STAT 139: Introduction to Linear Models

Fall 2023

Meeting Time: Tues/Thurs 1:30-2:45 PM (Science Center Hall D)
Fri 1:30-2:45 PM (Science Center Hall E)
Course Site: https://canvas.harvard.edu/courses/130511

Instructor: James Xenakis jxenakis@g.harvard.edu

Course Description

STAT 139 is an in-depth introduction to statistical methods with linear models and related methods. Topics covered will include group comparisons (t-based methods, non-parametric methods, bootstrapping, analysis of variance), linear regression models and their extensions (ordinary least squares, ridge, LASSO, weighted least squares, multi-level models), basics of machine learning via regression trees and random forests, model checking and refinement, model selection/comparison, and cross-validation. The probabilistic basis of all methods will be emphasized.

Prerequisites

Mathematics 21a and 21b or equivalent, and Statistics 110 (Multivariable Calculus, Linear Algebra, and Theoretical Probability). Statistics 111 (Theoretical Inference) is *highly* recommended; having taken Statistics 104 or 109 will suffice. Concurrently taking Math 21b is allowed.

Textbooks

There will be recommended reading for each class session. While this reading is not mandatory, it is highly recommended as it will enable more effective learning during class. Although not required, the recommended text is:

- Julian J. Faraway. *Linear Models with R, 2nd Edition*. CRC press, 2014.

Other useful texts are:

- Gelman, Andrew, and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge university press, 2006.
- Ramsey, Fred, and Daniel Schafer. *The Statistical Sleuth: A Course in Methods of Data Analysis*. Cengage Learning, 2012.

Computing

This course will heavily use the R language and environment, as well as RStudio and RMarkdown. We will be using the Posit Cloud service, which allows us do all our computing online through a browser. This has many advantages - for example, there will nothing for you to configure, and all you need is a browswer installed on your machine to do all the computing for this course.

General familiarity with R is required for this course, but no other coding experience of knowledge is necessarily (although it will be very helpful). Having taken STAT 111 will provide you with enough background in R. Further, these basic tutorials can behelpful for getting you started or as a basic review:

- DataCamp: https://www.datacamp.com/courses/free-introduction-to-r
- Rstudio.com: https://education.rstudio.com/

If you would like to install the software on your own machine (I do recommend doing this at some point, though it is not required for the class), R and RStudio are freely available for all common operating systems and can be downloaded at http:///www.rstudio.com/.

Learning Objectives

By the end of the course, students should be able to evaluate the strengths and weaknesses of a variety of statistical techniques. Given a dataset, students should be able to:

- state applied hypotheses,
- explore the data using statistical software,
- determine which statistical model may be appropriate,
- apply corresponding inferences,
- check the assumptions behind these tests and models,
- interpret the results of the analysis to draw conclusions about the hypotheses

This course is designed to prepare students for further coursework in Statistics (such as Stat 131, Stat 140, Stat 149, Stat 160, Stat 183, Stat 186, and others) or for drawing conclusions from data in any field.

Learning Environment

Lectures

Lectures will be conducted in a 'semi-flipped' format; new content will be presented for about 45-55 minutes via annotated lecture slides, and the remainder of the class will be an interactive lab, where students will work through handout problems (a mix of conceptual, mathematical and R-based problems) in small groups. Teaching staff will be present to provide support. The class labs are not graded and the solutions will be posted immediately after class.

Labs

Friday labs will consist entirely of an interactive lab, where students will work through handout problems in small groups (with the support of teaching staff). These labs will emphasize applied problems, with a greater focus on R than the short after-lecture labs. Friday labs are also not graded and the solutions will be posted immediately after class.

Sections

Optional, but strongly recommended TF-led sections will be held throughout the course. Sections schedule will be announced on the course website and there will be a range of day/time options to accomodate students' schedules. Sections will be devoted to review of concepts, practice problems, and problem set preparation/questions.

Discussion Forum

Discussions through Ed are highly encouraged: https://edstem.org/. You will be able to link to Ed through the course website.

Office Hours

I will have regular weekly office hours (time and location TBD), where you are free to ask questions about the course material or problem sets. More general statistics questions are also encouraged. In addition, you will be able to schedule one-on-one office hour appointments with me (15 minutes per session). Please do not email me to schedule an appointment; instead, you can directly book an appointment using *Calendly*.

