EPSY 887 Computational Statistics

Spring 2013

Instructors: Jason Bryer & Robert Pruzek
Email: jason@bryer.org, rpruzek@albany.edu
Phone: 518-542-0375
Website: github.com/jbryer/CompStats

Class Time: Monday 6:00 - 9:00
Class Location: Ed 13
Office Hours: by appointment.

Course Description This seminar will provide an introduction to statistical programming for data analysis with an emphasis on the analysis of large datasets. With the increased availability of large national and international datasets (e.g. PISA, TIMMS, NAEP, ECLS) there is a great opportunity and potential for researchers to address important questions. However, the analysis of large datasets requires special analytical procedures not found in commercial statistics software. Utilizing the open source statistical software R, students will be introduced to the tools and procedures for analyzing large datasets with an emphasis on conducting transparent and reproducible research.

Why learn to program? (Shalizi, 2012, `http://www.stat.cmu.edu/~cshalizi/statcomp/`)

- *Independence*: otherwise, you rely on someone else always having made exactly the right tool for you, and giving it to you.
- *Honesty*: otherwise, you end up distorting the problem to match the tools you happen to have.
- *Clarity*: turning your method into something a machine can do forces you to discipline your thinking and make it communicable.

Course Objectives
At the completion of the seminar students should be able to analyze large datasets using reproducible techniques. This includes utilizing analytic approaches not available in commercial statistical software.

Prerequisites
Two graduate level statistics courses (e.g. EPSY 530 & EPSY 630).

Grade Policy
This course will be graded on a pass/fail basis. Successful students will actively attend and participate in class and the course website. During the first two weeks students will identify a dataset and a general research question that will serve as their primary project for the semester. Each week students should be prepared to provide an update on their progress of their data analysis. Students will give a presentation at the end of the semester on their results (even if preliminary). A document outlining their procedures and annotated bibliography will also be required.

Course Outline:

	Topic	Reading
Jan 28	Introduction to R (e.g. data input, recoding, etc.)	Kabacoff Ch 1 & 2
Feb 4	Advanced R	
Feb 11	Reshaping data	Kabacoff 5.6.3 Wickham (2007)
Feb 18	Data visualization vis-à-vis a grammar of graphics	`http://had.co.nz/ggplot2`
Feb 25	Introduction to programming for data analysis (e.g. loops, conditional statements, functions, etc.)	Matloff Ch 7
Mar 4	Object oriented programming (S3, S4, and Reference Classes)	Chambers
Mar 11	Missing data	Kabacoff Ch 15
Mar 18	No class - Spring Break	
Mar 25	Analysis of complex surveys (e.g. use of replicate weights and multiple plausible values)	Lumley
Apr 1	No class	
Apr 8	Document preparation and typesetting with LaTeXand Sweave	
Apr 15	R package development	
Apr 22	Software project management principles as applied to data analysis (e.g. source countrol, progress tracking, versioning, Github, R-Forge, etc.)	
Apr 29	Propensity Score Analysis	
May 6	TBD	

Other topics as identified by students and appropriate for analysis of large datasets. Topics may include propensity score analysis, multilevel modeling, IRT, random forests, regression trees, cluster analysis, factor analysis, etc.

References

Braun, W.J., & Murdoch, D.J. (2007). *A First Course in Statistical Programming with R.* Cambridge, UK: Cambridge University Press.

Gentle, J. E. (2009). *Computational Statistics.* New York, NY: Springer.

Kabacoff, R.J. (2011). *R in Action: Data Analysis and Graphics with R.* Shelter Island, NY: Manning.

Kolata, G. (July 7, 2011). How bright promise in cancer testing fell apart. *The New York Times.* Available from `http://www.nytimes.com/2011/07/08/health/research/08genes.html?_r=1&pagewanted=print`

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R.* Hoboken, NJ: John Wiley & Sons, Inc.

Matloff, N. (2011). *The Art of R Programming.* San Francisco, CA: No Starch Press.

Organization for Economic Co-Operation and Development (2009). Programme for International Student Assessment (PISA) 2009 Framework. Available from http://www.pisa.oecd.org

Rizzo, M.L. (2008). *Statistical Computing with R.* London: Chapman & Hall/CRC.

Unwin, A., Theus, M., & Hofmann, H. (2006). *Graphics of Large Datasets: Visualizing a Million.* New York, NY: Springer.

Wilkinson, L. (2005). *The Grammar of Graphics* (2nd Ed). New York, NY: Springer.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* New York, NY: Springer.

Software
This course will rely heavily on the use of R. R is a free, open source software programming language focused on data analysis and is quickly becoming the *de facto* standard among statisticians and data analysts*. You can download R for Mac, Windows, and Linux from the Comprehensive R Archive Network (CRAN) at `http://cran.r-project.org`. Windows users should also download and install Rtools from CRAN as well. It is also recommended that you install RStudio, an Integrated Development Environment (IDE) from `http://rstudio.org` as this provides a vastly superior experience for interacting with R.

Data
One of the major purposes for computational statistics is to make better use of data. In some instances we will simulate data, but to get the most out of this seminar we will use real datasets. Class examples will make extensive use of PISA. A list of available data sets is available on the course wiki at `https://github.com/jbryer/CompStats/wiki/Data-Sources`.

Course Website
We will utilize Github and Dropbox for sharing information. Github functions primarily as a source control repository for storing and tracking changes in documents and source code (e.g. R scripts). It also provides a wiki that we will use for sharing information and issue tracking that we will use for questions and answers.

Academic Integrity
Whatever you produce for this course should be your own work and created specifically for this course. You cannot present work produced by others, nor offer any work that you presented or will present to another course. If you borrow text or media from another source or paraphrase substantial ideas from someone else, you must provide a reference to your source.

The University policy on academic dishonesty is clearly outlined in the Student Bulletin, and includes, but is not limited to plagiarism, cheating on examinations, multiple submissions, forgery, unauthorized collaboration, and falsification. These are serious infractions of University regulations and could result in a failing grade for the work in question, a failing grade in the course, or dismissal from the University. `http://www.albany.edu/undergraduate_bulletin/regulations.html`

Reasonable Accommodation
Reasonable accommodations will be provided for students with documented physical, sensory, systemic, cognitive, learning and psychiatric disabilities. If you believe you have a disability requiring accommodation in this class, please notify the Director of Disabled Student Services (Campus Center 137, 442-5490). That office will provide the course instructor with verification of your disability, and will recommend appropriate accommodations. For more information, visit the website of the UAlbany Office for Disabled Student Services. `http://www.albany.edu/studentlife/DSS/guidelines/accomodation.html`

*See the following articles by Bob Muenchen on the popularity of R: `http://r4stats.com/articles/popularity/` and `http://r4stats.com/2012/05/09/beginning-of-the-end/`.