Last Updated: October 2018

Introduction to Computational Statistics

Northeastern University
PPUA 5301, Fall 2018

Class Times & Location:	Thursdays 5:15-8:35 Kariotis Room 209
Instructor:	Ieke De Vries Churchill Hall 400D Violence and Justice Research Laboratory i.devries@northeastern.edu
Office Hours:	Tuesdays 3:00-5:00 *Or by appointment*

Required Textbooks:

- Lander, Jared P. (2014). *R for Everyone.* Addison Wesley. https://www.jaredlander.com/r-for-everyone/
- Schumacker, *Randall, E. Learning Statistics Using R.* Sage.
  Additional readings can be found online and/or will be posted on the Blackboard web page for this course.

Course Description: Quantitative research has become increasingly interdisciplinary and now merges the worlds of – among other - social scientists, statisticians, data scientists and computer scientists. Much of this collaborative work is motivated by the nature and quantity of data that becomes available due to technological advances. Such data require diverse data gathering and processing skills and both inferential and predictive techniques need to be in a contemporary researcher's toolkit. The current course provides an introduction to the fundamental techniques of quantitative data analysis, which include data acquisition and management, scripting and sampling, probability and statistical tests, econometric models, machine learning, and data visualization. These skills will be developed utilizing the open-source R Statistical computing language, which has become the dominant statistical tool for modern data analysis.

The course is comprised of four blocks starting with an introduction to computational skills in R (*R and scripting*) followed by the application of these skills in the context of probability and statistics, data gathering and processing, analysis and visualization. The course addresses various statistical distributions, sample guidelines, and statistical tests (*probability and statistics)* and seeks to develop skills related to collecting, cleaning, analyzing and visualizing data utilizing ordinary least square regressions, regressions for categorical or count dependent variables (*data gathering, processing, analysis and visualization)* and more advanced methods like unsupervised and supervised machine learning techniques (*advanced methods: machine learning).* As such, the course serves as a foundation to more advanced classes in econometrics, machine learning, or network science, for instance.

Course Learning Goals: The course aims to provide students with the fundamental techniques of quantitative data analysis such that:

- Students understand the concepts and math behind data description and visualization, probability and statistics;
- Students can apply statistical techniques, including some of the more advanced topics such as machine learning, to real-world data utilizing the R statistical computing language.
- Students learn to engage in analysis of complex data relevant to a wide variety of fields such as business and economics, health and medicine, marketing, public policy, computer science, engineering, and more.
- Students are ready to apply the analytic methods taught in this course to their own data problems and can present their results to nonexperts.
- Students leave the course ready to participate in more advanced course work and data analytics.

Course 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 on Thursdays from 5:15 to 8:35, with a break in the middle. Courses typically start with a 1-2 hour lecture covering new material, followed by brief hands-on activities, reviewing past assignments and explaining new assignments. Laptops are necessary for in-class activities.

Course Activities and Assignments: There are weekly homework assignments, as well as a midterm and take-home final. All assignments are individual assignments; there are no group projects. Students are encouraged to give each other feedback but students cannot submit the same assignments or part of assignments. At the beginning of the semester, each student will be assigned to an interdisciplinary reflection group that serves as an opportunity to engage in productive discussions and feedback with other students in the course.

Course Grading: It is the student's responsibility to become familiar with the assigned readings prior to each class. Your final grade will be calculated based on the following:

Requirement	*Due Dates*	*%*
Class Discussion and Participation	-	10%
Weekly Homework Assignments	Every next Thursday 12:00 pm.	50%
Mid-Term Exam	**October 25, 2018**	**20%**
Final Exam	December 13, 2018	20%
Final Grade	-	100%

Class Discussion and Participation: Participation in class, in the form of discussion, helping and collaborating with other students (e.g. by demonstrating code) is essential and contributes to 10% of the course grade.

Weekly Homework Assignments: There are weekly homework assignments to become familiar with each week's new material. Homework assignments will be made available through Blackboard under Assignments and should be submitted as PDFs to Blackboard for grading by noon on Thursdays. Each new assignment is due the next Thursday at 12pm.

Exams: Mid-term and Final Exams will be made available through Blackboard under Assignments and should be submitted as PDFs to Blackboard by the due date. There is no final project though some homework assignments allow you to collect, analyze and present data of your choosing.

Course Resources:

Lecture Notes: Will be made available to students.

Other Resources: https://www.datacamp.com; https://stackoverflow.com; many R Tutorials online (e.g. https://www.rstudio.com/online-learning/).

Blackboard and Communication: A discussion board for this course is available on Blackboard. The discussion board can serve as a platform to ask course-related questions and help fellow students.

If you need to contact to me, use the email provided above. Please put "PPUA5301" in the subject line of the email.

Class Policies:

Early Departures and Absences
No early departures or absences are permitted unless previously discussed with me. Systematic tardiness or early departures (defined as being late for class or leaving class early more than 2 times) will lead to a deduction from your participation grade.

Late and Missing Assignments
I must be notified in advance if you anticipate missing an assignment for a valid reason (e.g. serious emergencies). Documentation may be requested and I reserve the right to approve or deny any such requests. Missing assignments should be turned in at a later point in time to be agreed upon with your instructor. Late assignments without a valid reason that was notified to the instructor in advance results in a grade deduction.

Respect
The course involves students from different disciplines and with varying familiarity to programming and/or statistical analyses. This demands an open attitude, respect and a willingness to help fellow students.

Sports-Related Absences Policy
All student-athletes are required to notify me at the beginning of the course about all sports-related absences.

Students with Disabilities

Any student who may require special accommodations for this course should notify me as soon as possible. You may need to register with the university's Disability Resource Center (DRC). The DRC can provide students with services such as note-takers and extended time for taking exams. The DRC is located in 20 Dodge Hall and can be reached at 617-373-2675.

