Course Information

GR5206 Statistical Computing and Introduction to Data Science
417 International Affairs Building
Fridays 2:40pm - 5:25pm

Instructor

Name:	Linxi Liu
Email:	ll3098 [at] columbia [dot] edu
Office hours:	Thursdays, 5:00pm - 6:30pm

Overview

Statistical computing is an essential element of modern statistics curricula as solid programming skills and good computational understanding are necessities for current statisticians. Statisticians are routinely expected to gather data from disparate sources and implement the most current methodologies, both of which require computational fluency. This course is an introduction to the basics of statistical programming, targeted at entering statistics master students and senior undergraduate students with minimal prior programming knowledge. Examples from data science will be used throughout the course for demonstration. Students will be introduced to basic machine learning topics such as classification, regression, and clustering methods, resampling techniques including the bootstrap, cross-validation, and permutation tests, as well as the basics of optimization. At the end of the semester students will have:

- The ability to read and write code for statistical data analysis,
- An understanding of programming topics such as functions, object, data structures, debugging, etc.,
- An introduction to statistical learning methods applied to real-word data.

The class will be taught in the R language using the RStudio interface.

Prerequisites for GR5206: STAT GR5204 and GR5205 or the equivalent. Students will also be expected to have basic knowledge of linear algebra, elementary probability, and multivariate calculus.

All resources from class will be posted on Canvas https://courseworks.columbia.edu/welcome/. Check the web site often for any important course-related announcements. You need your UNI and password to log into the system.

Textbook

Note that all text books are optional. I highly recommend the the first three.

- *R for Data Science*; Garrett Grolemund and Hadley Wickham.
- *An Introduction to Statistical Learning*; Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
- *Computational Statistics*; Geof Givens and Jennifer Hoeting.
- *Advanced Data Analysis from an Elementary Point of View*; Cosma Shalizi.
- *The Art of R Programming: A Tour of Statistical Software Design;* Norman Matloff.
- Maybe more..

Teaching Assistants

We have a course mailing list: gr5206_course_staff [at]columbia [dot] edu

For any course-related inquiries, please send them to the mailing list. Please DO NOT email the instructor or the TAs in person. Any email directly sent to the instructor or the TAs WILL NOT get replied.

Name: Owen Ward
Section: 002
Lab: Wednesdays, 6:10pm-7:25pm
Office Hour: Mondays, 6:10pm-7:25pm

Name: Rishabh Dudeja
Section: 003
Lab: Tuesdays & Thursdays, 11:40am-12:55pm

Name: Chaoyu Yuan
Section: 004
Lab: Mondays, 8:40am-9:55am
Office Hour: Wednesdays, 8:40am-9:55am

Name: Ding Zhou
Section: 005
Lab: Thursdays, 8:40am-9:55am
Office Hour: Tuesdays, 8:40am-9:55am

Office Hours

Monday	Owen Ward	6:10pm - 7:25pm	411 International Affairs Buiding
Tuesday	Ding Zhou	8:40am - 9:55pm	702 Hamilton Hall
Wednesday	Chaoyu Yuan	8:40am - 9:55am	313 Fayerweather
Thursday	Linxi Liu	5:00pm - 6:30pm	*TBA*

Graders

TBA

Grading

Your overall course grade will be determined as a weighted average of the following categories:

10%	Lab	*The lowest score will be dropped.*
20%	Homework assignments	*The lowest score will be dropped.*
30%	Midterm exam	**Friday, Oct. 19, 2018, 2:40pm - 5:25pm,** *in lecture*
40%	Final exam	**Friday, Dec. 14, 2018, 1:10pm - 4pm,** *location TBA*

Software

R and RStudio will be used throughout the course and the assignments. R is open-source statistical software that can be downloaded at https://www.r-project.org and RStudio at https://www.rstudio.com. We expect that students will have the software downloaded before class begins.

Academic Honesty

The university expressly prohibits academic dishonesty such as cheating, plagiarism, etc. It provides for a number of rather unpleasant consequences for students who are caught in violation of its academic honesty policies. Any suspected cheating on examinations will be referred to the Dean's Discipline process, possibly resulting in course failure or College dismissal.

Exams

In general, **NO MAKE-UP EXAMES** are granted. Make-up exams will be given only in rare cases of emergency. If an emergency occurs on the exam day, you must contact the instructor *before* the exam (or arrange for someone else to do so). We will not approve any exam reschedueling requests based on personal reasons such as travel, leisure, or to ease exam week schedules. We will not approve any exam rescheduleing requests for students who take another class whose lectures or final exam occur at the same time as those of our class. No make-up exams will be granted to a student who contats us after the exam is over. No special accommodations will be made for students who arrive late to exams, regardless of the reason (missing a bus; overslept; sick; etc.). If you need to miss an exam due to a sudden severe illness, injury, traumatic event, etc., after consulation with the instructor it is possible that you will be given an **Incomplete** in the course and asked to complete the course in a future semester.

Assignments

There are eight assignments in total. According to the tentative schedule, all of them will be due on Thursday. Please submit your homework electronically from the course web site. Note that we

DO NOT accept late homework. However, the lowest score will **NOT** be counted in your final grade.

Tentative Schedule

Lecture	**Assignment**	**Lab**
09/07: Introduction to R and RStudio. Working with data in R.		
09/14: Working with data in R continued including: data frames, iterative coding.		Lab 1
09/21: R base graphics. Linear algebra review. Multiple linear regression. Bootstrap procedure.	HW 1 due	Lab 1
09/28: Character strings. Regular expressions. Web scrapping.	HW 2 due	Lab 2
10/05: Writing functions. Basic classification methods.		Lab 2
10/12: Split/Apply/Combine.	HW 3 due	Lab 3
10/19: Midterm.		Lab 3
10/26: Tidyverse and ggplot.	HW 4 due	Lab 4
11/02: Random number generation. Simulation. Monte Carlo integration.	HW 5 due	Lab 4
11/09: Simulation continued.		No labs
11/16: Distributions as models.	HW 6 due	Lab 5
11/23: Thanksgiving, no class.		
11/30: Optimization.	HW 7 due	Lab 6
12/07: Optimization continued. Logistic regression.	HW 8 due	Lab 6