BIST P8157:

Analysis of Longitudinal Data

Fall 2023

Columbia Mailman School of Public Health

Basic information:

Instructor:	Zhonghua Liu
	zl2509@cumc.columbia.edu
Lectures:	Tue & Wed 8:30 - 9:50 a.m.
	Mailman 8th Floor Auditorium
Office hours:	By appointment
TA:	Ruiyang Li and Anja Shahu
	rl3034@cumc.columbia.edu;
	as6798@cumc.columbia.edu
Office hours:	TBD
	TBD

Course webpage: https://courseworks2.columbia.edu/courses/184521

- course handouts/slides
- lab materials
- homework assignments
- datasets
- code

Computing: Majority (if not all) will be in `R`

Important dates:

- November 7th (Tuesday): Election day
- November 22th (Wednesday): Fall break
- December 13th (Wednesday): Last day of class

Recommended texts:

- Diggle P, Heagerty P, Liang KY, Zeger S. *Analysis of Longitudinal Data*. 2nd Edition. Oxford University press, 2013.
- Garrett M Fitzmaurice, Nan M Laird, and James H Ware. *Applied longitudinal analysis*. Wiley; 2nd edition
- Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. 2nd Edition. Springer, 2009.

BIST P8157: Overview

Analysis of longitudinal data, with an emphasis on regression modeling.

I. Introductory topics in longitudinal data analysis

II. General linear models for dependent data

III. Linear mixed-effects models

IV. Marginal models

V. Generalized linear mixed-effects models

VI. Missing data

VII. Time-dependent covariates

Comments

- This course considers settings where responses are *multivariate* and/or exhibit (residual) *dependence*

- Questions:
  - ⋆ what do we mean by a 'multivariate' response?
  - ⋆ what does it mean for responses to be 'dependent'?

- At the outset, we'll find that naïve application of methods from generalized linear regressions results in incorrect inference
  - ⋆ standard errors will be typically be too small
  - ⋆ intuitively, when we naïvely assume the responses to be independent we think we have more 'information' than we really do have

- Eventually we'll find that in trying to fix problems of inference, we add complexity
  - ⋆ choice of model
  - ⋆ choice of strategy for estimation/inference
- Throughout, a number of themes will be emphasized
  - ⋆ the importance of model interpretation
  - ⋆ the importance of understanding operating characteristics
  - ⋆ the use of simulation as a means to evaluating operating characteristics
- Estimation and inference will be approached primarily within the frequentist paradigm
- There'll be plenty of technical detail but an important goal should be to develop an intuition
  - ⋆ at some point in your career you'll faced with explaining these ideas to non-statisticians!

- Grades in the class will be based on the following:

Homework	40%
Mid-term exam	30%
Final project	30%

- There'll be 4 homework assignments during the semester
  - ⋆ usually take 1 or 2 weeks
- The midterm will be in the form of a one-week take-home exam
- The final will consist of a 3-4 week project
  - ⋆ topic that is not covered in class
  - ⋆ 10-page written report that will, in part, serve as resources to everyone in the class

- Topics might include:
  - ⋆ Bayesian estimation/inference, including variational Bayes
  - ⋆ correlated survival outcomes
  - ⋆ joint modeling of longitudinal and survival data
  - ⋆ genome-wide association studies
  - ⋆ mediation in the longitudinal context
  - ⋆ networks
  - ⋆ outcome-dependent sampling
  - ⋆ principal components analysis
  - ⋆ spatial modeling and prediction
  - ⋆ informative cluster sizes
  - ⋆ asymptotics for GEE
  - ⋆ small-sample bias corrections
  - ⋆ cluster-randomized trials
  - ⋆ resampling procedures (e.g. the bootstrap)
  - ⋆ clustering methods

- For the projects in this class, there are a number of options:
  - ⋆ choose one from a list that I will compile and make available closer to the time the projects will start
  - ⋆ a topic of your choice
  - ⋆ a detailed analysis of an interesting dataset