# Statistics 244

## Linear and Generalized Linear Models

Fall 2022

**Lectures:** Mon/Wed, 9:00am–10:15am, Sci Ctr lecture hall B
**Instructor:** Mark E. Glickman, Sci Ctr 605
**E-mail:** *glickman@fas.harvard.edu*
**Course web site:** https://canvas.harvard.edu/courses/104351
**Office Hours:** Wed 4:00-5:00pm on Zoom, Fri 10:30-11:30am on Zoom, or by appointment
**TFs:** Zeyang Jia (*zeyangjia@fas.harvard.edu*), Sean Ty (*seanty@college.harvard.edu*),
RunLin Wang (*rwang3@college.harvard.edu*), and Christy Huo (*rhuo@g.harvard.edu*).

### Objectives and Prerequisites:

This course presents the theory and application of linear and generalized linear models. Topics include ordinary linear models that usually assume a normally distributed response variable, models for binary and multinomial response data, models for count data, quasi-likelihood and compound models for overdispersed data, and an introduction to generalized linear mixed models. The class of generalized linear models contains many models commonly used in statistical practice.

The main prerequisite for this course is Stat 211 or a strong understanding of Stat 111 or a comparable course. The course also assumes you have some background in applied linear models (regression and ANOVA). The linear models portion of the course assumes a linear algebra background, some of which will be reviewed early in the course. If you feel weak in this area, Harvard has a great resource: an internet connection to the MIT OpenCourseWare linear algebra course .

### Comparison with Stat 139 and 149:

This course shares some topics with both Stat 139 (Linear Models) and Stat 149 (Generalized Linear Models), especially the latter. To be more concrete, the material taught in Stat 139 and most undergraduate or masters level regression courses focuses on the application and implementation of linear models. Stat 244, in contrast, will focus on deeper concepts in linear modeling, with an emphasis on linear model spaces and decompositions through projections. It will not be application-focused. I expect that the linear models portion of Stat 244 will have little to no overlap with Stat 139, though it is important as a prerequisite that you have been exposed to linear models. The generalized linear models material in Stat 244, on the other hand, will have a fair amount of overlap with that of Stat 149, though taught at a somewhat deeper level. I am assuming that students may have seen some applied examples before taking Stat 244 of logistic regression or maybe log-linear models, but this background is not required. If you have taken Stat 149, you will definitely experience *deja vu*, as quite a bit of the Stat 244 presentation of generalized linear models is aimed at the use of various models, but more emphasis in Stat 244 will be placed on the mathematical underpinnings. Also, the problem sets for the generalized linear models material will concentrate mainly on methodology, compared to Stat 149 where the problems tended to be application-focused.

Outline of topics:

The following is an outline of material covered in the course.

1. Introduction to linear and generalized linear models
2. Relevant linear algebra, model spaces, orthogonality, estimability
3. Properties of least-squares estimates, projection matrices, orthogonal decompositions, Gauss-Markov theorem
4. Cochran's theorem, general linear hypothesis, least-squares inference, collinearity, diagnostics
5. Computation of least-squares estimates via Cholesky and QR decompositions
6. Exponential dispersion family models
7. Generalized linear models: Model fitting and inference
8. Models for binary and count data
9. Overdispersion, compound models, quasi-likelihood methods
10. Model selection
11. Mixed effects models (time-permitting)

Textbooks:

Agresti A (2015). Foundations of Linear and Generalized Linear Models. Wiley. ISBN-13: 978-1118730034. ISBN-10: 1118730038. (*Required textbook*)

McCullagh P, and Nelder JA (1989). Generalized Linear Models (2nd ed). Chapman and Hall/CRC. ISBN-13: 978-0412317606. ISBN-10: 0412317605. (*Optional reference book*)

The Agresti book and the McCullagh and Nelder book should be on sale at the Harvard Coop, and from online vendors.

Lectures:

Harvard's mask policy currently is that face coverings are optional. I will probably be wearing a mask while teaching.

Office hours:

Your instructor will be holding remote office hours, at least initially, through scheduled Zoom meetings. Your teaching fellows will likely have mixed in-person and remote office hours.

Sections:

One-hour weekly sections will begin the second week of the course. Attendance is not mandatory, but sections will be the place to work through examples, review difficult lecture material, and solve problems.

Discussion platform:

We will make use of the platform "Ed Discussion" for Stat 244 offline discussions. This platform is integrated into Canvas, so you can access Ed Discussion directly on the left panel of the course Canvas page. This is a great tool to have online discussions with other classmates, and ask questions about course material, logistical information, homework problems, and so forth. We will make all course announcements through Ed Discussion, so please plan to pay attention to the discussion on this platform.

