Syllabus
Linear and Generalized Linear Models
Spring (2015)

Professor: Skyler Cranmer
Office: 2032 Derby Hall
Email: cranmer.12@osu.edu
Office Hours: Wednesdays 0900-1100
Meeting Place & Time: TuTh 1100-1220 Room: Derby 150
Course Web Site: Carmen

TA: Drew Rosenberg
Office 2043 Derby Hall
Email: rosenberg.1108@buckeyemail.osu.edu
Office Hours: 0900-1100 on Thursdays
Recitation: 1100-1200 Wednesdays in the Spencer Room

Course Description

This course builds on what you did in the first course of the methods core sequence and focuses on the concept of regression modeling using the linear and generalized linear models (their proper terms, though for historical reasons people in political science often call them "ordinary least squares" and "maximum likelihood models" respectively; this is odd because those are means of estimating models rather than models themselves, so don't do that). The principle we will care about is how to build a standard regression framework with a linear predictor such that a tremendously broad class of outcome variables and data structures can be modeled. We will consider how to perform regression analysis with the following types of outcome variables: continuous, counts, dichotomous outcomes, ordered categorical outcomes, unordered categorical outcomes, bounded variables, and more. There is a strong theoretical basis for the models that we will use. Also, the bulk of the learning in the course will take place outside of the classroom by watching the video lectures, reading, practicing using statistical software, going to lab, doing problem sets, and studying. Keep in mind that the skills attained in this course are those that the discipline of political science expects of any self-declared empirically oriented researcher (basically, everyone other than theorists). In other words, you should approach the course thinking not about a grade, but about preparing for the rest of your career.

A second theme of the course involves the statistical package `R`. `R` is completely free and works on Mac, Unix, Linux, and even Windows. You can download `R` at CRAN, the Comprehensive `R` Archive Network. `R` is an implementation of the `S` language, which is the default computational tool for research statisticians around the world. Quite simply, `R` is the most powerful, extensively featured, and capable statistical computing tool that has ever existed on this planet. And as mentioned, its free. `R` is a theme in so far as we will cover all models theoretically (math) and then apply them using `R`. So a great deal of the coursework will involve the use of the `R` language. We will not use Stata; don't ask.

Prerequisites

Coursework in probability theory and linear models or my permission. The course also assumes a working knowledge of calculus and linear algebra. For those of you within the department, you need to have taken the first course in the methods core sequence.

Evaluation

Your final grade will be based on two components: weekly problem sets (50%) and two exams (worth 25% each). The problem sets will be a combination of analytical and computational assignments. All quizzes will be in class, closed-book, closed-notes. You should complete the scheduled reading *before the class listed!*

Course Norms

- Teamwork and collaboration is *highly encouraged* on nearly every aspect of the course (e.g. on homework and exam preparation). However, everyone must write out their own homework (no group submissions or just changing the name) and list who they worked with. Obviously, collaboration is not allowed on the quizzes.
- <u>*All*</u> homework assignments must be written in LaTeX. Assignments not written in LaTeX (or `sweave` if you want to be really fancy) will be returned without a grade.

Recommended Texts

Despite a large number of attempts, there is no one book that I think is really good at doing everything. As such, we will use material on electronic reserve from many books. The following may be especially helpful however.

- Gelman, Andrew and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press. *Note: We will only use a few chapters out of this book, but it is one of the best stats books out there. One would very much benefit from reading the rest of it! Further note, someone put a copy on the internet here.*
- Greene, William H . *Econometric Analysis. Note: 5th edition. Someone scanned it and put a pdf here. I will assign readings out of the 5th edition, even though the book is onto its 7th edition (for a whopping $220).*
- Faraway, Julian J. *Linear Models with R* CRC Press. Some dude put it on the web here
- Faraway, Julian J. *Extending the Linear Model with R*, CRC press. *Note: The faraway books are a bit light on the stats theory, but very good treatments of how to do all this stuff in R. As far as 30 seconds with google could reveal, no one has put this book on the net; but this is also the least important book on this list for our purposes.*
- Kennedy, Peter. *A guide to Econometrics*, MIT Press. *Note: This book is very basic, but is a good resource if you are struggling with the theoretical material. Note that someone has apparently scanned the entire book and put it on the interwebs here*
- King, Gary. *Unifying Political Methodology.* Cambridge. Online here

Additional Readings

Come talk to me if the other assigned reading is not working for you.

Tentative Outline

Note well: Some re-working of the below outline is basically inevitable. Homework assignments and required reading will be posted on the Carmen site each week and that is what we will go by officially. The below is just to give you a basic outline of the intended structure.

The logic of the topic progression is as follows: We will spend the first several weeks refreshing some basic concepts and going over maximum likelihood and generalized linear model (GLM) theory in a fairly abstract way. This may seem removed from political science applications, but you'll find that once you have a handle on the general theory, picking up specific models is surprisingly easy. Once we're done with the theory portion of the class, we'll go through a wide variety of GLMs which are standards in the field (including the LM). This segment of the class will be much less theoretical and highly focused on applications in `R` as well as the presentation and interpretation of statistical results. We will then move on to discuss a series of special topics with as much range and detail as time permits. While what follows is a tentative outline, we'll try to stick to it conditional on everyone following along. If we run short on time, we'll drop some of the stuff at the end.

