van Buuren, pp TBD