Computing:

All course documents, including homeworks, supplementary material, etc., will be available on the course web site. The course lecture material will show some computational examples, but for the Fall 2022 term the homework and exam material will not require you to analyze data. The course project, on the other hand, will likely benefit from your implementing the methods you examine and applying them to actual data (see the course project description).

Homework:

Homeworks are the place to really learn the course material, so please take them seriously. The assignments will be made available on the course web page for you to download. You will be told when the assignment is posted online.

We will use the online tool Gradescope (and not Canvas) as the way in which homeworks are submitted. To submit a homework, navigate to gradescope.com and select "Login with school credentials" and find "Harvardkey" to log in through Harvard. Stat 244 should show up on the dashboard, and from there you can submit assignments. Alternatively, you can access Gradescope as a link from the Canvas course page. An important distinction between Gradescope and Canvas is that Gradescope asks you to assign each problem in the homework to a page in your submission (e.g., problem 3 is on page 5). This step is important to help us grade more easily and efficiently!

We will have a total of six homework assignments that will be due approximately every 2 weeks. The assignments are to be turned in by 10:00pm on the due date by submitting a pdf version to Gradescope. The homework due dates will be

- Homework 1: due Monday, September 19, 2022
- Homework 2: due Monday, October 3, 2022
- Homework 3: due Friday, October 21, 2022
- Homework 4: due Monday, November 7, 2022
- Homework 5: due Monday, November 21, 2022
- Homework 6: due Monday, December 5, 2022

You are free to discuss and work on homework problems with other students, but you should write up your solutions independently (see the collaboration policy statement at the end of the syllabus). Only a sample of problems will be graded for each homework, though you will be provided solutions to all homework problems. To ensure objectivity of homework assessment, homeworks will be graded via anonymous grading on gradescope. Make sure you do not include identifying information on your submitted homework assignment.

The official course policy is that the top five out of the six homeworks will count for your final grade. This means that you are free to submit only five problem sets without penalty. It also means that there will be no late homeworks – if you cannot get an assignment in on time, that will be your dropped homework which will not count for your final grade. However, it is to your advantage to submit all of your homeworks because working through the homeworks will likely result in a better learning experience, which will translate into better exam performances.

Exams:

The course will have two 1.25-hour closed-book exams during the semester. The dates for the exams are:

**First Exam:** Wednesday, October 26, 2022

**Second Exam:** Online exam administered on gradescope - TBA

Course project:

The course project will involve your writing a short paper on a topic related to the course material. The projects can be done individually or by a team of two or three students. The written report should be no more than 10 pages (double-spaced, 12pt font, 1-inch margins) for a 1-person report, and 15 pages for a 2 or 3-person report. The project is due Wednesday, December 7 at 10pm, the final day of reading period.

Following are examples of possible project areas; these are broad topics, and you should focus on a particular aspect of such an area. You are not limited to these topics. You are welcome to discuss with me or the TF any other ideas you might have.

- Survival models
- Diagnostics for checking GLM assumptions
- Generalized additive models
- Parameterized link functions
- Addressing high-dimensional data
- Regression models for Beta-distributed response data

The only topic I ask students to avoid is Bayesian approaches to GLMs, a topic that is covered in Bayesian statistics classes in the Statistics department.

By Friday, November 11, 2022 at 10pm, you should submit a paragraph describing your intended project (including references you are reading). I will plan to provide feedback on your submission.

You will receive a project guidelines document later in the course. The document will explain precisely the details of the project, suggested structure for the report, and grading advice. I will make this document available around the midpoint of the course.

Grades:

Course grades will be determined by the following components, with the weights shown:

| | |
|---|---|
| Homework assignments | 25% |
| First Exam | 25% |
| Second Exam | 25% |
| Course Project | 25% |

Collaboration policy statement:

University policies against plagiarism will be strictly enforced. You are encouraged to (orally) discuss problem sets with your classmates, but each student must write up solutions separately. Be sure that you have worked through each problem yourself and that the answers you submit are the results of your own efforts. You should not copy or paraphrase others' solutions. In addition, you must cite any books, articles, websites, lectures, etc., that have helped you with your work using appropriate citation practices. Similarly, you must list the names of students with whom you have collaborated on problem sets. You also may not share or view another student's homework, or allow another student to view your homework. A good rule of thumb: if a fellow student asks if you would like to discuss a homework problem, we encourage you to say "yes"; if a fellow student asks to see your answer to a homework problem, the answer is "no."