Special Topics in Applied Biostatistical Practice

Public Health 298 [CCN 76708], Spring 2016
Facilitators: Lucia Petito and Kelly Street
Instructor of Record: Sandrine Dudoit

Facilitator contact info

Lucia Petito	Kelly Street
lucia.petito@berkeley.edu	kstreet@berkeley.edu

Class info

Meeting place: 89 Dwinelle Hall, Mondays 5-6pm
Office hours: 114 Haviland Hall, Fridays 10-11am
https://piazza.com/berkeley/spring2016/ph298/

Prerequisites

Some statistical theory and computing, like
Stat 134/135/201A/201B/133/243.

Course description

This course is a rigorous presentation of the practical issues encountered when applying statistical methods to public health and biological data. Course material includes software to enable collaborative research, make use of survey data, and implement computational algorithms. All topics covered incorporate hands-on data analysis with computer software (primarily R).

Traditional lectures will alternate with workshop sessions in which students will complete guided computer lab activities. These workshops, to be completed individually or in groups of up to size 3, are designed so students will implement the technique(s) presented in class the previous week. Readings will be assigned to supplement the material presented in lecture. Class will meet for one hour every week. Two long-term projects will be assigned throughout the semester to give students an opportunity to apply appropriate methods to a data analysis of their choosing.

Course learning objectives

By the end of this course, students will be able to:

- Facilitate collaborative research, via use of GitHub in conjunction with RStudio, as well as through creation of an R package
- Visually communicate statistical findings, through use of effective graphics in R (using the packages lattice, ggplot2, maps)
- Implement multiple imputation procedures in R (write code themselves, use packages mi and mice)
- Execute analyses of survey data
- Summarize large data sets with dimensionality reduction techniques
- Build and assess effective prediction models through bagging and boosting
- Fit non-linear models with smoothing and kernel methods

Reference texts

All required readings will be provided to students as PDFs through the course website.

Some supplemental texts that might be of interest are:

- Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York.
- Lumley, T. (2010) *Complex Surveys: A Guide to Analysis Using R.* Wiley, New York
- Wickham, H. (2009) *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag, New York.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.

R Software

All analyses in this class will be done in R, a set of open-source statistical software. R is free, and can be downloaded at http://www.r-project.org. A basic familiarity with R is expected of all students in the class.

Evaluation procedures

To earn a pass, a student must:

- Attend all class meetings (exceptions will be considered on a case-by-case basis)
- Participate in every in-class workshop
- Design and execute 2 projects, which can be done in groups of up to size four, to be approved by the facilitators

Additional policies

Personal computer use is encouraged, especially during the in-class workshops. Cell phone use is discouraged, but not banned. There will be no extra credit offered.

Academic integrity is a requirement for this class. For more information on the official university policy, see http://sa.berkeley.edu/sites/default/files/UCB-Code-of-Conduct-new%20Jan2012_0.pdf

Students with disabilities should first register with the Disabled Students' Program and then arrange a private meeting with the course facilitators to make particular arrangements. See http://www.dsp.berkeley.edu for more information.

Class schedule

week	date	topic	reading due	assignments due
1	1/25	Workflow scenarios to facilitate collaborative statistical research	---	---
2	2/1	Making attractive graphics in R	Tom Preston-Werner. “The Git Parable” Friedrich Leisch. “Creating R Packages: A Tutorial”	Project 1 Assigned
3	2/8	Workshop: basic plot, mapping, ggplot2, RShiny, create an R package, install and use GitHub	Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis, Chapter 2: “Getting started with qplot”	
------Feb 15: PRESIDENT’S DAY-------				Project 1 proposal due
4	2/22	Multiple imputation for missing data	Andrew Gelman, Jennifer Hill. Data Analysis Using Regression and Multilevel/Hierarchical Models. Chapter 25: “Missing-data imputation”	---
5	2/29	Statistical estimation and inference for survey data	Sarndal et al. Model Assisted Survey Sampling. Chapter 2: “Basic Ideas in Estimation from Probability Samples”	---
6	3/7	Workshop: mi and mice R packages, NHANES Data and survey package in R	Thomas Lumley. Complex Surveys: A Guide to Analysis Using R. Chapter 3: “Cluster Sampling”	---
7	3/14	In-class project presentations	---	Project 1 Due
------March 21: SPRING BREAK-------				
8	3/28	Dimensionality Reduction Techniques: PCA, Factor Analysis	---	Email Burke (burke@berkeley.edu) for cluster account Project 2 Assigned
9	4/4	Introduction to parallelization and how to use the cluster (GrizzlyBear)	Burke’s biostat cluster FAQs	
10	4/11	Workshop: Boostrapping and cross validation in parallel	---	Project 2 proposal due
11	4/18	Bagging (Random Forest) and boosting (AdaBoost)	Breiman. “Random Forests,” 2001. Freund & Schapire. “A Short Introduction to Boosting,” 1999.	
12	4/25	Workshop	---	
13	5/2	Kernel smoothing and local regression	Jacoby. “Loess: a nonparametric, graphical tool for depicting relationships between variables,” 2000 (sections 1-8).	---
14	5/9	Project Presentations	---	Project 2 Due