Statistics 470: Data Analytics and Statistical Computing
Syllabus Fall 2015

Organization:

- Instructor: A. Buja, JMHH 471
- TA: M. Sklar
- Lectures: Mon/Wed 12noon-1:30pm, JMHH F55
- Software: “R” from http://www.r-project.org/
- Readings:
  1. “The R Language” (Venables and Ripley)
  2. “Advanced R” (H. Wickham)
  3. Documentations and vignettes of various R packages

- This course is cross-listed as Statistics 503 for graduate students outside of the Statistics department. Apologies to MBAs – cross-listing as a 700-level course will have to wait for next year.

Prerequisites:

- Stat 111/112 or Stat 101/102 or Stat 431.
- No programming experience required.

General outline:

This course offers introductions to

- programming in the widely used R language,
- principles and techniques that are broadly useful for statistical data analysis,
- techniques for collecting and cleaning data,
- data visualization.
Detailed but tentative list of materials:

Sections 1 and 2 reinforce the fundamental notions of statistical variability and its consequences for estimation and inference. Section 3 deals with data acquisition through web scrubbing, as well as the dirty business of data cleaning and feature extraction. Section 4 on data visualization will introduce some powerful graphical tools to reveal overall as well as detailed structure in complex data. Finally, Section 5 summarizes aspects of the R language that will be taught interwoven with the contents of Sections 1-5.

The material below will not be presented in the order shown. The order reflects the logic of the content, not the pedagogy.

1. Fundamentals of simulation:
   (a) uniform random numbers and the inverse cdf mapping
   (b) simulating draws from simple probability models
   (c) simulating draws from regression models (prelude for parametric bootstrap)
   (d) random sampling of finite populations
   (e) simulation of simple stochastic processes (e.g., Brownian motion)
   (f) accuracy of simulation estimates

2. Nonparametric statistical inference based on simulation:
   (a) Permutation tests of simple association
   (b) Standard errors from nonparametric bootstrap
   (c) Inferential diagnostics for regression models using parametric bootstrap
   (d) Visual inference with the line-up method

3. Data janitorial work:
   (a) web scrubbing, text data
   (b) data cleaning
   (c) feature extraction
   (d) network extraction (social networks)
   (e) simple imputation of missing values (fixed & multiple)

4. Data visualization:
   (a) Univariate plots for all variable types
   (b) Bivariate plots for all combinations of variable types
   (c) Coplots for higher-dimensional data
   (d) 3D visualization
   (e) Network visualization
5. R material, interwoven with the above:

(a) Syntax
(b) Atomic data types: numeric, character, logical, missing
(c) Data structures: vectors, dataframes, matrices, arrays, lists
(d) Data queries: enumeration, exclusion, Boolean selection, associative arrays
(e) Input/output
(f) Vectorization for efficient numeric computing
(g) Natural language tools: pattern matching, regular expressions
(h) Programming principles:
   i. expressions and evaluation
   ii. lexicographical scope
   iii. functional programming
   iv. simple object-oriented programming and generic functions
(i) Model language and model formulae as used in
   i. statistical models
   ii. plotting
(j) Visualization:
   i. low-level base graphics and the construction of efficient complex plots
   ii. the lattice package for coplots
   iii. high-level graphics with ggplot2
   iv. animations with simple loops
   v. interactive graphics through event handling
(k) The world of R packages: a selection
   i. Big-data structures in R
   ii. Web scrubbing, handling html and xml