Political Data Analysis

University of Illinois at Urbana-Champaign

Sample syllabus

Instructor

Gustavo Diaz
diazdia2@illinois.edu
420 David Kinley Hall MC-713
1407 W Gregory Drive
Urbana, IL 61801

Course Description

This course introduces quantitative research methods from the lenses of political and social science. We start with a discussion of two different goals of data analysis: Inference and prediction. We will learn the theory and application of core statistical techniques for data analysis in social science research and industry. For example, we will try to answer questions such as: What explains whether and for whom people vote? Do government handouts help alleviate poverty? Can we predict who will win the next election? Later, we will introduce standards to determine the quality of the statistical evidence we produce.

This course involves a combination of lecture and hands-on work. We will learn the communication and programming skills to produce and share data analysis. Throughout the course, you will read two textbooks, answer problem sets, and work towards a research project that you will present as a poster.

Note: This course currently assumes a 15-week semester. The main ambition is to complement the usually academia-oriented social science research training with an introduction to machine learning. In a quarter system, I would teach these as two separate courses.

Audience

Who should take this course? The short answer is: Everyone! Data analysis nowadays is a must for any kind of job, even if you are not the person doing data analysis, you will engage with and interpret the work of someone who is doing it.

The long answer is: This course is intended primarily for majors in political science, social science, and related fields. Students from other concentrations who are interested in politics will also benefit greatly from this course. While everyone should benefit from taking a course like this, note that

there are many other courses in other departments that cover similar material in a way that may better suit your interests and career goals (List alternative courses here).

Structure

In a normal week in this course, you will read the required material in anticipation of the lecture (depending on whether TAs are available, lecture will be one or two days a week). The goal of lecture is to practice the language of quantitative social science research, some come prepared to participate and ask questions.

Our weekly schedule will also include a workshop session in which we will learn and practice statistical programming (we will call these "labs"). Early on, lectures and workshops will be unrelated. As we progress in the semester, the two will converge as we cover data analysis techniques and discuss research project ideas. **If possible, please bring your laptop to the labs.** If you don't have regular access to a laptop, please let us know so we can make the necessary arrangements.

Every week will also include a considerable portion of work outside of the classroom. In some weeks, you will complete problem sets that emphasize applying that week's material to statistical programming. In other weeks, you will write a short report detailing the progress of your research project, we will refer to these as "project milestones."

Course Goals

By the end of this course, students will acquire three essential job market skills:

1. Knowledge of statistics
2. Statistical programming
3. Written and verbal communication in the language of the social sciences

Usually, a course of this type emphasizes data analysis from the perspective of academic social science research, which emphasizes understand complex social and political phenomena. I recognize that, even in an audience of primarily political science majors, not everyone will pursue this path. Therefore, a secondary goal of this course is also to expose students to quantitative social and political science applied to government, civil society organizations, and industry. We will learn that the concepts and tools are often the same, but the language and purposes vary. In broad terms, I aim to help you build skills that will prove useful regardless of your career goals.

Requirements and Expectations

There is no formal requirement to take this course. While most of the material assumes high school math, we will fill the gaps as we go. All of the techniques in this course involve calculating a one number summary for a collection of observations in a data set. Sometimes they have different names or involve more complex calculation, but in the end we are mostly dealing with averages and standard deviations.

This course is about building and practicing skills. To maximize space for this goal, we have weekly assignments rather than exams. I expect you are willing to put the time and effort to stay on top of the material and assignments. The only tried and true method to learn statistics is through practice

and repetition. We won't have the luxury of time, so I also expect you to be proactive in asking for help when you are stuck and to help others who are struggling. More than anything, I am concerned with building a foundation that will open doors for future learning opportunities. You are welcome to collaborate on problem sets, but your answers should reflect your own learning process.

Materials

Software

We will use R (https://www.r-project.org/) and RStudio (https://rstudio.com/) to learn statistical programming and perform data analysis. The advantage of R is that is free and open source, meaning that you will be able to apply everything you learn in this course anywhere else. The disadvantage is a somewhat steep learning curve. I believe the investment is worthwhile.

If you can't install software on a personal computer, a viable option is RStudio Cloud (https://rstudio.cloud/), which works on your browser and will let you access your work from any device. A free account will suffice for the purposes of this course. If resources permit, we will create a dedicated course space in RStudio Cloud so you can access the course materials everywhere.

Required Readings

This course has two required textbooks:

- Kaplan, Daniel T. 2012. *Statistical Modeling: A Fresh Approach.* Project Mosaic (SM2, read for free at: https://dtkaplan.github.io/SM2-bookdown/)
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning.* New York: Springer (ISL, available for free at: http://faculty.marshall.usc.edu/gareth-james/ISL/)

Both are starting to show their age, but they are still excellent for the purposes of this class (and they are free!). SM2 is mostly about inference, which the book refers to as statistical modeling. ISL is somewhat more advanced and covers prediction, which they call supervised learning.

