INSH 5301 Intro Computational Statistics Spring 2020 (Tentative)

Credit Hours: 4 **Term**: Spring 2020 Instructor: Hao-En Kao	**Format:** Online **Office Hours:** by appointment E-mail: h.kao@northeastern.edu

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.

Students will finish with the basic skills needed to (1) manipulate data, (2) answer hypotheses statistically, and (3) 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 Structure and Requirements: This course has a number of different components to it ------- some that will be familiar to you from other courses you have taken before and some that are more unique to the online format. We follow a weekly structure, where each week consists of a single "module" of materials. All the online materials are contained within our course's Blackboard website, and that is the primary platform through which we will conduct the class. Each week, you will complete the materials and tasks within a module. These include:

(1) Reading the module **Overview, Learning Objectives**, and **Key Concepts**;

(2) Viewing the **Lecture slides**, which are the heart of the course;

(3) Reading the **textbook** chapters;

(4) Viewing any **videos** or **external materials** that are indicated; doing the

(5) **Homework** and Participating in **discussions**.

In addition to the modules, there will also be a Midterm Exam and a Final Exam, each of which are take-home and have a week devoted to them without other lecture materials or homework, yeah!

The course runs a total of 16 week, including starts from Jan 6 to April 26, with 12 modules, 2 (midterm and final) exams, 1 week of reviewing and 1 week of Spring break.

Weekly Schedule: See below for a detailed schedule of the course content. Please note that all due dates and times are specified according to the Eastern Standard Time zone (EST).

Module launch: Each week's module will be launch on 1 minute before Monday, that is Sunday 11:59 PM at the previous week, other than week 1, which is on exactly Jan. 1st of 2020 12:01 AM. For example, Week 2 will be on Jan. 12th 11:59 PM on Sunday, 1 minute before Monday of week 2.

Homework: Each week's homework will appear at the time module launch, and the homework will be due on the 11:59 PM Tuesday the following week. Please do NOT send HW with email, upload the as PDF assignment to BB. Each week's homework will grade and return by BB the following week. Solutions are given next Sunday at 11:59 PM after deadline. For example, week 1 HW is due on Jan 14 (Tue) at 11:59 PM and the solution will be shown on Jan 19 at 11:59 PM (Sun).

Discussion Boards: All questions regarding course material or homework should be posted on the week's Bb Discussion Board, The instructors will reply to unanswered questions and provide further clarifications using the discussion boards. Participation is graded just as it is in a normal class: if you contribute to discussions on a fairly regular basis, and occasionally help out your fellow students with their questions, that's great, and you'll get full points. Don't be afraid of making errors – for discussion board, I care only about participation and helpfulness, not about whether everything is precisely correct. Also, check if anyone has the problem before you post, there is a chance that someone have already asked the question.

Email: Please used direct email only for private communication about grading, appointments and other private matters. Everyone will be better off by seeing a post in public, even if you think the problem is very specific to you, someone else might have the same problem. Once again, do not send assignment by email! It's extremely difficult to grade!

Appointment: If you wish to see me in person, I am at 3rd floor of Renaissance Park almost all weekdays in the afternoon. But please email me before coming so I would not go eating or taking some rest.

Exam: Midterm and final exam are there at a module of the week and follow the schedule of the week.

Assignments: Guidelines for completing and submitting each assignment are posted along with the assignment in Blackboard. Please note that if you are unable to complete an assignment within the time period provided, a documented compelling excuse is required. For excuses to be valid they have to anticipate the due date, except for unavoidable emergencies. Late submission that are not properly excused will be discounted 20% each day, so after 5 days after the assignment is due, it's basically an automatic zero point.

Textbooks
R for everyone: Jared P. Lander; Adison Wesley, 2014
Learning Statistics Using R: Randall E. Schumacker, Sage, 2015

Grading

Exams and homework will be graded on a 100 point scale. General comment will be annotated upon grading, especially when taking off points. For detailed explanation, please refer to solution the Sunday after submission is due.

Course Grading Criteria:

Homework: 50%

Midterm Exam: 20%

Final Exam: 20%

Participation: 10%

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.

Weekly schedule and topic outline

Week	Module	Topics	Subtopics
Part 1: Data Analysis Using R			
Jan 5	1	Introduction to R	a. Variable types and basic math b. Vectors, matrices and data frames; data import and export.
Jan 12	2	Scripting and Graphics in R	a. Scripting, conditionals, loops, functions, and vector operations b. Visualizing data with ggplot2
Part 2: Probability and Statistics			
Jan 19	3	Probability	a. Discrete/continuous distributions; marginal/conditional prob. b. Binomial, Poisson, normal, and other common distributions.
Jan 26	4	Statistics	a. Samples and populations; population parameters. b. Central Limit Theorem; standard errors; T distribution.
Feb 2	5	Individual Hypothesis Test	a. Significance, p-values, alpha level, type I and type II errors. b. Means tests and difference in means tests.
Feb 9	6	Joint Hypothesis Test	a. F test and ANOVA. b. Chi-square test.
Feb 16	Midterm Exam		
Part 3: Regression			
Feb 23	7	Bivariate Regressions	a. Correlation and partial correlation. b. OLS; significance tests; R^2
Mar 1	Spring break!		
Mar 8	8	Bivariate Regression	a. Interpreting coefficients and regression results b. Causal inference
Mar 15	9	Multiple Regression 2	a. Quadratic terms; b. Interactions.
Mar 22	10	Advanced Regression Methods	a. Categorical dependent variables b. Time series
Part 4: Advanced Analytic Methods			
Mar 29	11	Unsupervised Machine Learning	a. Categorical dependent variables b. Time series
Apr 5	12	Supervised Machine Learning	a. Shrinkage methods and elastic net b. Support vector machines
Apr 12	13	Review	
Apr 19	Final Exam		