STA 5820: Introduction to Data Science
Winter 2023

CRN:	26129 (Sec 001)
Instructor:	Kazuhiko Shinki
E-mail:	shinki@wayne.edu (OR seikibunpu@gmail.com)
Course Web:	We use the Canvas.
Lectures:	M/W 6:30-7:45p in Zoom (Online synchronous throughout the semester)
Office Hours	M/W 5:30-6 pm (in Zoom only); Tu/Th 2:45 -3:30 pm (1189 FAB and in Zoom).

Prerequisites

STA 5030 or STA 5800. STA 5030 is recommended. Or consent of the instructor.

Previous experience in R language is not required, but students who have not used it before may go over Chapters 1-6, 11 and 13 of STA 5030 slides as needed. The course may briefly review these materials, depending on the student body.

Overlap with other courses

MAT 5890 (Introduction to Data Science) in Winter 2019 and 2020 was the same course as this course (while some materials are added). Do not register this course if you have taken the course before.

DSA 6000 also has similar course materials as this course as well. Consult with your advisor if you plan to count both courses for graduation.

Lectures

Attend all lectures and turn on your video, unless special reasons. I look forward your active participation.

- The link to lecture is:
  `https://wayne-edu.zoom.us/j/95486778388?pwd=dWFPOXZ1a2dqbUt0Qnc5bXR4RkF4UT09`
  Meeting ID: 954 8677 8388. Passcode: 776754. One tap mobile:
  `+13092053325,,95486778388#,,,,*776754# US`.
- Recorded lecture will be uploaded to:
  `https://www.youtube.com/channel/UCn8RuLL52FqHzc6puHhgMxw/playlists`

Office hours

- Monday/Wednesday 5:30-6 pm (Zoom only)
  `https://wayne-edu.zoom.us/j/96885257970?pwd=MFkwOUdtM04wTXZHYjRDVmxBN3ZNdz09`
  Meeting ID: 968 8525 7970. Passcode: 546132. One tap mobile:
  `+13017158592,,96885257970#,,,,*546132# US`
- Tuesday/Thursday 2:45-3:30 pm (1189 FAB and in Zoom)
  `https://wayne-edu.zoom.us/j/99164473556?pwd=VEVPRXFRY2tWcTVJQkVFbmprQ0FTdz09`
  Meeting ID: 991 6447 3556. Passcode: 709090. One tap mobile:
  `+16468769923,,99164473556#,,,,*709090# US (New York)`

Textbook

"Introduction to Statistical Learning with Applications in R (2nd edition)," G. James, D. Witten, T. Hastie, R. Tibshirani, Springer. Download the book at `https://statlearning.com`, or get a printed version.

Computer environment

The course requires a computer (Windows, Mac OSX or Linux) with the R language and the RStudio installed. **Download and install R and RStudio into your computer at your earliest convenience**.

R is an open-source statistical language available for Windows, Mac OS X and Linux. It is available on the Comprehensive R Archive Network (CRAN) website at `http://cran.us.r-project.org/`.

RStudio is an integrated development environment (IDE) for R. It is available on the RStudio Website at `https://rstudio.com/products/rstudio/`. To knit PDF files in RStudio, installing LaTeX (a software system for document preparation) before installing RStudio helps. For Mac OSX users, install MacTeX beforehand.

Course outline

This is a 3 credit hour introductory course of applied statistical learning, designed for upper level undergraduate students and graduate students in mathematics and other quantitative fields.

The course covers statistical models which are more advanced than traditional ones such as linear and logistic regressions. It responds to surging demands for more advanced statistical models in science, engineering and business in recent years.

Topics includes: bias-variance trade-off, regression, classification, cross-validation, bootstrap, model selection, regularization, splines, generalized additive models, tree-based methods, support vector machines, principal component analysis and clustering. As time permits, neural network may be covered.

The course consists of three components: statistical theory, computing and data analysis. R and RStudio will be used as the tool for computing and data analysis. Each student will implement their own data analysis project and make a final presentation in class.

Learning outcomes

As a result of mastering the material in this course, you will be able to

- understand concepts and mathematical foundation of statistical learning models,
- use the R language to apply the statistical learning models, and
- choose and implement appropriate models for real-world data analysis problems.

Quizzes and homework

3-5 quizzes will be assigned every lecture in Canvas and due in 2-3 days. It requires understanding of concepts as well as computer implementation.

Homework will be assigned 2-4 times a semester and one week prior to due date. Work independently. Submit your work as a PDF or MS-Word file created by RStudio (or a format mimic to it if you have trouble to create a file by R Studio). Full credits are given only when your work has a professional-level quality. Work not complying with the specified format (e.g., code without outputs & figures) will not be graded.

Midterm exam

One take-home exam. Work independently. The problems will be given one week prior to the due date.

Final Project

The final exam is replaced by final projects. Implement a data analysis project using R on real data of your choice, make a 12-minute presentation in Zoom (Wed. 4/19, Mon. 4/24 and Mon. 5/1), and submit your presentation slides. The presentation schedule will be decided after the Midterm.

Submit a project plan by Friday, March 31. The project plan is a 1-2 pages summary of your data and analysis plan. Attach data sample in additional page(s) or as a CSV/Excel file. The plan will be returned with comments. A more detailed guidance will be given later.

Grading

Quizzes and homework (37%), midterm (24%), final project plan (3%), and final project (36%).
No show in presentation will not get course credits. At least, following final grades are guaranteed based on your weighted score.

Grade	A	A-	B+	B	B-
Score	$[80, 100]$	$[75, 80)$	$[70, 75)$	$[65, 70)$	$[60, 65)$

The score for the final project is 100 at maximum. Outstanding presentations (up to 35% of projects) will get 90-100 points; good presentations will get 80 points (a vast majority is good in the past); good presentations with some weakness will get 60-79 points; presentations with serious weakness will get 0-59 points.

Important dates

Date	Schedule
2/24 (F)	Take-home midterm due
3/13-17	Spring break (no classes, no office hours)
3/31 (F)	Final Project Plan Due
4/19 (W), 24 (M)	Final Presentations
5/1 (M)	Final Presentations (**6:30-8:45pm**)

References

Students who are not familiar with R can refer the following books. Most of them can be found free on the Internet.

1. Michael J. Crawley, **"The R book (2nd edition),"** Wiley.

2. Julian J. Faraway, **"Practical Regression and Anova using R,"** available at `cran.r-project.org/doc/contrib/Faraway-PRA.pdf`.

3. Julian J. Faraway, **"Linear Models with R,"** Taylor & Francis.

4. Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining, **"Introduction to Linear Regression Analysis (5th edition)"**, Wiley.

5. Peter Dalgaard, **"Introductory Statistics with R (2nd edition)"**, Springer (ISBN: 978-0-387-79053-4). (the STA 5030 textbook).