Winter 2023 | Lecture: Tues./Thurs., 2:00pm-3:20pm, | Room: Hinds 184 (Lecture) | Units: 100

PLSC 40502: Data Analysis with Statistical Models

Anton Strezhnev

Office: Pick Hall 328, 3rd floor
Office Hours: Tuesdays 4pm-6pm or schedule an appointment by e-mail
astrezhnev@uchicago.edu
http://www.antonstrezhnev.com

Last Updated: 1/19/2023

Course Overview

Statistical models provide a structure for the analysis of data. Often many scientific questions revolve around drawing statistical inferences about some parameter, such as a regression coefficient. Models also allow researchers to generate predictions on new or out-of-sample data. Understanding the fundamentals of how to define, estimate and validate a statistical model is essential to the process of quantitative empirical research.

This course is part of the second year of the Quantitative Methodology sequence in the Department of Political Science and builds on the first year sequence (PLSC 30500, 30600, 30700). It will introduce students to likelihood and Bayesian inference with a focus on multilevel/hierarchical regression models. The overarching framework of this class is model-based inference for description and prediction – a complement to the design-based framework of PLSC 30600 Causal Inference. Students will learn both the theory behind Bayesian modeling as well as how to implement common estimators (e.g. Expectation-Maximization, Markov Chain Monte Carlo (MCMC)) in the R statistical programming language. Applied examples will be drawn from across the political science literature, with a particular emphasis on the analysis of large survey data (e.g. the American

National Election Survey (ANES), the Cooperative Election Survey (CES), the European Social Survey (ESS)).

This course will involve a combination of lectures and problem sets as well as final research project. Lectures will focus on introducing the core theoretical concepts being taught in this course as well as providing illustrations through worked applied examples. Problem sets will contain a mixture of both theoretical and applied questions and serve to reinforce key concepts and allow students to assess their progress and understanding throughout the course. The final project consists of an 8-12 page research note applying the methods taught in the course to an actual data analysis task.

Assignments will involve analysis of data using the R programming language. This is a free and open source language for statistical computing that is used extensively for data analysis in many fields. Prior experience with the fundamentals of R programming is required.

Prerequisites

This course assumes that you have both a background in the core concepts of probability, statistics and inference as well as prior exposure to linear regression models. Completing the first three courses in the political science graduate methodology sequence should prepare you for the material in this class. However, there are no strict, specific course pre-requisites as many different disciplines and departments offer introductory statistics classes that cover the relevant material.

If you are unsure of whether you meet the requirements, skim/read through the first six chapters of *Regression and Other Stories*, one of the books being used by this course. You should find most of the concepts behind the material relatively familiar, aside from the references to Bayesian models (which will be covered in this course).

Please contact the instructor at (astrezhnev@uchicago.edu) if you are interested in enrolling but are unsure of the requirements.

Logistics

Lectures: Tuesdays/Thursdays from 2:00pm - 3:20pm – Location: Hinds 184

Disucssion Forum: We will use a private StackOverflow forum as a course discussion platform at https://stackoverflow.com/c/plsc-40502-statistical-models/. See the Canvas page for more details on how to join.

Course Materials: Lecture materials, assignments and section code will be posted on the course GitHub page at https://github.com/UChicago-pol-methods/plsc-40502-statistical-models/.

Readings will be listed on the syllabus. I will also post links to any non-textbook readings on the Modules page on Canvas.

Textbooks

The course will involve readings from a variety of different textbook chapters and published papers. The class will not require the purchase of any textbook as they are available online. However, you may wish to obtain a paper copy for your own personal use or reference.

The two primary textbooks from which many readings will be drawn are:

- Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories.* Cambridge University Press. (An introduction to regression and multilevel modeling from an applied perspective) https://avehtari.github.io/ROS-Examples/index.html
- Gelman, A., Carlin, J., Stern H., Dunson, D., Vehtari A., & Rubin, D. (2013). *Bayesian Data Analysis.* 3rd Edition. Chapman and Hall/CRC. (A more advanced text on Bayesian modeling) http://www.stat.columbia.edu/~gelman/book/

Requirements

Students' final grades are based on three components:

- **Final Project** (55% of the course grade). The primary goal of this class is for students to guide students in completing an independent quantitative research project that could be developed into an academic publication. The final project is a paper in the style of a research note – approximately 8-12 pages in length – that answers a clear, well-defined social science research question and applies techniques related to what is covered in the course. As the methods discussed in the class are particularly popular and widely used in the analysis of survey data, I would encourage students to consider paper topics involving the analysis of data from a source such as the General Social Survey (GSS), American National Election Study (ANES), Cooperative Election Study (CES), or any other broad public opinion survey. However, this is merely an encouragement and you should consider writing on any topic that you are interested in as long as it involves some of the methods from this class. Replications and analyses of prior studies and datasets are also encouraged as long as the paper provides an original contribution and extension of the replicated work.
  - *Timeline*: You should begin brainstorming topics early and consult with me regarding your project. Come to office hours, e-mail or just drop by

Pick Hall! To make sure that the projects develop in a timely manner, you should be aware of the following three deadlines.

