Political Science 271–"Advanced Statistical Applications"
(aka Maximum Likelihood)

Lecture/Discussion: Friday, 9-11:50am SSB 104

Instructor: Prof. John S. Ahlquist
email: `jahlquist@ucsd.edu`
Office: 1411 RBC
Office Hours: M 1-3PM (PhD priority); Tu,W 1:30-2:30 (MIA/MPP priority); & by appointment

Course Description

This course introduces PhD students to the theory and practice of likelihood inference for statistical models, as applied to the social sciences. Models include those designed for binary, nominal, ordinal, count, and continuous random variables. We will discuss theory, estimation, interpretation, and presentation of results. Time and student interest permitting, we will also introduce models for spatially and temporally correlated data; survival and event history models; missing data imputation; and model validation and selection.

Motivation

Many, if not most, social science research problems do not easily conform to the standard linear regression model. The likelihood framework provides a powerful set of tools through which to study many social phenomenon while also subsuming the standard linear regression model. Likelihood-based models are now readily incorporated into most statistical packages, so estimation is often trivial. Nevertheless, interpretation and evaluation of likelihood-based models requires care and nuance. The principles of likelihood theory and maximum likelihood estimation are foundational for continued methodological study.

This course serves as an introduction to maximum likelihood methods with an applied social science focus. We will discuss the selection of appropriate statistical methods for a variety of typical research questions in the social sciences while providing the theoretical and computational foundations for maximum likelihood estimation. Our goal is to empower you to develop your own statistical models, tailored to fit your research problems.

The syllabus is intended to provide an overview over the course. You cannot claim any rights from it. In particular, scheduling and dates may change. While the syllabus should be a reliable guide for the course, official announcements are always those made in class by the instructor.

Expectations and Student Responsibilities

I assume that you have the following:

1. recent exposure to basic linear regression and elementary probability theory.
2. a working knowledge of multivariate calculus, especially differentiation and optimization (finding maxima and minima and knowing when you have one or the other).
3. a working knowledge of basic linear (matrix) algebra.
4. some experience with statistical computing in $\mathcal{R}$.

If any of the preceding does not describe you, please come and see me immediately.

I expect you to attend all lectures, to have done the relevant readings, and to complete all assignments in a timely fashion. I also expect you to be self-directed in choosing a topic for your final project, though I am happy to provide feedback on paper topics and suggestions for finding data.

Course Deliverables

- Quasi-weekly Homework
- Replication and extension of an article of interest that focuses on non-OLS techniques
  - Presentation of a poster on **March 17**, detailing your project and your findings
  - 10-15 page writeup of the "Methods" and "Results" sections of a paper on your project. No theory, no introduction, no conclusion, no summary. Due at **noon, March 20.**
- Comment and feedback for colleague on his/her project.
- Establish online archive for sharing your data and work with others.

Grading

Your grade will be composed of the following components:

- 2.5% Office hours visit
- 2.5% data archive set up
- 35% homework assignments. Grading will be assigned based on three categories {fail=1, unacceptable=2, acceptable=3, exceptional=4}. Each is worth four points. An assignment not submitted will receive a 0. The lowest homework grade will be dropped.

- 10% partner comments
- 50% Poster & final write up

There will be no incompletes given in the this class. I will not accept late papers.

Class meetings

We meet weekly. Each week will have a text chapter and a selection of actual applied research articles. I expect that you will come to class meetings prepared, having already read the required texts and developed questions.

Expect to spend the first half of our class meetings reviewing the days readings, answering questions, and sometimes having a short lecture. The second half of class will be more of a lab and tutorial session focused on working through homework problems. You should bring your laptop to class.

A note on reading articles for this class: Each week we will read a selection of research articles applying the tools we are studying that week. In reading these articles will ignore the introduction, literature review, theoretical model, and conclusion. We will focus exclusively on the data, measurement, and modeling sections, including the discussion of results and findings. Come to class prepared to discuss modeling and presentation of results. For this class we don't care about theory.

The course text

The main text for this course is a textbook manuscript that is forthcoming at Cambridge University Press. You get to work with the penultimate, unofficial version of the book for free.

