

Biostatistics 880

Statistical Analysis with Missing Data

Fall 2024

8:30 – 10:00 Tues and Thurs, 1755 SPH I

First meeting: August 27, 2024

Instructor: Professor Donglin Zeng
dzeng@umich.edu

Statistical analysis of data sets with missing values. Pros and cons of standard methods such as complete-case analysis, and imputation. Likelihood theory with missing data. ML, Bayes inference, and semiparametric estimation equations for common statistical problems when data are missing, including regression, repeated-measures analysis, and contingency table analysis. Approaches for missing not at random (MNAR) missingness mechanisms. Sensitivity analysis. Applications to surveys, clinical trials, etc. Computational tools include the EM algorithm and extensions, the Gibbs' sampler, and multiple imputation. Furthermore, analysis of partially missing data such as censored and truncated data is discussed. Application to causal inference and dynamic treatment regimes is introduced in terms of a missing data framework.

3 hour credit, letter grade.

Required Text:

Little, R.J.A. and Rubin, D.B. (2019). *Statistical Analysis with Missing Data*, 3rd edition. New York: John Wiley.

Recommended Text:

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

Hernan, M. and Robins, J. (2020). *Causal Inference*. Boca Raton: Chapman & Hall/CRC.

Chakraborty, B. and Moodie, E. (2013). *Statistical Methods for Dynamic Treatment Regimes*. New York: Springer.

Prerequisites: Biostatistics: 601, 602, 650, 651, 653, + at least one of 617, 619, 682, 690, 695 suggested. Statistics: 500, 501, 510, 511. Students with good quantitative skills in epidemiology, psychology, business, economics and other social sciences are welcomed, with the consent of the instructor. Knowledge of standard statistical models such as multivariate normal, multiple linear regression, contingency tables; matrix algebra, calculus, and basic maximum likelihood for common distributions.

Grading: Homework (biweekly) 60%, Midterm exam 15%, Final presentation 15%, Class attendance and participation 10%. Homework will receive a full grade if handed in on time with showed effort. There will be an in-class midterm, and then, in the latter part of the semester, individual paper on missing data topics. The last few meetings will be student presentations based on their readings. There will be no final exam. **Class attendance is required and participating in class discussion is strongly encouraged.**

Office Hours: 3-5pm Friday or by appointment @M4126

TA: Yize Hao (yizehao@umich.edu) office hour: 1:30-2:30 Wednesday (location TBA)

COURSE OUTLINE

Topics highlights include:

- Introduction and overview. What is a missing data problem? Patterns and mechanisms of missing data.
- Available-case and complete-case analysis, nonresponse weighting
- Single imputation, propagating imputation uncertainty
- Likelihood theory for complete and incomplete data
- ML and Bayes for monotone patterns: factored likelihood methods
- ML for general patterns: EM algorithm and extensions
- Bayes and multiple imputation for general patterns: MCMC approaches.
- Normal models with missing data; measurement error as missing data; robust approaches.
- Models with categorical data; mixed continuous and categorical data
- Missing not at random models and approaches
- Inverse probability weighted estimating equation and augmentation
- A general framework for targeted inference based on semiparametric efficiency theory

- Partial missing data: right- or interval- censoring
- Missing data framework for causal inference
- Heterogeneous treatment effects and dynamic treatment regimes
- Missing data in machine learning

Lecture Schedule (subject to change):

	Date	Description	Reference
1	Aug 27	Introduction	LR Chapter 1
2	Aug 29	Introduction ctd.	LR Chapter 1
3	Sep 3	Complete-case analysis/weighting	LR Chapter 3
4	Sep 5	Imputation	LR Chapter 4
5	Sep 10	Imputation uncertainty	LR Chapter 5
6	Sep 12	Maximum Likelihood Theory	LR Chapter 6
7	Sep 17	Bayesian inference	LR Chapter 6
8	Sep 19	Factored Likelihood Methods	LR Chapter 7
9	Sep 24	EM algorithm	LR Chapter 8
10	Sep 26	Bayes and MI by Chained Equations	LR Chapter 10
11	Oct 1	Missing not at random	LR Chapter 15
12	Oct 3	Missing data in other problems	
13	Oct 8	Mid-term exam (in class, open book))	
14	Oct 10	IPW estimating equation	
15	Oct 15	Fall study break no class	
16	Oct 17	IPW estimating equation	
17	Oct 22	General semiparametric EE I	Tsiatis Chapter 3
18	Oct 24	General semiparametric EE II	Tsiatis Chapter 9,10
19	Oct 29	Right censored data	Cox (1972; 1975)
20	Oct 31	Interval censored data	Zeng and Lin (2016)
21	Nov 5	Causal inference	RH (2020)
22	Nov 7	Dynamic treatment regimes 1	MG(2018)
23	Nov 12	Dynamic treatment regimes 2	MG (2018)
24	Nov 14	Review of homework problems	
25	Nov 19	Dynamic treatment regimes 3	MG (2018)
26	Nov 21	Missing labels in machine learning	Wang et al. (2013)
27	Nov 26	Final presentations 1	
28	Nov 28	Thanksgiving break no class	
29	Dec 3	Final presentations 2	
30	Dec 5	Final presentations 3	
31	Dec 17	Final presentations 4	(10:30-12:30 pm)