Introduction to Computational Statistics

Northeastern University
PPUA 5301, Fall 2017
Thursday 5:15-7:45pm, 237 Richards Hall

Professor: Nick Beauchamp
Email: n.beauchamp@neu.edu
Website: nickbeauchamp.com
Office Hours: Th 3-5, and by appt. (RP 931)

Course Description: Modern data analysis increasingly faces an embarrassment of riches: abundant and complex data, along with increasingly sophisticated techniques for modeling these data and building and testing theories. This course provides an introduction to the fundamental techniques of quantitative data analysis with an emphasis on the diverse skills needed for contemporary work: data acquisition and management, scripting and sampling, probability and statistical tests, econometric models, machine learning, and data visualization. These diverse skills are developed using the open-source R statistical computing language, which has become the dominant statistical tool for modern data analysis.

The course begins with an exploration of R and its use in data analysis, programming, and visualization. This is followed by a review of probability and statistics, followed by statistical tests; OLS regression; categorical dependent variables; and time series. The course finishes with an introduction to analyzing more complex data using machine learning methods. Throughout, there will be an emphasis on the challenges and limitations of modeling complex data, and students will finish with the basic skills needed to manipulate data, answer hypotheses statistically, and present their insights to non-experts. The course should also serve as a solid introduction to more advanced classes in econometrics, machine learning, or network science, for instance.

Prerequisites: This course proceeds from the ground up, and introduces all of the necessary concepts along the way. However, the steep learning curve means that students will be better off coming in with at least some familiarity with either statistics or programming. But students of all backgrounds are welcome if they are ready to put in the work to acquire new skills on a weekly basis.

Course Format: The course meets weekly for 2.5 hours, with a break in the middle. The first 1-2 hours consists of lecture and discussion, while the remaining time consists of hands-on activities with R, including practice coding, reviewing past assignments, and working through new assignments.

Course Activities and Assignments: There are weekly homework assignments, as well as a take-home midterm and final. Homeworks and exams are available on Blackboard under

Assignments, and should be submitted as PDFs to Blackboard for grading, generally by noon on Thursdays (see the Assignments page for weekly details). There is no final project, although some of the later homework assignments will allow you to apply techniques and present results using data of your choosing.

Participation in class, in the form of discussion, helping fellow students, and demonstrating code, is also essential and is 10% of the course grade. As part of your participation, please keep an eye out for interesting statistics-related stories in the news, and if something catches your eye, we will be spending the first few minutes of each class discussing such items, their strengths, flaws, etc.

Course Grading Criteria:

- Homework – 50%
- Midterm Exam – 20%
- Final Exam – 20%
- Participation – 10%

Required Textbooks:

- *R for Everyone*. Jared P. Lander; Addison Wesley, 2014.
- *Learning Statistics Using R*. Randall E. Schumacker; Sage, 2015.

Class Schedule & Topical Outline

Weekly reading assignments and more detailed "blueprints" for each weekly module are available in the Syllabus section on Blackboard. Homework guidelines and other crib sheets are available in Course Resources. Note that the following schedule is subject to change, as we may require more time in for various topics, such as probability (week 3) or multiple regression (week 9), or less time (R, weeks 1 & 2).

Part 1	Data Analysis Using R
Part 2	Probability and Statistics
Part 3	Regression
Part 4	Advanced Analytic Methods

Date	Week	Topics	Subtopics
Sept 7	1	Introduction to R	a. Variable types and basic math b. Vectors, matrices and data frames; data import and export
Sept 14	2	Scripting and Graphics in R	a. Coding, loops, and vectorized operations b. Visualizing data with ggplot2
Sept 21	3	Probability	a. Discrete and continuous distributions; marginal and conditional probabilities. b. Binomial, Poisson, normal, and other common distributions

Sept 28	4	Statistics	a. Samples and populations; population parameters b. Central Limit Theorem; standard errors; T distribution
Oct 5	5	Statistical Tests 1	a. Significance, p-values, alpha level, type 1 and type 2 errors b. Means tests and difference in means tests
Oct 12	6	Statistical Tests 2	a. F test and ANOVA b. Chi-square test
Oct 19	7	Bivariate Regression	a. Correlation and partial correlation; b. OLS; significance tests; R^2
Oct 26	8	Multiple Regression 1	a. Interpreting coefficients and regression results b. Causal inference
Nov 2	9	Multiple Regression 2 **(Take-home Midterm due)**	a. Quadratic terms; b. Interactions.
Nov 9	10	Advanced Regression Methods	a. Categorical dependent variables b. Time series
Nov 16	11	Unsupervised Machine Learning	a. Factor and principal component analysis b. Clustering
Nov 23		(Thanksgiving)	
Nov 30	12	Supervised Machine Learning	a. Shrinkage methods and elastic net b. Support vector machines
Dec 7	13	Review	
Dec 14		**Take-home Final due**	

Academic Honesty: Students are expected to do their own work for both homework and exams. For homework assignments, students are welcome to discuss problems and issues with each other using the online forums, but all submitted work should be the student's own. Students are not allowed to discuss the midterm or final exam with anyone, and all questions about the exams should be addressed to the instructors. Plagiarism, copying from other students, or submitting the work of someone not in the program are grounds for expulsion from the course.

Honor Code: All students must adhere to the Northeastern University honor code available here: http://www.northeastern.edu/osccr/academic-integrity-policy and in the graduate student handbook.

Special Accommodations: If you have specific physical, psychiatric or learning disabilities that may require accommodations for this course, please contact Northeastern's Disabilities Resource Center (DRC) at (617) 373-2675. The DRC can provide you with information and assistance to help manage any challenges that could affect your performance in the course. The University requires that you provide documentation of your disabilities to the DRC so that they may identify what accommodations are required, and arrange with the instructor to provide those on your behalf, as needed.