The text is still being refined. You are in a great position to offer helpful suggestions both during and after the course. Please feel free to share any and all feedback you have! More specifically, with probability approaching 1 you will encounter typos or other errors in the manuscript. We are desperate to catch these errors. To that end I am offering a bounty. Here are the rules:

- All errors are fair game: typos/misspellings, grammatical errors, broken links or references, mathematical or notational inconsistencies, mathematical errors, inconsistent references for tables, figures, chapters, etc.
- Only the first person who brings the error to my attention (via email) gets the bounty. Please state page and line. Also state what you believe to be the correction.
- If multiple people bring the error to my attention at exactly the same time (as reported on the email time stamp), then I will randomly pick one to get the bounty.
- The bounties are as follows:

- 1 bonus homework point for each of: spelling/typo/minor grammatical error in main text; broken link; missing citation or reference; minor (math) notational error or inconsistency. Maximum bonus of 4 points.
- 2 bonus homework points for each of: major grammatical error (e.g., incomplete or nonsensical sentence or a section of text that clearly does not belong); spelling or typo in a table or chart; mathematical or other inconsistency in a table or chart; mathematical or notational error in a definition or derivation; typo that breaks computer code. Maximum bonus of 4 points.
- 2 point bonus on your final grade if you show a major error in logic in a proof. You must provide the corrected proof.
- If you send in valid corrections but are not picked because someone beat you to it you will get partial credit as follows: for every 5 valid corrections you send in you will receive a 1 bonus homework point up to a maximum of 3 points. Points are awarded discretely, i.e., only in perfect multiples of 5 valid corrections.
- For every five *invalid* corrections you send in you will lose one homework point (or bonus point). Points are reduced discretely, i.e., only in perfect multiples of 5 invalid corrections.

Homework assignments

There will be periodic ($\sim$7) homework assignments during the class designed to get you to work with the mathematical concepts and real data, producing and interpreting results. All homework MUST be typewritten and submitted electronically in PDF format via the TED course site. You should also include your (documented) $\mathcal{R}$ code as a *separate* `.txt` or `.R` file. You can work in groups on the homework but all write-ups, results, and coding must be individual work.

Homework is due at noon on Thursday each week it is assigned. This will give me a chance to read it in time to cover any items that are necessary before we move on to subsequent topics.

Solutions for all problem sets will be posted. Your lowest homework grade will be dropped.

Reproduction and extension project poster & paper

Your final project for this course will involve reproducing and extending a piece of published, peer-reviewed empirical research. You will identify a published paper (or, less likely, book) that uses some type of model based on maximum likelihood methods. You will acquire the author's data and reproduce the author's original results. You will then think of an extension of this model. This extension could involve updating the data to include newer data; re-estimating the model under different modeling assumptions; altering the assumptions around missing data; and comparing the author's model to other alternatives. Some key points:

- **You must find your chosen article/data and secure my approval for your project by Week 4.**
- **You should have the data and successfully open it in $\mathcal{R}$ by week 5.**
- You will be asked to make progress reports on this research over the course of the quarter, so you should come to class each week prepared to describe your project and where you stand.
- Everyone in class will be assigned a partner. Your partner will function as your shadow co-author and sounding board. Your co-author will provide written comments on one draft of the paper, due to both me and your partner on **Monday, 6 March at 9AM**. You should have your comments back to your partner by 5pm that Wednesday.
- If you wish to combine the final project for this and another class you must receive approval from me and the other instructor.

The final course project has two components: a paper and a poster.

paper

The most familiar part of the project is a final paper. The paper is meant to be a draft of an empirical article to be submitted to a social science journal. This draft should follow the style of an article in the *American Political Science Review*. I want *only* the data analysis section of the paper that describes the (statistical) model, the data, the results, and your conclusions. I do *not* want the introduction, literature review, theory, or motivation. If you are writing the rest of the paper as part of your work for another class that's fine. But I will only grade the data analysis and interpretation sections.

