STAT 4700/5030

Data Analytics and Statistical Computing

Spring 2024 Syllabus

Time and Location:

- Mon/Wed 10:45 – 11:15 am and 1:45 - 3:14 pm
- JMHH F90 1/22/2023 – 5/1/2023
- Note: Instructor away 2/12/23 and 2/14/23, recorded content will be uploaded to Canvas.

General Content:

- This course provides an introduction to programming for non-programmers in the widely used R language.
- The course also presents data visualization, data wrangling, stochastic simulation, basic statistical inference and regression, statistical inference based on simulation and the bootstrap, numerical optimization, and machine learning using R.
- Learnings outcomes. By the end of the course you should be able to:
    - Write/debug basic programs in R
    - Use these to develop your own estimators if those you need are not available.
    - Develop simulation or bootstrap methods to assess the accuracy of uncertainty estimates or provide uncertainty estimates for your results.
    - Use tools in the tidyverse and ggplot to efficiently clean, organize and display data.

Prerequisites:

- STAT 111/112 or STAT 101/102 or STAT 431 or STAT 613 or ECON 103/104.
- No programming experience is required. This may not be the best course for experienced programmers since they can easily pick up R using online resources.

Organization:

- Instructor: Giles Hooker
    - E-mail: ghooker@wharton.upenn.edu
    - Office Hours: Thursday, 2- 4 pm or by appointment
- TA: Jinho Bok
    - E-mail: jinhobok@wharton.upenn.edu
    - Office Hours: Tuesday 1:30 – 3:30

- A number of graders will also be available for office hours before homework is due. Details to be posted.
- Canvas will be used to post announcements, lecture slides, R scripts, R markdown files and html renderings, data sets, homework, online quizzes, and take-home exams. The materials will be organized in weekly modules.
- Ed Discussion will be used for online discussions.

Course Materials:

- Textbook: None but see resources below.
- Lecture Slides (to be posted on Canvas)
- The open-source R statistical software and the R studio IDE (Integrated Development Environment) for R (https://www.rstudio.com/products/rstudio/download/)
- MAC Users: If RStudio does not work for you, you may have to download the raw version of R from https://cran.r-project.org
- R scripts, R markdown files and html renderings, and data sets (to be posted on Canvas)

Course Requirements and Grading:

- 20% Midterm Exam (in class)
- 30% Final Exam (in class)
- 40% Homework
- 10% In class quizzes

Quiz and Exam Policies:

- Quizzes will be given on most Mondays in class (missing Jan 22, Feb 12).
- Quizzes are to be taken in class unless otherwise specified.
- To be completed individually, but you are free you use any of the course materials or online sources but may not use Generative AI tools.
- Grades will be assigned based on 8 best results. No make-ups.

Exams will be open-book but closed-internet. You may take in any printed materials you feel you need – summary notes will also be provided to everyone.

Homework Policies and Collaboration:

- Homework should be completed in Rmarkdown format. You will need to submit your rendered pdf file on Gradescope and your Rmd source file on Canvas (instructions and link to be provided later). However, only the pdf file will be graded using the Gradescope platform. Hence, you need to ensure that you can knit your Rmd file to a pdf or at least to an html file. If you render it as an html file, you will need to print it to a pdf file before you can upload it to Gradescope.

- You are expected to complete and turn in your assignments on time. Points will be deducted for late work and homework that is three days late will no longer be accepted unless there is a special circumstance (e.g., illness or family emergency) that prevented you from submitting your assignment on time. Please notify me by e-mail (ghooker@wharton.upenn.edu) to request an extension without penalty.
- You may discuss the homework with other students. However, you are expected to prepare your homework by yourself (e.g., write your own code). In addition, you are expected to acknowledge any collaboration by including a sentence of the following form at the beginning of each question: "I worked on this question with _____." Verbatim copying of another student's code or failure to acknowledge collaboration may result in a zero for the assignment.

Generative AI (AKA ChatGPT) Policies

Generative AI is an important tool in coding and this class will include using ChatGPT. However, it is not (yet) so flawless that you can use it uncritically. Therefore:

- You may use ChatGPT or any other generative AI assistant unless otherwise stated. It also makes another excellent source of explanations, although you should double-check what you understand from it.
- Some homework questions will bar the use of these assistants; complying with this will help you develop skills that will be tested in exams.
- Some homework questions will specify the use of ChatGPT 3.5 (freely available); please respect this for these questions, but note that ChatGPT 4.0 (subscription required) is significantly more reliable.
- If you use a Generative AI aid for a question, you should briefly note this and describe how you used it (e.g. provided question as a prompt and used the code it produced as a starting point).

Classroom rules of conduct:

- Be respectful and helpful to fellow students. Helping your classmates debug their code will also help you become a more proficient programmer.
- Please use your laptops or tablets only for purposes related to our class. No email, texting, and social media during class time.

Students with Disabilities:

- If you have a documented disability, please contact me as soon as you can, preferably within the first two weeks of class.

Resources:

Wickham + Grolemund: R for data science mostly about data wrangling and structures.

Irizarry: Intro to data science focusses more on statistical models and visualization, but also provides some material on wrangling data and programing.

Scott: A Gentle Introduction to Data Science goes a little more into inference and the bootstrap.

Thulin: Modern Statistics with R covers most topics at a slightly more advanced level.

Towards Data Science is a good source for high-level explanations of statistical methods and ideas.

Course Content: We will cover the following topics and demonstrate how to perform various computing and data analysis tasks using R:

I. Introduction to R Programming
   a. Syntax and atomic data types
   b. Data structures (vectors, matrices, arrays, data frames, lists, factors)
   c. Numerical and graphical summaries of data, empirical distributions
   d. Writing your own functions
   e. Flow control and iteration (conditionals and loops)

II. Introduction to Simulation
   a. Random number generation
   b. Probability simulations and Central Limit Theorem demo
   c. Monte Carlo integration

III. Data Visualization
   a. Grammar of graphics of the ggplot2 package
   b. Choropleth maps and heat maps
   c. Time series plots
   d. Visualization of multivariate data

IV. Data Wrangling
   a. Grammar of data wrangling using the dplyr package
   b. Using the pipe operator as an alternative to function composition

V. Review of Basic Statistical Inference

VI. Review of Simple and Multiple Linear Regression Models

VII. More on Stochastic Simulation
   a. Monte Carlo methods in inference
   b. The bootstrap
   c. Permutation tests

VIII. Introduction to Numerical Optimization Using R
   a. Overview of optimization in Statistics and Machine Learning

b. R packages for Linear Programming and Nonlinear Optimization
c. Application to Numerical Maximum Likelihood Estimation and Simulation-Based Optimization

IX. Overview of Machine Learning methods in R
a. Predictive inference
b. Representing text and image data
c. Dimension reduction.