Academic Integrity Policy:

- All students must follow Northeastern University's procedures regarding academic integrity. Commitment to the principles of academic integrity is essential to the mission of Northeastern University. Northeastern University expects students to complete all examinations, tests, papers, creative projects, and assignments of any kind according to the highest ethical standards as set forth in the Northeastern University Student Handbook. It is the student's responsibility to become familiar with their rights and responsibilities.
- A detailed explanation of what constitutes academic cheating, plagiarism, and facilitating academic dishonesty, and how such cases are handled by Northeastern University, is in the Student Handbook (pp. 40-42; http://www.northeastern.edu/osccr/wp-content/uploads/2016/06/2016-2017-UG-Handbook.pdf). I have summarized a few points below (but this is by no means exhaustive):
    - Cheating includes handing in the same paper for more than one course without explicit permission from the instructors.
    - Cheating includes storing notes in a portable electronic devise for use during an examination.
    - Plagiarism can occur accidentally or deliberately. It is defined as using as one's own the words, ideas, data, code, or other original academic material of another without providing proper citation or attribution. Forgetting to document ideas or materials taken from another source does not exempt one from plagiarizing.
    - Participation in academically dishonest activities includes misrepresenting oneself or one's circumstances to an instructor
    - Facilitating academic dishonesty is defined as intentionally or knowingly helping or contributing to the violation of any provision of the Northeastern University Student Handbook.
        - This includes doing academic work for another student.
        - This includes making available previously used academic work for another individual who intends to resubmit the work for credit.
- In this course, cheating or plagiarizing on an assignment, as defined by Northeastern University's Academic Integrity Policy, will result in receiving a "0" on that assignment, meaning it may result in a failing grade for the course. This conduct will also be reported to Office of Student Conduct and Conflict Resolution (http://www.northeastern.edu/osccr/).
- If you have any questions of whether you should be citing to a source or paraphrasing a source in a different way, please let me know. Often situations implicating academic integrity can be avoided as they come about due to confusion regarding appropriate citations.

Grading Scale:

A	930-1000	B+	870-899	C+	770-799	D+	670-699	F	<= 599
A-	900-929	B	830-869	C	730-769	D	630-669		
		B-	800-829	C-	700-729	D-	600-629		

Class Schedule and Topical Outline

Subject to change at the discretion of the instructor. Changes will be announced in class and through Blackboard. It is your responsibility to be aware of such changes.

Part 1	Introduction to R and Scripting
Part 2	Probability and Statistics
Part 3	Data gathering, processing, analysis and visualization
Part 4	Advanced Methods: Machine learning

Date	Week	Topics	Subtopics	Bonus Topics (*if time permits*)*	Readings
Part 1 Introduction to R and Scripting					
Sep 6	1	Introduction to R	a. Variable types and basic math b. Vectors, matrices and data frames; data import and export	*Ethical Concerns* *Group activity bonus topic interests*	Lander 1-2; 4.1-4.4; 5-6 (except 6.7)
Sep 13	2	Scripting and Graphics	a. Coding, loops, and vectorized operations b. Visualizing data with ggplot2	*Data cleaning*	Lander 3, 7-10, 11.1; 11.2, 12.2, 13 Schumacker 8
Part 2 Probability and Statistics					
Sep 20	3	Probability	a. Discrete and continuous distributions; marginal and conditional probabilities b. Binomial, Poisson, normal, and other common distributions	*Sampling hidden or hard-to-reach populations*	Lander 14 Schumacker 3, 5 (pp. 66-81)

Sep 27	4	Statistics	a. Samples and populations; population parameters b. Central Limit Theorem; standard errors; T distribution		Lander 15.1 Schumacker 4, 5 (pp. 64-66), 6 (pp. 106-112)
Oct 4	5	Statistical Tests 1	a. Significance, p-values, alpha level, type I and type II errors b. Means tests and difference in means tests		Lander 15.3.1-15.3.2 Schumacker 10, 13
Oct 11	6	Statistical Tests 2	a. F test and ANOVA b. Chi-square test		Lander 15.4 Schumacker 11, 14
Part 3 Data gathering, processing, analysis and visualization					
Oct 18	7	Bivariate Regression	a. Correlation and partial correlation; b. OLS; significance tests; R^2	*Data sources: offline and online, primary and secondary*	Lander 15.2, 16.1 Schumacker 15-16
Oct 25	8	Multiple Regression 1	a. Interpreting coefficients and regression results b. Causal inference **Midterm Exam Due**	*Missing Data*	Lander 16.2, 18.1, 18.2, 18.5 Schumacker 17
Nov 1	9	Multiple Regression 2	a. Quadratic terms; b. Interactions	*Mediation*	Lander 20.1 Schumacker 19
Nov 8	10	Advanced Regression Methods	a. Categorical dependent variables b. Count dependent variables	*Replication and Validation (Panda video)* *Time Series or Count*	Lander Chapter 17.1-17.3 Schumacker 18

				dependent variables	
Nov 15	11	Time to meet with reflection group and work on assignment			
Nov 22	12	**No class – Thanksgiving**			
Part 4 Advanced Methods: Machine learning					
Nov 29	13	Unsupervised Machine Learning	a. Factor and Principal component analysis b. Clustering		Lander Chapter 22.1, 22.3
Dec 6	14	Supervised Machine Learning	a. Shrinkage methods and elastic net b. Support vector machines	*What's next: Advanced computational techniques (e.g. text or networks as data)*	Lander Chapter 18.3, 19.1, 20.4, 20.5.
Dec 13	15	**Final exam due Dec. 13**			

*Bonus topics will not be part of exams. Ideas for bonus topics? Email me!