Further Reading

The following books are not required, but you may be interested in reading them after taking this course to broaden your horizons.

- Angrist, Joshua D. and Jörn-Steffen Pischke. 2015. *Mastering 'Metrics: The Path from Cause to Effect.* Princeton University Press
- Grolemund, Garret, and Hadley Wickham. 2016. *R for Data Science.* O'Reilly Media, Inc (Read for free at: https://r4ds.had.co.nz/)
- Efron, Bradley and Trevor Hastie. 2016. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science.* Cambridge University Press (Available for free at: https://web.stanford.edu/~hastie/CASI/)
- Imai, Kosuke. 2018. *Quantitative Social Science: An Introduction.* Princeton University Press

Assignments

Your final course grade will depend on the following assignments. All assignments will be graded with a score from 0 to 100, with 100 being the highest possible grade. Assignments may change depending on roster size.

Problem Sets (40%)

We will complete 11 problem sets. Problem sets will usually involve a set of coding exercises, followed by some short answer questions. Problem sets are released on Monday and are due by Friday of the same week (TBD specific times). Only your best 8 problem sets will count towards your final grade. Any problem set not submitted gets 0 points.

Project Milestones (20%)

In weeks 3, 9, and 15 you will deliver a report indicating the progress on your research projects. We call these project milestones. These are intended to spread the workload of producing a research poster throughout the semester. We will release a template for each project milestone a week before it is due. We won't have problem sets on project milestone weeks. Along with a grade, you will receive feedback on how to progress with the project.

Poster Presentation (30%)

By the end of the semester, we will hold a poster session in which you will present the results of your research project. Your work will be graded by three external judges and the instructional team following the same rubric. Your poster presentation grade will be the average of all these grades, excluding the judge that assigned you the lowest score.

Students that anticipate not being able to participate in the poster session should contact me as soon as possible to make appropriate arrangements.

Participation (10%)

Finally, you will also be graded on your participation in this course. Half of the participation grade will depend on the quality of your participation during lecture, labs, office hours, and other formal communications. We all have different learning styles, so I will keep an open mind about what constitutes good participation. At a minimum, I expect you to attend lecture and labs. From there, any intervention conducive to a positive learning environment for yourself and your peers will be considered good participation.

The second half of your participation grade will depend on the quality of the feedback to your peers. At every project milestone, you will be randomly assigned to provide written feedback on the report of two of your peers following a template. Only the instructional team will know who you got assigned to, your peers will receive anonymous feedback. Your peers will then grade the quality of your feedback following a rubric. If you do not grade the feedback you receive for your peers, your participation grade for that project milestone will be 0.

TBD project milestone feedback and feedback grade deadlines.

Grading

TBD. This section will detail the conversion from points to letter grades.

Learning Tips and Resources

TBD. This section will include campus and online resources to assist writing, coding, and statistical analyses.

Policies

TBD. This section will outline course policies on late assignments, academic honesty, classroom behavior, study habits, special needs, extra credit etc.

Schedule

Week 1: Introduction and Overview

Read: SM2 chapters 1 and 2

Lab: Introduction to R and RStudio

Due: Practice problem set

Week 2: Understanding Variation

Read: SM2 chapters 3 and 4

Lab: Data processing

Due: Problem set 1

Week 3: The Language of Models

Read: SM2 chapter 6

Lab: Data visualization

Due: Project milestone 1

Week 4: Linear Models for Inference

Read: SM2 chapter 7

Lab: Bivariate linear models

Due: Problem set 2

Week 5: Multivariate Models I

Read: SM2 chapter 8

Lab: Multivariate linear models

Due: Problem set 3

Week 6: Multivariate Models II

Read: SM2 chapter 10

Lab: Interactive linear models

Due: Problem set 4

Week 7: Uncertainty in Inference

Read: SM2 chapters 5 and 12

Lab: Sampling distributions and mathematical approximations

Due: Problem set 5

Week 8: Hypothesis Testing

Read: SM2 chapter 13

Lab: Experiments and hypothesis testing

Due: Problem set 6

Week 9: Evaluating Models for Inference

Read: SM2 chapters 9 and 14

Lab: Discussing project ideas

Due: Project milestone 2

Week 10: Placeholder for mid-semester break

Week 11: Linear Models for Prediction I

Read: ISL chapters 2 and 3

Lab: Regression for prediction

Due: Problem set 7

Week 12: Linear Models for Prediction II

Read: ISL chapter 4

Lab: Logistic regression

Due: Problem set 8

Week 13: Evaluating Models for Prediction

Read: ISL chapter 5

Lab: Prediction contest

Due: Problem set 9

Week 14: Model Selection

Read: ISL chapter 6

Lab: Making a poster

Due: Problem set 10

Week 15: Non-linear Models

Read: ISL chapters 8 and 9

Lab: Having fun with non-linear models

Due: Project milestone 3

Placeholder for poster session dates