DATA 2020 Syllabus

Instructors:
Roberta De Vito
Email: roberta_devito@brown.edu
Office Hours: Fridays 12:30-1:30 DSI

Teaching Assistants:
Kun Meng
Email: kun_meng@brown.edu
Office Hours: TBD DSI

Amy Liu
amy_liu1@brown.edu
Office Hours: TBD DSI

Lectures: Tuesdays and Thursdays 2:30 - 3:50 pm, CIT 227
Labs: Tuesdays 4:00 - 5:00 pm, CIT 227

Course Goals

This course provides a modern introduction to inferential methods for regression analysis and statistical learning, with an emphasis on application in different contexts and settings. Topics will include the basics of linear regression, variable selection and dimension reduction, and approaches to nonlinear regression. Extensions to other data structures such as longitudinal data and the fundamentals of causal inference will also be introduced. At the end of the course, students should be able to:

1. Describe the statistical underpinnings of regression-based approaches to data analysis.
2. Use R to implement basic and advanced regression analysis on real data.
3. Develop written explanations of data analyses used to answer scientific questions.
4. Provide a critical appraisal of common statistical analyses, including the choice of method and assumptions derived from it.

Course Logistics

Prerequisites: You should be comfortable with the concepts seen in DATA 1010.

Course Website (Canvas): The canvas site will contain all the information for this course including this syllabus, office hours, homework assignments and solutions, and posted grades. Make sure that you are enrolled on the course site and that you check the site regularly.

Reading: Most of the course will adhere closely to the material in the first book listed below. Additional topics may be covered with corresponding notes and references provided. For those interested in advanced material, the recommended texts can be consulted.

(Required) James G, Witten D, Hastie T, Tibshirani R (2013). *Introduction to Statistical Learning, with Applications in R*. Springer. `http://www-bcf.usc.edu/~gareth/ISL/index.html`

(Required for EDA) Irizarry, RA (2018). *Introduction to Data Science* . `rafalab.github.io/dsbook`

(Required for Generalized Linear Model) Dunn P.K., Gordon K.S (2018). *Generalized Linear Model With Examples in R*. Springer. `https://link.springer.com/book/10.1007/978-1-4419-0118-7`

(Required for Missing Data) Little R. J. A., Rubin D. B. (2019). *Statistical Analysis with Missing Data*. Wiley. https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119013563

(Required for Causal Inference) Hua H., Pan W., Ding-Geng C. (2016). *Statistical Causal Inferences and Their Applications in Public Health Research*. Wiley.

https://link.springer.com/book/10.1007%2F978-3-319-41259-7

(Recommended) Hastie T, Tibshirani R, Friedman J (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition)*. Springer.

(Recommended) Morgan SL, Winship C (2015). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press.

Course Assessment

Your course grade will be based on homework sets, two exams, and a final project:

Homework	20%
Midterm Exam	30%
Final Exam	30%
Final Project	20%
Total	100%

Homework Students will be asked to complete assignments approximately every two week. Assignments will consider theory-based, written assignments, and data analysis exercises. When writing your homework, be sure to show your reasoning. Correct and well-written reasoning is as important as a correct solution.

All assignments will be due at 11:59 pm on their due date and must be submitted through canvas. The two lowest homework grades will be dropped, and no late assignments will be considered.

Exams The midterm exam will be March 17th, and the final exam date will be May 14th at 2 pm. Requests for extensions on exams will be considered only if accompanied by a written memo from a Dean, Health Services, or SEAS.

Final Project Students will be asked to address a specific question from a dataset chosen among the ones provided. The project will require students to submit a report that contains (i) formulation of the question at hand in terms of a statistical objective; (ii) description and justification of methods used for analysis; (iii) visual and numerical summaries of key components of the analysis; and (iv) an expository section providing interpretation of results in context. More information to come.

Credit Hours and Time Expectations Over 14 weeks, students are expected to spend 4 hours per week in class and lab, 2 hours on assigned reading, and 3 hours on homework assignments (126 hours total). Preparation for the midterm exam, final exam, and data analysis project is expected to take 54 hours, for an overall total of 180 hours.

Additional Information

Academic Integrity You are expected to maintain the highest level of academic integrity in the course. Any violation will be penalized according to the Brown academic integrity code: https://www.brown.edu/academics/college/degree/policies/academic-code. In particular, sharing materials outside of this class will be treated as an academic code violation.