The paper should be no longer than 20 pages double spaced, and ideally it should be a bit shorter. Do not include your references in this total, but do include your tables and graphics. A good summary of how to construct the paper can be found here: `http://gking.harvard.edu/files/paperspub.pdf`

poster

The second portion of your final project is the production of a poster summarizing your research and conclusions. Posters are the most common form of communication at academic conferences in the natural sciences, becoming increasingly important in the social sciences. The last class will be dedicated to poster presentations. Political science faculty and other grad students will be invited to come and see the results of your work. Your poster should conform to the criteria for a PolMeth poster session, as given here: `http://polmeth.rice.edu/poster-session/`. The Department of Political Science has generously offered to help defray the cost of poster printing; more on this later.

Office Hours

I insist that you come talk to me in office hours at least once prior to week 3 of the semester. By this date you will be required to get my consent for your choice of final project. You will be asked to turn in a 1 page summary of your proposed final project, as well as set up an online data archive.

The vast majority of your learning will occur as you struggle with the problem sets and your project. For this reason I will be holding more than the usual number of official office hours for this course. You are strongly encouraged to come and talk to me about any questions or problems you might be having. If you feel lost or that you are falling behind, come see me sooner rather than later so that we can improve your comprehension and the content of labs and lectures. In general, it is a good idea to bring a laptop when you come to OH.

If you have concerns with lectures, the format of the course, or any other issue PLEASE bring them to my attention.

Email & website policy

I will be using email regularly to communicate with the class. You should feel free to contact me via email. I will do my best to respond to email queries within 12 hours during the week and 24 hours on weekends. Please do NOT expect me to be at your beck and call on Sundays.

This course will have a dedicated website. Course notes, assignments, and solutions will be posted regularly on the TED site.

How to Succeed in this Course

This is a course in learning how to fit and interpret a variety of statistical models using maximum likelihood estimation. Each approach will have least two examples presented, one in the text/class and a different one as graded homework. Most all of these will be real, not toy, examples that I have taken from contemporary social science research.

You should not expect to understand everything the first time through. It is unlikely that you will be able to solve the assigned problems in their first pass. Thus, tenacity is crucial. You will get negative feedback. The point is to learn from it, not to stop at it. It is also unlikely you will understand all the concepts and details the first time through but I promise you that as you work these models repeatedly in your own work and as you revisit the notes and readings over time you will come to have a greater appreciation for the material covered in this course.

There is not a great deal of reading for this course (relative to, say, the comparative politics field seminar), but I have assigned readings in two forms. First, there is text material that should help you understand the techniques. Second, there are real research examples of each technique, so that you will see how scholars use these approaches in practice. Carefully

read all of the text readings before class each week. It is important to read the relevant methods and analysis sections of the examples before class, as I will be discussing them in some detail and will refer to them in my handouts and notes. It probably wouldn't hurt to read them again after each class, though that is entirely up to you.

Often times, your course colleagues will have an insight or explanation that far exceeds the instructor's, or even your own. So collaboration is one way to take advantage of these insights among your fellow travelers in this course. It is also a way to share your own. You should feel encouraged to collaborate in this class. However, when you get to it, sit down and do your own homework, as a synthetic testament to how much *you* learned.

Required materials

Books

The core texts for the class are:

1. Ward, Michael D. and John S. Ahlquist. forthcoming. *Maximum Likelihood for the Social Sciences: strategies for analysis.*

2. Ward, Michael D. and Kristian Gleditsch. (updated edition) *Spatial Regression Models.* Sage.

Suggested reference and alternative texts

Statistical theory

1. Cameron and Trivedi. *Microeconometrics.* New York: Cambridge University Press

2. Cleveland, William. 1993. *Visualizing Data.* Hobart Press

3. Eliason, Scott R. 1993. *Maximum Likelihood Estimation: Logic and Practice*, Thousand Oaks, California: Sage.

4. Fox, Jonathan. 2008. *Applied Regression Analysis and Generalized Linear Models* second edition. Sage

5. Greene, William. *Econometric Analysis* sixth edition. New York: Prentice Hall

6. King, Gary. 1998. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*, Ann Arbor, Michigan: University of Michigan Press.