    * **February 2nd: Paper topic memo due**. You will submit a short one to two page memo outlining your research proposal. In it, you should describe the main theoretical and/or empirical puzzle that the paper aims to resolve, a brief discussion of the motivations behind this question, and a description of the proposed data being used and the analyses that you will undertake. You do not need to have any actual results done by this point, but you should have acquired and begun working with the data.
    * **February 28th/March 2nd: In-class presentation**. You will present a brief (10-15 minute) summary of your research results in the style of a presentation at a large APSA/MPSA-style conference.
    * **March 9th: Final paper due**. You will submit the final paper as well as a replication archive containing your data and code to the Canvas website.

  – *Collaboration policy*: Co-authoring for the final project is encouraged, however each student is expected to contribute to the final product. I encourage students to use collaboration tools such as the GITHUB version control system for managing edits to code and OVERLEAF for collaborating on the writing of the paper.
  – *Examples*: Below are a handful of short papers that are good models for the kind of project that this class is designed to help you develop. These may be helpful guides for you in formulating your question, finding your dataset and choosing your analytical method:
    * Butz, A. M., & Kehrberg, J. E. (2016). Estimating anti-immigrant sentiment for the American states using multi-level modeling and post-stratification, 2004–2008. *Research & Politics*, 3(2).
    * Butters, R., & Hare, C. (2020). Polarized networks? New evidence on American voters' political discussion networks. *Political Behavior*, 1-25.
    * Smith, C. W., Kreitzer, R. J., & Suo, F. (2020). The dynamics of racial resentment across the 50 US states. *Perspectives on Politics*, 18(2), 527-538.

- **Problem sets** (35% of the course grade). Students will complete a total of three problem sets throughout the quarter. Problem sets will primarily cover topics from the lecture and section for that week and the previous week.

  The goal of the problem sets is to encourage exploration of the material and to provide you with a clear and credible means of assessing your understanding and progress through the course. As such, problem sets are *designed* to be challenging and I expect students to find some questions difficult. Because problem sets are mainly a tool for communication between the me and you, they will not be graded in the usual fashion.

A few days after the release of the problem set, I will be posting the solution key. You should work through the problems *without* the solutions for as long as you can, but should use the solutions after having completed your write-up to check your work. In addition to submitting your assignment write-up, you will also submit responses to a brief open-ended survey outlining which questions or topics you found easy or difficult, where you needed to consult the solutions, and what problems remain unclear or confusing. I will use these surveys to inform which topics to spend more time reviewing and to update the course to focus more on areas that students might be struggling with.

Problem sets will be graded on a (+/✓/-) scale with a + awarded for complete and near-perfect work, a ✓ awarded for generally good work with clear effort shown but with some errors, and a - awarded for significantly incomplete work with major conceptual errors and little effort shown. Since the solutions will be available to you, there is essentially no reason to not get a + if you put in the effort and complete the assignment and the survey. Although the solutions will be posted, I *highly encourage* you to not simply copy and paste solutions and code from our files as this ultimately defeats the entire purpose of the problem sets.

- *Collaboration policy*: I strongly encourage collaboration between students on the problem sets and highly recommend that students discuss problems with each other either in person or via Stack Overflow. However, each student is expected to submit their own write-up of the answers and any relevant code.
- *Office hours and online discussion*: Students should feel free to discuss any questions about the problem sets with me during class and during office hours. I also strongly encourage students to post questions about both the problem sets and the readings on the course Stack Overflow board and respond to other students' questions. Responding to other students' questions will contribute to your participation grade.
- *Submission guidelines*: Problem sets will be distributed as `PDF` and `Rmarkdown` files (`.Rmd`). You should submit your answers and any relevant R code in the same format: including an `Rmarkdown` file (`.Rmd` extension) and a corresponding rendered `.pdf` file as your submission. `Rmarkdown` combines the text formatting syntax of Markdown markup language with the ability to embed and execute

chunks of R code directly into a text document. This allows you to present your code, graphical output, and discussion/write-up all in the same document. I highly recommend that you edit the distributed Rmarkdown assignment file for each problem set directly to make organization easier.