You are encouraged to brainstorm with other students in this course, but **all assignments must be written up alone without reference to any group work**. For example, if a group works

out a problem together on a whiteboard, each student must write the homework up alone from scratch. Using someone else's code is considered a violation of Brown's Academic Code. If you use code chunks to implement specific tasks for a data analysis, you must properly attribute them.

Public Health Competencies (Masters Level)

Demonstrate a foundation in statistical theory and methods for standard designs and analyses encountered with biomedical data. Identify and implement statistical techniques and models for analysis of data. Acquire knowledge and skills in research methodologies to collaborate with substantive investigators. Understand the advantages and disadvantages of randomized and non-randomized studies to measure effects of interventions. Apply programming skills to analyze data and develop simulation studies. Develop proficiency in making oral, written and poster presentations of work to statistical and non-statistical colleagues.

Accessibility and Accommodations

Brown University is committed to full inclusion of all students. Please inform us early in the term if you have a disability or other conditions that might require accommodations or modification of any of these course procedures. You may speak with me after class or during office hours. For more information, please contact Student and Employee Accessibility Services at 401-863-9588 or SEAS@brown.edu. Students in need of short-term academic advice or support can contact one of the deans in the Dean of the College office.

Schedule

Week	Lab/Lectures	Reading	Assignments
Week 1	Lecture 1(1/23): What is Statistical Learning?	ISL Ch. 1	
Week 2	Lecture 2 (1/28): Exploratory Data Analysis I Lab 1 (1/28): Introduction to R Lecture 3 (1/30): Exploratory Data Analysis II	ISL 2.1-2.2 IDS Ch. 1-2 IDS Ch. 10-11	**Assignment 1 Due (2/7)**
Week 3	Lecture 4 (2/4): Linear Regression Basics Lab 2 (2/4): EDA Lecture 5 (2/6): Linear Regression More Advance	ISL Ch. 3.1-3.2 IDS Ch. 7-8 ISL Ch. 3.3-3.4	
Week 4	Lecture 6 (2/11): Building the "best" model Lab 3 (2/11): Linear Regression Lecture 7 (2/13): Cross Validation and Bootstrapping	ISL Ch. 6.1 ISL Ch. 5	**Assignment 2 Due (2/21)**
Week 5	Lecture 8 (2/20): Logistic Regression Basics	ISL Ch. 4.1-4.2	
Week 6	Lecture 9 (2/25): Logistic Regression More Advance Lab 4 (2/25): Logistic Regression Lecture 10 (2/27): Generalized Linear Models	ISL Ch. 4.3 GLM Ch. 8.1-8.2-10.1	**Assignment 3 Due (3/6)**
Week 7	Lecture 11 (3/3): Automatic Model Selection Lab 5 (3/3): Model Building Tools Lecture 12 (3/5): Nonlinear Models	ISL Ch. 6.2 ISL Ch. 7.4-7.5	
Week 8	Lecture 13 (3/10): Missing Data I Lecture 14 (3/12): Tree-based Methods	SAMD Ch.1 ISL Ch. 8.1-8.2	
Week 9	**In-Class Midterm (3/17)** Final Project Discussion (3/19)		
Week 10	Lecture 15 (3/31): Support Vector Machines I Lecture 16 (4/2): Support Vector Machines II	ISL Ch. 9.1-9.2 ISL Ch. 9.3-9.4	**Assignment 4 Due (4/10)**
Week11	Lecture 17 (4/7): Longitudinal Data I Lecture 18 (4/9): Longitudinal Data II	LDA Ch. 1.1-1.2 LDA Ch. 1.3-1.4	
Week 12	Lecture 19 (4/14): Unsupervised Learning I Lecture 20 (4/16): Unsupervised Learning II	ISL Ch. 10.1-10.2 ISL Ch. 10.3	**Assignment 5 Due (4/24)**
Week 13	Lecture 21 (4/21): Casual Inference I Lecture 22 (4/23): Casual Inference II	SCI Ch. 1-2 SCI Ch. 3-4	
Week 14	Lecture 23 (4/28): Open Review Session		
Week 15	**Final Exam (5/14 2pm)**		