7. McCullagh & Nelder. 1989. *Generalized Linear Models*, Second Edition. Chapman & Hall.

8. Pawitan, Yudi. 2013. *In All Likelihood.* Oxford University Press.

9. Wooldrige, Jeffry M. *Econometric Analysis of Cross Section and Panel Data.* Cambridge: MIT Press

Computation in $\mathcal{R}$

1. Fox, Jonathan. *An R and S Companion to Applied Regression.* Beverly Hills: Sage
2. Dalgaard, Peter. 2002. *Introductory Statistics with $\mathcal{R}$*, New York: Springer Verlag.
3. Venables, William N. and Brian D. Ripley. 2002. *Modern Applied Statistics with S, fourth edition*, Springer Verlag

Math

1. Moore, William and David Siegel. 2013. *A Mathematics Course for Political and Social Research.* PUP
2. Simon & Blume. *Mathematics for economists.* Norton

Computation

$\mathcal{R}$ Software

The analysis we undertake in this course can be accomplished in a variety of computing platforms. Feel free to use whatever computational tools will best suit your own work.

That said, $\mathcal{R}$ will be the primary statistical package used by the instructor. All worked examples in the lecture notes will have $\mathcal{R}$code associated with them. Expertise in other programs like Matlab, Python, or STATA will be useful, but all assignments and solutions will be given for implementation in $\mathcal{R}$. If you are capable of programming in lower level languages like C, C++ that's great, but I am not a programmer and will not be able to offer much help there.

$\mathcal{R}$ is a free, open source statistical computing environment. You can download and install the most recent version of from CRAN at `http://cran.r-project.org/`. $\mathcal{R}$ runs on most operating systems, including Windows, Linux, Unix and the Apple OS/X (and earlier) operating systems. While $\mathcal{R}$ is at the cutting edge of statistical software, it takes time to master. Lab sessions for the course will emphasize the nuts-and-bolts of statistical computation in $\mathcal{R}$.

LaTeXand Markdown

TeX is a typsetting system that is very commonly used by technical writers (i.e., those making extensive use of equations, numbers, tables, and graphics). Most in political science have settled on the LaTeX "dialect" of TeX. TeX is freely available via the TeX Users Group at http://www.tug.org/ or via the Comprehensive TeX Archive Network (CTAN), primary repository for TeX-related software: http://www.tug.org/ctan.html. There is a good, free

implementation of LaTeX for Windows via MikTex: `http://www.miktex.org/`. If you are working in the Mac world, you have a few choices, several of which come pre-installed with OSX.

The learning curve for LaTeX is fairly steep but I strongly encourage you to use this course to learn and practice writing in LaTeX. While I do not require that you adopt TeXfor your own purposes, there are numerous benefits to doing so, among them easier formatting of your doctoral dissertation and job market signaling. I will be available to help you with some LaTeX related problems. While there are numerous books and online references for TeX the reality is that most find it useful to start with a file template someone else created that already works replace their stuff with your stuff, tweak it a bit, and declare victory.

Moving beyond LaTeX, we can now use Rmarkdown (combined with Pandoc and the `knitr` package) to write and compile documents that incorporate text, $\mathcal{R}$code, and output in the same file. This document can be compiled directly, incorporating changes in analysis and graphics directly and immediately into the final report. If you are interested in learning to work with these tools you should feel free, but with the caveat that native LaTeXis still easier for the production of posters.

You are ultimately free to use whatever text processor you wish; if you are using WYSIWYG oriented word processors such as Microsoft Office, StarOffice, WordPerfect or other similar products for this task, you will need to learn about (and possibly purchase) MathType (`http://www.dessci.com/en/products/mathtype/`) in order to adequately display quantitative materials.

Text editors

Text editors are stand alone programs meant to aid in writing, retaining, and (sometimes) compiling computer code. Regardless of the statistical package you choose to use, you must keep a record of the code you ask the computer to execute. Working in LaTeX also requires you to enter both your text and the appropriate LaTeX commands into a text editor. You should therefore find a text editor that you are comfortable working in. While the options are myriad, Sublime 3 is my preferred text editor, although it is not free. WinEDT is and excellent editor in a Windows environment, working very nicely with the MikTeX. If you are interested in a freeware option, you can get copies of the EMACS editor (www.xemacs.org), which links to R directly via the Emacs-Speaks-Statistics (ESS) package. Aquamacs is the commonly used free text editor in for Mac OSX. TextWrangler is a popular option. Finally, RStudio comes pre-loaded with a text editor and the ability to compile Markdown files into both HTML and pdf (via LaTeX.)

