DATA 2020: Probability, Statistics, and Machine Learning

Instructors:

Joseph Hogan
Carole and Lawrence Sirovich Professor of Public Health
Professor and Chair of Biostatistics
jwh@brown.edu

Christopher Schmid
Professor of Biostatistics
christopher_schmid@brown.edu

Alice Paul
Postdoctoral Fellow in Data Science and Biostatistics
alice_paul@brown.edu

Teaching Assistants:

Jiabei Yang
PhD Student in Biostatistics
jiabei_yang@brown.edu

Office Hours:

Jiabei Yang (Lab): Tues 4-5, SPH 636
Alice Paul: Wed 12-1, DSI

Course overview

This course provides a modern introduction to inferential methods for regression analysis and statistical learning, with an emphasis on application of the methods in practical settings. Regression methods are developed in the context of learning relationships from observed data. Methods include basics of linear regression, variable selection and dimension reduction, and approaches to nonlinear regression such as polynomial regression, spline-based methods, and generalized additive models. Extensions to multilevel data structures such as clustered and longitudinal data are also described. Fundamentals of causal inference are introduced, together with statistical methods for inferring causal relationships in different study designs.

Course objectives

At the end of the course, students should be able to do the following:

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 in context.

4. Formulate scientific questions in terms of causal quantities, and select an appropriate statistical method for inferring causal relationships.

5. Provide a critical appraisal of common statistical analyses, including choice of method and assumptions underlying the method.

Pre-requisites

This course is designed for students in the Data Science masters program; as such DATA 1010 and 1030 are required. Nearly all data analysis examples described in class will use R.

Readings

The course will adhere closely to material in the first two books listed below. For those interested in advanced methods of statistical learning and causal inference, the recommended texts can be consulted. Journal papers and supplemental reading will be assigned as needed.

(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.
`http://www.stat.columbia.edu/~gelman/arm/`

(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.

Method of evaluation

Homework assignments (40%) Students will be asked to complete assignments approximately every two weeks. Each assignment will combine written answer and data analysis exercises.

Take-home exams (60%) Two data-analysis projects will be assigned, one at the mid-term (20%) and one at the end of the term (40%). For each project, students will be asked to address a specific question from a dataset supplied by the instructors. Each exam 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.

Policy for handing in assignments. All assignments must be submitted online by the due date. Late assignments will be considered on a case-by-case basis. Requests for extensions on exams will be considered only if accompanied by a written memo from a Dean, Health Services or SEAS.

Academic code. All students must familiarize themselves with the Academic Code.

`https://www.brown.edu/academics/college/degree/policies/academic-code`

Students may discuss content of assignments but are not to work in groups in preparing solutions. Each student must prepare his or her own assignment. Using someone else's computing code is considered a violation of Brown's Academic Code. If you use code chunks to implement specific tasks in the course of a data analysis, you must properly attribute them.

Credit hours and time expectations

Over 14 weeks, students are expected to spend 4 hours per week in class and lab, 3 hours on assigned reading, and 3 hours on homework assignments (140 hours total). The midterm project is expected to take 15 hours, and the final project 25 hours, for an overall total of 180 hours.

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 me 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.