Statistics 149
Statistical Sleuthing through Generalized Linear Models

Spring 2016

Lectures: Mon/Wed, 1:00pm–2:30pm, Sci Ctr E
Instructor: Mark E. Glickman, Sci Ctr 605
E-mail: *glickman@fas.harvard.edu*
Course web site: `https://canvas.harvard.edu/courses/10751`
Office Hours: Mon 2:30-3:30pm, Fri 10:30-11:30am, or by appointment
TFs: Zach Branson (*zbranson@g.harvard.edu*), Maxime Rischard (*mrischard@g.harvard.edu*), Callin Switzer (*cswitzer@fas.harvard.edu*)

Objectives and Prerequisites:

Arguably one of the most fundamental problems in applied statistical work involves modeling a response variable from a collection of predictor variables for the purpose of explaining, describing, or forecasting a random phenomenon. Least-squares regression may be the most well-known approach, and was the subject of Statistics 139, but it can be very restrictive and often inappropriate for many situations. For example, when the response variable is not continuous, or when the response distribution is heavily skewed or heavy-tailed, other approaches should be considered. An alternative that addresses such limitations is the class of "generalized linear models" (GLMs), of which least-squares regression happens to be a member. Simply put, GLMs extend the least-squares regression framework by allowing a variety of probability distributions (not just normal) to be assumed for the response variable. This course will explain the basic principles of GLMs, familiarize students with common GLMs, and demonstrate the use of GLMs in a wide variety of data situations. Applied examples from various fields including education, political science, psychology, and biology will be presented.

In addition to GLMs, the course will introduce students to several other approaches to relating a response variable with a set of predictors. These will include models accounting for over-dispersed responses, non-parametric smoothing and generalized additive models, and tree-based models (including random forests).

Stat 149 assumes a background in linear models (e.g., having successfully completed Stat 139) as the main pre-requisite. The course will assume students are comfortable with basic differential calculus and matrix algebra.

Stat 149 vs. Stat 244:

Stat 149 is primarily aimed at upper-level undergraduates and masters students in statistics, as well as graduate students not specializing in statistics, and will focus mainly on the development and application of GLMs and related statistical tools. Stat 244, which is intended primarily for Statistics PhD students, combines the material of Stat 139 and 149 into one Fall semester course, and is intended to be a balanced mix of the application and theory of least-squares regression and GLMs. Students in Stat 244 are assumed to have a more substantial math and statistics background. While Stat 149 will present occasional mathematical details, the course will emphasize the application of the methods.

Outline of topics:

The following is a lecture-by-lecture outline of material covered in the course. No guarantees are being made that we will keep to this schedule with 100% precision, however.

Date	Topic
Jan 25	Course introduction
Jan 27	Intro to maximum likelihood estimation, common probability models
Feb 1	Exponential dispersion family (EDF) and generalized linear models (GLMs)
Feb 3	Binary response models, link functions, logistic regression
Feb 8	Latent variable representation, Inference for logistic regression
Feb 10	Predictive inference
Feb 17	Categorical predictors, contrast coding, odds ratios
Feb 22	Likelihood ratio tests, profile likelihood
Feb 24	Predictor interactions, collinearity
Feb 29	Analysis of deviance
Mar 2	Hosmer-Lemeshow test for lack of fit, diagnostics
Mar 7	Design issues, paired comparison models, introduction to missing data
Mar 9	Midterm exam
SPRING BREAK	
Mar 21	Poisson log-linear models, offsets
Mar 23	Analysis of deviance, multiway tables
Mar 28	Multinomial logit regression, connections to Poisson model
Mar 30	Inference and prediction for multinomial logit regression
Apr 4	Ordinal response models, proportional odds models
Apr 6	Estimation and prediction for ordered response models
Apr 11	Overdispersion, negative binomial and beta-binomial regression
Apr 13	Gamma and inverse-normal models
Apr 18	Smoothers
Apr 20	Additive and generalized additive models (GAM)
Apr 25	Classification and regression trees (CART)
Apr 27	Random forests

Textbooks:

(*Required*)

> Faraway J (2006). Extending the Linear Model with R. Chapman & Hall/CRC. ISBN-10: 158488424X — ISBN-13: 978-1584884248

(*Reference*)

> James G, Witten D, Hastie T, and Tibshirani R (2013). An Introduction to Statistical Learning. New York: Springer. ISBN 978-1-4614-7137-0

The Faraway textbook should be on sale at the Harvard Coop. The James et al. textbook is available for free download at `http://www-bcf.usc.edu/~gareth/ISL/`.

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 computer assignments using a statistics computing package. The official computer package for this course is R, which runs on Windows, MacOS, and Linux systems. The software is free and available online through `www.r-project.org`. You will be taught everything you need to know how to run R. R is straightforward to learn, but is sufficiently powerful and versatile to be useful for projects that you might carry out after this course. Several reference guides on R will be placed on the course web site.

When you run analyses using R, you should either cut and paste the relevant output into a document containing your homework, or print the output and cut the relevant pieces to include on your homework. Given that we are now in the second decade of the 21st century, the former approach is both preferable and easier.

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 how to use the statistics package R 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. Homework assignments will be due approximately every 2 weeks and 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).

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.

Homework assignments will be graded on a scale from 1 to 5. Homeworks are graded in large part on the clarity of your presentation of the solutions, not just their correctness. Homeworks that are generally clear and correct will earn scores of 4 or 5; those less so will earn a 3. Sloppy and/or incomplete homeworks will receive a 1 or 2. All homeworks will count toward your course grade – we will not drop any homework grades. At the end of the course, homework scores are converted into a 0-100 scale. A homework score of 5 is converted to 100; a score of 4 is converted to 90; a score of 3 is converted to 80; a score of 2 is converted to 65; and a score of 1 is converted to 50. Also, a homework score of 0 (not handing in your homework) is converted to a 0 on a 100 point scale. In other words, if x is a homework score, then the converted score y is given by

$$y = \frac{5}{12}x^5 - \frac{35}{6}x^4 + \frac{365}{12}x^3 - \frac{445}{6}x^2 + \frac{595}{6}x$$

Exams:

The course will have an in-class midterm exam during the semester, and a final exam. The dates for the exams are:

Midterm Exam: Wednesday, March 9, 2015

Final Exam: Not yet scheduled

Both exams will be closed-book, with the exception of a single sheet of notes. This sheet can be double-sided. You should plan to bring a calculator to the exams.

Course project:

The course project will involve students working singly, or in groups of up to four students, to develop a statistical model that makes accurate predictions based on a supplied data set. Students will be provided with a subset of a large data set, and asked to develop statistical models using methods learned in the course. The models will then be applied to the remaining portion of the data to assess the predictive accuracy. Students will also be expected to submit a short report documenting their modeling efforts and prediction methods. The project details will be distributed in a separate document.

The prediction exercise will be carried out online through an educational resource connected with the `kaggle.com` web site. Instructions for uploading predictions will be provided. The site will allow multiple prediction submissions as the course proceeds, and the predictive accuracy of each submission on a portion of the withheld data subset will be available for all to see. Students will not be graded on predictive accuracy of their models (as long as the models are more predictive than naive benchmark models), but will be graded on the effort, thoroughness, logic of the methods, and the clarity of their presentation in the reports. The report is an opportunity to demonstrate mastery of material taught in Stat 149. As inspiration to make accurate predictions, the student or student team with the best predictive accuracy will win a prize at the end of the course (to be announced). Can a statistics course project be any more fun?

Grades:

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

Homework assignments	20%
Midterm Exam	20%
Course Project	30%
Final Exam	30%

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 students computer code, submit output from another students 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.”