Attendance

Class attendance is not mandatory. Lectures and labs will be recorded (but not sections or office hours) and available on the course website shortly after they have taken place.

Collaboration

You are encouraged to discuss homework with other students, but you must write your final answers yourself in your own worlds. Solutions that are copied or paraphrased from someone else's work are not accepatable and will be treated as Honor Code violations. All computer output you submit must come from work that you have done yourself. You must list at the top of each problem set the names of all other students with whom you worked. All exams must be entirely individual work.

Course Work

Problem Sets

There will be approximately nine problem sets throughout the semester, which will be posted to the course website approximately 7-10 days before they are due. Currently, all problems sets are

scheduled to be due on Fridays (see accompanying calendar). These dates might change depending on how quickly we progress through the course material. All assigments should be submitted on Gradescope by 11:59pm EST on the day they are due. Solutions will typically be posted to the website two days after the due date, and graded homeworks will be returned before the next problem set is due.

While you should attempt to complete all assigned problems, it is possible that only a subset of the problems will be graded. This subset will be: a) the same for all students, b) selected before the graders view any of the submitted problem sets, and c) not announced in advanced. Please always refer to the posted solutions to check your work on the ungraded problems.

Late problem sets will be accepted up to 48 hours after they are due, but will be penalized 25% of the total points per day. That is, if the problem set is worth 100 points, you would lose 25 points per day, where time is measured discretely - a problem set submitted at 12:00am will be considered one day late. But, to give flexibility, each student will be allowed 3 late days throughout the semester, with at most 1 day (24 hours) applied to any one problem set. Late days cannot be applied to take-home portions of midterms (if we have them) or to the final project. In addition, your lowest problem score will be dropped from your homework average.

In the event of serious illness of unexpected extenuating circumstances, I can provide additional flexibility. You will need to provide documentation from University Health Services or your resident dean.

Exams

There will be two midterms exams, the first of which is (tentatively) scheduled for 10/12/23. The second exam will be take-home and is currently scheduled to be due on 11/17/23 (there will possibly be an in-class component as well). There will be no final exam.

Final Project

In lieu of a final exam, a group project will be due during the exam period, which will be based on a data analysis of your choice, and for which the final project will be 6-8 page paper (single-spaced). This project will be due at 5pm on our assigned exam day (TBD).

Grading

Your final raw score for the course will be computed using the weights in the table below and an associated letter grade assigned based on the class distribution. The grades are curved to allow for more challenging exams without the worry of low raw scores automatically translating to low letter grade. You will not receive a grade lower than that associated with the "standard" scale (with standard rounding applied). For example, if your weighted score is 89.5, you will be guaranteed at least an A-, regardless of the distribution of scores. That is, if your 89.5 happens to be the *lowest* score in the class, you will be guaranteed an A-.

Component	Weight
Problem Sets	35%
Midterm 1	20%
Midterm 2	25%
Group Project	20%
Total	100%

Regrade Requests: All regrade requests must be made within a week of receiving your initial score. A regrade request submission consists of a written explanation detailing which questions are the main focus of the regrade request and explaining why your answer merits additional credit. Please submit any regrade requests directly to me.

Accessibility

Harvard College is committed to working with all students. To request accommodation for a disability, please contact the Accessible Education Office (AEO) (https://aeo.fas.harvard.edu/). Advance notice and appropriate documentation (in the form of a Faculty Letter from the AEO) are required for accommodations. Please speak with me by the end of second week of class (9/15/23) so that we can respond in a timely manner. All discussions will, of course, remain confidential.

Generative AI

This course encourages students to explore the use of generative artificial intelligence (GAI) tools such as ChatGPT to gain conceptual and theoretical insights, as well as assistance with coding. Any such use must be appropriately acknowledged and cited. However, the text in any homework responses must be entirely your own, in our own words, reflecting your own understanding and reasoning. Cutting and pasting text generated by GAI will be considered plagiarism, and therefore academic misconduct.

We encourage you to share your experience with these technologies on Ed, especially in relation to how you think it is making you a more effective learner, instances where it might have led you astray (these technologies can and do "hallucinate"), etc.

The use of Generative AI on exams is forbidden.

We draw your attention to the fact that different classes at Harvard could implement different AI policies, and it is the student's responsibility to conform to expectations for each course.