DATA 2020 Syllabus

Instructors:
Roberta De Vito
Email: roberta_devito@brown.edu
Office Hours: Thursday 16:00-17:00

Teaching Assistants:
Zhe Huang
Email: zhe_huang1@brown.edu
Office Hours: Tuesday 7 - 9 pm
Lei Shiqi
Email: shiqi_lei@brown.edu
Office Hours: Monday 7 - 9 pm

Ruya Kang
ruya_kang@brown.edu
Office Hours: Fridays 3:30 - 5 pm

Lectures: Tuesdays and Thursdays 10:30 - 11:50 am

Course Goals

This course provides a modern introduction to inferential methods for statistical learning and regression analysis, 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. Also, a knowledge of the R software is advisable (here the link for a nice class for R).

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) Gelman A, Hill J (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

(Highly Recommended for EDA) Irizarry, RA (2018). *Introduction to Data Science* . `rafalab.github.io/dsbo`

(Recommended 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`

(Recommended 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`

(Recommended 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, and a final project:

Homework	30%
Midterm Exam	20%
Final Exam	20%
Final Project	25%
Participation	5%
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.

You are allowed 3 late days (total) for homework assignments. If you want to use a late day(s), you need to let the Instructors and TAs know at the time the assignment is due and after you turn the assignment in. If want to request additional time due to some sort of personal emergency, you should directly contact the course instructor BEFORE the assignment is due. Assignments handed after their deadlines will score a zero.

Exams The midterm exam will be March 11, and the final exam date will be April 23. Both midterm and final exams will be online. More information will be available later. 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 analyze a dataset chosen among the ones provided (in the folder FinalProject_datasets). The students will have to be organized by themselves in groups of 4 and communicate both the dataset chosen and the composition of the group to the instructor asap. The analysis performed by each group has to be detailed and including at least one method presented during the course. Each group will present its analysis 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; (iv) an expository section providing interpretation of results in context; and (v) contribution of each member of the

group. The draft presentation files from each group will have to be submitted to the instructor at 11.59 pm of April 11. The presentations from each group will be either on April 13 or 15 (the order will be randomly chosen), for the duration of 10 minutes with 5 minutes of questions from the instructor and other students (online presence is required for both dates).

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(1/21): Syllabus and Overview	ISL Ch. 1	
Week 2	Lecture 2 (1/26): Exploratory Data Analysis I Lecture 3 (1/28): Exploratory Data Analysis II	IDS 7-9 IDS Ch. 10-11	**Assignment 1 Due (2/9)**
Week 3	Lecture 4 (2/2): Linear Regression Basics Lecture 5 (2/4): Linear Regression Advance	DAUR Ch. 3 DAUR Ch.4	
Week 4	Lecture 6 (2/9): Building the "best" model Lecture 7 (2/11): Cross Validation and Bootstrapping	ISL Ch. 6.1 ISL Ch. 5	**Assignment 2 Due (2/23)**
Week 5	Lecture 8 (2/18): Logistic Regression Basics	DAUR Ch. 5.1-5.4	
Week 6	Lecture 9 (2/23): Logistic Regression Advance Lecture 10 (2/25): Generalized Linear Models	DAUR Ch. 5.5-5.9 DAUR Ch. 6	**Assignment 3 Due (3/9)**
Week 7	Lecture 11 (3/2): Nonlinear Models Lecture 12 (3/4): Casual Inference I	ISL Ch. 7.4-7.5 DAUR Ch. 9	
Week 8	Lecture 13 (3/9): Open Review Session **On-Line Midterm (3/11)**		
Week 9	Lecture 12 (3/16): Casual Inference II Lecture 15 (3/18): Unsupervised Learning I	DAUR Ch. 10 ISL Ch. 10.1	**Assignment 4 Due (3/30)**
Week 10	Lecture 16 (3/23): Unsupervised Learning II Lecture 17 (3/25): Unsupervised Learning III	ISL Ch. 10.2 ISL Ch. 10.3	
Week11 `	Lecture 20 (3/30): Tree-based Methods Lecture 18 (4/1): Longitudinal Data I	ISL Ch. 10.3 DAUR Ch. 11-12	**Assignment 5 Due (4/13)**
Week 12	Lecture 19 (4/6): Longitudinal Data II Lecture 20 (4/8): Open Review Session	DAUR Ch. 13-14	
Week 13	Final Project Presentation (4/13) Final Project Presentation II (4/15)	ISL Ch. 8.1-8.2	
Week 14	**Final Exam (4/23 9am)**		