Course Outline

Week 1: Introduction to the class; assignments; projects; $\mathcal{R}$; LaTeX; Good computing practice.

Readings

- Nagler, Jonathan. 1995. "Coding Style and Good Computing Practices." `http://www.nyu.edu/classes/nagler/quant2/coding_style.html`
- Ward & Ahlquist Chapter 1.
- LaCour, Michael and Don Green. 2014. "When contact changes minds: An experiment on transmission of support for gay equality." *Science*. 346(6215). 1366-69.
- Brockman, David, et al. 2015. "Irregularities in LaCour 2014." `https://web.stanford.edu/~dbroock/broockman_kalla_aronow_lg_irregularities.pdf`

Week 2: Likelihood Theory

Readings

- Ward & Ahlquist Ch. 1, 2, 4.1. Scan chapter 1 again then dig in to chapter 2.
- Mebane & Sekhon. 2002. "Coordination and Policy Moderation at Midterm." *APSR.*

Week 3: Binary Data

Readings

- Ward & Ahlquist ch.3 and rest of chapter 4
- Choose at least one of:
  - Burden et al. 2014. "Election Laws, Mobilization, and Turnout" *AJPS*.
  - Barber et al. 2016. "Ideologically Sophisticated Donors: Which Candidates Do Individual Contributors Finance?" *AJPS*.

Week 4: Out-of-sample prediction, model selection, and interpretation

Replication project must be approved by today!

Readings

- Ward & Ahlquist ch. 4-5
- Hainmueller, et al. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice"
- ProPublica investigation into racial bias in criminal justice predictive algorithms:
  - `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-se`
  - `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algori`
  - `https://github.com/propublica/compas-analysis`

Week 5: GLMs and ordinal data

Readings

- Ward & Ahlquist ch. 6-7
- Treier & Jackman. 2008. "Democracy as Latent Variable." *AJPS*
- Nicholson & Hansford. 2014. "Partisans in Robes: Party Cues and Public Acceptance of Supreme Court Decisions." *AJPS*.

Week 6: Nominal data

Readings

- Ward & Ahlquist ch. 8
- Choose at least one of:
  - Kalyvas & Balcels. 2010. "International System and Technologies of Rebellion." *APSR*
  - Olken. 2010. "Direct Democracy and Local Public Goods." *APSR*

Week 7: Counts

Readings

- Ward & Ahlquist ch. 9
- Choose at least one of:
  - MacDonald. 2010. "Limitation Riders and Congressional Influence over Bureaucratic Policy Decisions." *APSR*

- Dube et al. 2013. "Cross-Border Spillover." *APSR*
- Weghorst & Lindberg. 2013. "What drives the Swing Voter in Africa?"' *AJPS*.

Week 8: Missing data and imputation.

Readings

- Ward & Ahlquist ch. 11
- Nyblade & Reed. 2008. "Who Cheats? Who Loots? Political Competition and Corruption in Japan, 1947-1993." *AJPS*
- Choose at least one of:
  - Denny & Doyle. 2009. "Does Voting history Matter?" *AJPS*
  - Houle. 2009. "Inequality & Democracy." *World Politics*.
  - Slapin. 2008. "Bargaining Power at Europe's Intergovernmental Conferences" *International Organization*.

Week 9: Non-IID data: Spatial dependence.

Readings

- Gleditch & Ward

Week 10: Poster session.

Americans with Disabilities Act

Students requesting accommodations for this course due to a disability must provide a current Authorization for Accommodation (AFA) letter issued by the Office for Students with Disabilities (OSD) which is located in University Center 202 behind Center Hall. Students must present their AFA letters to Faculty (please make arrangements to contact me privately) so that accommodations may be arranged. Requests for accommodation must be made at least two weeks in advance of the midterm exam. Contact the OSD for further information: 858.534.4382; osd@ucsd.edu; http://disabilities.ucsd.edu