Modern Bio-Statistical Methods for Incomplete Data

COURSE SYLLABUS

Instructor: Dr. David Todem  
Department of Epidemiology and Biostatistics  
(Office) B644 West Fee Hall  
Phone: 353-8623 Ext. 171  
E-mail: todem@msu.edu

Meeting Times: Mondays and Wednesday: 1.50pm-4.40pm (7/1/2019 - 8/15/2019)  
Classroom: E111 Fee Hall  
Office Hours: Mondays after class (or by appointment)

Course Description: Missing data abound in both observational and experimental research. They reduce the precision of parameter estimates, and perhaps more critically, may introduce bias into the statistical analysis. This course examines a range of modern statistical tools for analyzing incomplete data generated from randomized and observational studies using mainstream statistical software. The course will discuss and provide a critique of conventional and simple methods such as complete case analysis and LOCF (Last Observation Carried Forward) before moving on to the more advanced techniques such as Direct likelihood, Weighted Estimating Equations or Inverse Probability Weighting, Multiple Imputation, the EM Algorithm, and Sensitivity Analyses for Informative Missing data. Because methods for handling missing data, for the most part, depend on the measurement data model being entertained, the course will start with a general overview of statistical models for fully observed independent and correlated (longitudinal) data, with a focus on likelihood and non-likelihood based inference methods. The course ends with a discussion of expert committee recommendations regarding The Prevention and Treatment of Missing Data in Clinical Trials.

The SAS software will be used for implementation, but other software such as R and SPSS will be discussed. Practical examples from the instructor’s own experience of analyzing incomplete data in real applications will be used to illustrate the methodologies. Specific data sets come from medicine, clinical trials, public health and related bio-sciences.

Prerequisites: Basic statistics courses including regression models for continuous and discrete data. No prior knowledge about missing data concepts or methods will be assumed.

Format: Classes will be a mixture of lectures and labs.
1. Required EPI 880 Course Notes  
Lecture notes in electronic form, mostly pdf files (developed by David Todem, MSU; Dr. Verbeke, G. and & Dr. Molenberghs, G., Hasselt University Belgium, and Dr Little, R and Dr Raghunathan, T, University of Michigan) will be posted on D2L.

Grading: Students will be evaluated based on their performance on the homework/project assignments (30%), the mid-term exam (30%), and the final-term project (40%).

Course Home Page: MSU D2L: https://d2l.msu.edu/
SELECTED TOPICS

- **Introduction and general overview**
  - Basic concepts and definitions
  - Missing data patterns (Monotone and Arbitrary patterns)
  - Missing data mechanisms (MCAR, MAR, MNAR) and implications on inference
  - Overview of simple and advanced analytic methods for missing data
  - Examples of missing data problems

- **Overview of classical models for fully observed (complete) data**
  - Models for independent data (generalized linear models)
  - Models for correlated data (likelihood and non-likelihood based models)
  - Model fitting using the SAS software

- **Unweighted and Weighted Complete Case analysis**
  - General analytic strategies
  - Unweighted Complete case (case deletion) analysis
  - Weighted Complete case analysis (Inverse Probability Weighting methods)
  - Available case analysis and Weighted GEE models under MAR
  - Implementation using the SAS software

- **Imputation Methods**
  - Simple imputation methods (LOCF and (un) conditional averages)
  - Imputation methods that account for imputation error (Resampling methods - bootstrapping and Jackknife; Multiple imputation)
  - Software for Multiple Imputation Analysis (SAS, Stata, R)
  - Limitations of imputation methods used in many commercial software

- **Likelihood theory and Computational tools for incomplete data**
  - Maximum likelihood and Bayes estimation (Direct likelihood inference)
  - Computational tools for monotone patterns (factored likelihood/posterior distribution)
  - EM algorithm for non-monotone (arbitrary) patterns
  - Missing data approach as a data augmentation strategy for estimation via the EM

- **Methods for Missing Not at Random data mechanism**
  - Joint models for missing data and measurement data (Selection model factorization, Pattern mixture factorization, shared random effect models)
  - Model identifiability and Unverifiable modeling assumptions
  - Using methods only consistent with MAR for bias reduction for MNAR data
  - Sensitivity Analysis for non-ignorable missing Data