Missing Data (RESCH-GE.2013)
Spring 2013

Course Time and Location:
First class: March 29th, 2013
Friday 9:30-12:15
Loc: TBD
Instructor: Jennifer Hill
Office Hours: Typically Wednesdays from 4-5PM

Course Description and Prerequisites:
This 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 strategies for multiple imputation in the context of multilevel models.

The prerequisite is at least two semesters of quantitative methods (for instance E10.2003 and E10.2004 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: the Binomial probability model, logistic regression, transformations, and use of regression diagnostics to diagnose lack of model fit. Computer code will be presented in the `R` statistical software language and `R` will be required for the multiple imputation work (`R` is freely available). 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 Blackboard site for the course). *Students must also show up on the **first** day with their own dataset (with missing data) to be used for assignments throughout the course.*

Assignments. Each week students will perform different data analyses that correspond the weeks readings and lecture. These will be performed both on a common dataset and on their 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 that will be created as an amalgamation of all the weekly assignments. Class participation will count as **15%** of the semester grade.

Reading materials

Required reading materials to be purchased
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.

Recommended reading materials available for purchase
Venables, W. N. (2009) *An Introduction to R*

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

Other online resources that might be of interest
For a nice review of basic statistical concepts:
http://oli.web.cmu.edu/openlearning/forstudents/freecourses/statistics

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

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

More specific R advice for Windows:
http://math.illinoisstate.edu/dhkim/rstuff/rtutor.html

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 Blackboard.

Topics and assigned readings

0) **Introduction to R.** **You will complete the following on your own *before the first class.***
Read the following and try out the provided code as you follow along.
Verzani, *simpleR*, p. 94 (installing R, external packages), 1-19, p. 94-100 (sample session)
[this document is available at cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf and on the Blackboard site under course documents]

1) **Getting comfortable with R.**
Missing Data Mechanisms. How are missing data generated and why should we care?
Complete case analyses.

Allison, pp. 1-6
McKnight et al., Chapters 2 and 3
Verzani, *simpleR*, pp. 41-46, 77-89

2) **Simple missing data fixes: available case, LVCF, mean imputation, dummy variable methods**
Allison, pp. 6-11
McKnight et al., Chapters 5, Chapter 7 (pp. 150-151), Chapter 9 (pp. 173-190)
JASA article on dummy variable strategy posted to course documents, week 2

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

4) **Building blocks and overview of multiple imputation (including regression imputation with noise)**
McKnight et al., Chapter 10
Allison, pp. 27-50

5) **Multiple imputation in practice**
Software in R, simple analyses, and diagnostics)

6) **Multiple imputation in practice**
More complicated models and considerations, more advanced diagnostics

7) **More advanced imputation and other missing data methods**
Brief presentations of individual projects