- **Participation** (10% of the course grade). I expect students to take an active role in learning in lecture. Engagement with the teaching staff by asking and answering questions will contribute to this grade as will interaction on the Stack Overflow board.

Computing

This course will use the R programming language. This is a free and open source programming language that is available for nearly all computing platforms. You should download and install it from http://www.r-project.org. Unless you have strong preferences for a specific coding environment, I also highly recommend that you use the free RStudio Desktop Integrated Development Environment (IDE) which you can download from https://rstudio.com/products/rstudio/download/#download. In addition to being a great and simple to use environment for editing code, RStudio makes it very easy to write and compile Rmarkdown documents: the format in which problem sets will be distributed. In addition to base R, we will be frequently using data management and processing tools found in the tidyverse set of packages along with basic graphics and visualization using ggplot2.

The course will also introduce the Stan language and software for specifying and estimating Bayesian models. Stan is written in C but has bindings for a variety of programming languages. We will use two interfaces for Stan in R: RStan and brms.

Schedule

A schedule of topics and readings is provided below. Each week will cover a single topic or group of topics. You should treat the readings as a reference and as a more detailed exposition of the topics discussed in lecture. Consult the readings when you want to know more or want a slightly different approach to explaining a particular topic.

Week 1: Introduction to Likelihood Inference (January 3)

- What are statistical models good for?
- What is a "parametric" model?

- The likelihood function
- Maximum likelihood estimation

Readings

- **Review**: "Regression and Other Stories" - Chapters 1-7

Problem Set 1 Assigned January 5, Due January 17

Week 2: Generalized Linear Models (January 10)

- Properties of maximum likelihood estimators
- Binary outcome models, Event count models, Duration models

Readings

- "Regression and Other Stories" - Chapters 13-14
- Box-Steffensmeier, J. M., & Jones, B. S. (1997). Time is of the essence: Event history models in political science. *American Journal of Political Science*, 1414-1461.
- Wooldridge, J. M. (1999). Chapter 8: "Quasi-likelihood methods for count data." In *Handbook of applied econometrics*, 2, 35-406.

Week 3: Bayesian Inference (January 17)

- Principles of posterior inference
- How to write a bayesian model
- Quantities of interest: Posterior Mode, Posterior Mean, Credible Intervals
- Estimation and inference via Markov Chain Monte Carlo

Readings

- "Regression and Other Stories": Chapter 9
- "Bayesian Data Analysis" Chapters 10-11

Problem Set 2 Assigned January 19, Due January 31

Week 4: Multilevel regression models (January 24)

- "Hierarchical" regression models – random slopes/random intercept models
- Estimation via MCMC in `Stan`
- Interpreting and analyzing results

Readings

- "Regression and Other Stories": Chapter 9, 11, Appendix A
- "Bayesian Data Analysis" Chapter 5, 15

Week 5: Working with survey data (January 31)

- How to approach population inference from non-probability samples: constructing and using weights
- Multilevel regression + post-stratification

Readings

- TBA

Paper Topic Memo Due: February 2nd

Week 6: Mixture Models and the EM Algorithm (February 7)

- Exploratory data analysis and clustering models
- MLE and MAP estimation via the "Expectation-Maximization" algorithm

Readings

- TBA

Week 7: Item Response Theory and Ideal Point Models (February 14)

- Latent variable models from a bayesian perspective
- "Ideal point" models for voting
- Extensions to models of networks

Readings

- TBA

Problem Set 3 Assigned February 16, Due February 28

Week 8: Regularization and Model Selection (February 21)

- Variable selection and penalized regression (Ridge, LASSO)
- Cross-fitting and out-of-sample validation

Readings

- TBA

Week 9: Research Presentations + Conclusion (February 28)

Final Paper Due March 9

Assignment Schedule

- Problem Set 1: Assigned January 5, Due January 17
- Problem Set 2: Assigned January 19, Due January 31
- **Paper Topic Memo Due**: February 2
- Problem Set 3: Assigned February 16, Due February 28
- **Research Presentation**: February 28/March 2
- **Final Paper Due**: March 9

Acknowledgments

This course is indebted to the many wonderful and generous scholars who have developed causal inference curricula in political science departments throughout the world and who have made their course materials available to the public. This course in particular has been heavily inspired by Gov 2001 and Gov 2003 at Harvard University as well as Quant III at MIT. In particular, I thank Matthew Blackwell, Brandon Stewart, Erin Hartman, Molly Roberts, Kosuke Imai, Teppei Yamamoto, Jens Hainmueller, Adam Glynn, Gary King, Justin Grimmer, and In Song Kim whose lecture notes and syllabi have been immensely valuable in the creation of this course.