Statistics 244
Linear and Generalized Linear Models

Fall 2016 - Draft Syllabus

Lectures: Tues/Thur, 2:30pm–4:00pm, classroom TBD
Instructor: Mark E. Glickman, Sci Ctr 605
E-mail: *glickman@fas.harvard.edu*
Course web site: `https://canvas.harvard.edu/courses/14638`
Office Hours: Tues 4:00-5:00pm, Fri 10:30-11:30am, or by appointment
TF: Zach Branson (*zbranson@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 smoothing and generalized additive models. The class of generalized linear models contains the models most 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 part of the course assumes some basic linear algebra that will be reviewed early in the course. If you feel weak in this area, you could read overviews on the internet such as the MIT linear algebra lectures at *http://ocw.mit.edu/courses/mathematics* on topics such as vectors spaces.

Outline of topics:

The following is an outline of material covered in the course. Because this is the first time I am teaching this course, I will forgo a lecture-by-lecture outline.

1. Introduction to linear and generalized linear models
2. Relevant linear algebra, least-squares theory, projections
3. Properties of least-squares estimates, collinearity, statistical inference
4. Computation of least-squares estimates via LU and QR decompositions
5. Least-squares diagnostics, robust methods for linear models
6. Exponential dispersion family models
7. Generalized linear models: Model fitting and inference
8. Models for binary data
9. Multinomial response models
10. Log-linear models for count data
11. Overdispersion, Compound models, Quasi-likelihood methods
12. Regularization in GLMs, computation in large data sets
13. Introduction to smoothing and generalized additive models

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*)

Both textbooks should be on sale at the Harvard Coop.

Computing:

You will be expected to perform data analyses as part of your course work, and you will also receive course announcements through e-mail. All course documents, including homeworks, supplementary material, etc., will be available on the course web site.

Homeworks will include occasional computer problems using the statistics computing package R. The course assumes you have had solid exposure to using R. In case you are rusty, several reference guides on R will be placed on the course web site.

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, solve problems, and learn R tips and tricks for implementing the various methods taught in the course.

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 print out. You will be told when the assignment is posted online. We will have a total of six homework assignments that will be due approximately every 2 weeks, typically on Tuesdays. The assignments are to be handed in by 4:00pm on the due date. You may hand in your assignment at lecture, or you may leave it with the course TF by 4:00pm on the due date. 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 on the following page).

Only a sample of problems will be graded for each homework. All homeworks will count toward your final grade. The official course policy is that *no late homework will be accepted*. In return for your timely submission of homework, we will make every effort to return graded homework promptly.

Exams:

The course will have two 1.5-hour exams during the semester. The dates for the exams are:

First Exam: Thursday, October 13, 2016

Second Exam: Course final exam date – not yet scheduled

Both exams will be closed-book.

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 students. The written report should be no more than 10 pages, double-spaced, for a 1-person report, and 15 pages for a 2-person report. The project is due Monday, December 12 at 4pm.

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.

- Bayesian methods for GLMs
- Random effects and hierarchical (multilevel) GLMs
- Survival models
- Diagnostics for checking GLMs
- Generalized estimating equations
- Parameterized link functions
- Addressing high-dimensional data

By Tuesday, October 25, 2016 you should hand in a paragraph outline describing your intended project (including references you are reading) based upon which I can provide feedback. When you are forming your ideas for a project, please feel free to talk with me or e-mail me about it.

The written report should be in your own words; do not simply extract sentences and formulas from articles or books that you read. Clarity of exposition is important. The report should preferably include a data example to illustrate the ideas. Whenever possible, use notation and terminology consistent with that used in the course lectures, as if you were presenting a lecture in the course. Explain all new notation and define all new terms. It is best not to wait until the last minute to work on the project, as it is expected to take a substantial amount of time.

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 also may not share or view another student's computer code, submit output from another student's computer session, or allow another student to view your code or output. 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 or R code, the answer is "no."