Every week, we'll follow the same basic structure. Tuesdays, we'll discuss the content of the video lectures in an informal style and clear up any questions. We'll also spend some time on "career skills" (like how to pick projects). Wednesday, Drew will hold recitation. Thursdays will be lab days. There will be no formal lecture or anything, but you should have had a chance to (a) read the assigned material, (b) watch the lectures, (c) discuss their content and clear up questions, and (d) start the assignment. During lab, Drew and I will help you through problems you've hit on the assignments. *Come with problems! Do not start the assignment in lab!*. Assignments will always be due at 5pm on Friday.

PART I: THEORY

Week 1 Week of Jan 9
Topic: *Uncertainty and Inference*
Detailed topics:

- Course introduction and outline
- Notation
- Forms of uncertainty
- Probability as a model of uncertainty
- Principles of simulation
- *Recitation: Using LaTeX*

Read before class: King chapters 1-2

Week 2 Week of Jan 16
Topic: *Teaser Introduction to the Basic Linear Model and Least Squares Estimation (+ probability distribution review)*
Detailed topics:

- Basic linear regression
- Estimation via least squares
- Commonly used PDFs/PMFs
- *Recitation: Understanding PDFs/PMFs via simulation in R*

Read before class: Kennedy chapters 2-3. Wikipedia pages on the Normal, Binomial, Poisson, and Exponential distributions.

Week 3 Week of Jan 23
Topic: *Likelihood Theory and its Properties*
Detailed topics:

- Likelihood and Bayesian inference vs Frequentism
- Likelihood theory
- Maximizing likelihoods
- Justifying standard errors (Fisher information)
- Properties of likelihoods and their maxima
- Assessing likelihood models: Wald, Likelihood-Ratio, Deviance, AIC/BIC
- The generalized linear model and the exponential family form
- *Recitation: Fun with calculus! Practice deriving and maximizing common GLMs*

Read before class: King chapter 4; Greene MLE chapter; Pawitan Chapter 2 (Carmen); McCullah and Nelder chapter 2 (Carmen)

Part II: Standard GLMs

Week 4 Week of Jan 30
Return to the Linear Model Part I
Detailed topics:

- The model (again)
- Estimation via LS and ML
- Gauss-Markov Theorem
- Inference
- Prediction
- *Recitation: Linear algebra review (analytic & with R) / matrix operations with R*

Read before class: Faraway Ch 2-4; Greene, Ch 2-3; Kennedy 3-4

Week 5 Week of Feb 6
Assumptions and Diagnostics
Detailed topics:

- Assumptions about the structure of the model
- Assumptions about the data
- Assumptions about the errors

- *Recitation: Better graphics with R*

Read before class: Faraway Ch 6-8; Greene, Ch 2.3; Kennedy 6-10

Week 6 Week of Feb 13
A very detailed example
Read before class: Gelman and Hill Ch. 3-4
Recitation: TBA.
Exam 1 on Thursday

Week 7 Week of Feb 20
Topic: *Models for Binomial Data Part I*
Detailed topics:

- Motivating example
- Logistic regression
- Interpretation (several techniques)
- Latent data formulation
- *Recitation: Better graphics with R part III*

Read before class: Gelman and Hill chapter 5.

Week 8 Week of Feb 27
Topic: *Models for Binomial Data Part II*
Detailed topics:

- A very detailed model construction exercise, including...
- Adding and evaluating predictors
- Interactions
- Centering and standardization
- Transformations
- Fit checking (error rates, Deviance/AIC/BIC
- *Recitation: bibtex*

Read before class: Gelman and Hill chapter 5.

Week 9 Week of Mar 6
Topic: *Count Models*
Detailed topics:

- Poisson regression
- Dispersion and dispersion parameters
- Rate parameters and offsets
- Linear rate models
- Negative binomial (not technically a GLM)
- *Recitation TBA*

Read before first class: Gelman and Hill 6.2 and 6.2.

Week 10 Week of Mar 13
Spring Break. No class.

Week 11 Week of Mar 20
Topic: *Ordered Categorical Models*
Problem Set 6: *Due at the beginning of class*
Detailed topics:

- Ordered logit and probit
- Derivation
- Interpretation
- *Recitation: beamer and its customization*

Read before first class: Gelman and Hill 6.5.

Week 12 Week of Mar 27
Topic: *Unordered Categorical Models*
Problem Set 7: *Due at the beginning of class*
Detailed topics:

- Multinomial GLMs
- Derivation
- Interpretation
- *Recitation: writing functions in R part 1*

Read before first class: Gelman and Hill 6.5.

Week 13 Week of Apr 3
Topic: *Survival Models*
Problem Set 8: *Due at the beginning of class*
Detailed topics:

- Survival analysis
- Survival curves and their estimators
- Parametric survival models (exponential and Weibull GLMs)
- A quick glimpse at the Cox Proportional Hazards Model
- *Recitation: writing functions in R part 2*

Read before first class: Reading assignment to be provided

Week 14 Week of Apr 10
Topic: *Special Topic! Missing Data (not a GLM, but very common and important)*
Problem Set 9: *Due at the beginning of class*
Detailed topics:

- Types of missing date: MCAR, MAR, NI
- Casewise deletion is evil
- The insufficiency of single imputation
- Multiple imputation (parametric, fully conditional, and hot decking)

- *Recitation: Advanced LaTeX*

Read before first class: Gelman and Hill chapter 25. Articles TBA

Week 15 Week of Apr 17
Topic: TBA
Exam 2 on Thursday