Economics 5/Political Science 5D
Introduction to Social Data Analytics
UC San Diego
Spring 2023

Professor: David Arnold (daarnold@ucsd.edu)

Teaching Assistants

- Anjali Pai (a1pai@ucsd.edu)
- Vedant Vohra (vevohra@ucsd.edu)
- Alex Garland (jgarland@ucsd.edu)
- Kurtis Gilliat (kgilliat@ucsd.edu)
- Leonardo Falabella (lfalabel@ucsd.edu)
- Parth Hiren Shah (phshah@ucsd.edu)
- Shunsuke Hori (shhori@ucsd.edu)

Course Email: econ5poli5ducsd@gmail.com

Overview

This course has three main goals. The first goal is to introduce you to interesting and important social science questions. Each chapter will highlight a different application, often highlighting research from faculty at UCSD. We will cover a wide range of topics, including how colleges promote intergenerational mobility, what motivates people to vote, how do we identify discrimination in labor markets, among others.

The second goal is to show you how data can be used to inform our discussion of these topics. We will be using a lot of different datasets in this course. Some come from experiments that have been run by researchers. Others come from administrative datasets collected by governments. For many of the datasets, we will only have time to scratch the surface of what is possible with this data.

The third goal is to give you the tools to perform data analysis. We will focus on three popular software: Excel (1 chapter), Stata (4 chapters) and R (5 chapters). While learning coding fundamentals in each of these programs, we will shed light on big social science questions.

Prerequisites

There are no prerequisites for this course. In particular, we don't expect any prior coding experience. There will be some math that involves interpreting linear equations.

Lectures and Labs

There will be two lectures per week. All in-person lectures will be recorded and posted after the lecture via podcast. In addition, all the lecture material, as well as bonus material, is covered in a series of videos on Canvas. In certain cases, I will ask you to watch a video before arriving at class. Make sure you watch this video so that you are prepared for the material in lecture. Additionally, there will be reading material that accompanies each lecture.

In addition to lecture, each week there will be a lab. Each lab will involve completing an Excel workbook, Stata Do-file, or R script. If you attend in person, you will work on sections of the lab together with your classmates that attend in person.

Each week you will need to answer a short quiz related to the lab. If you attend in person, the answers to the quiz will be discussed during the lab. If you cannot attend in person, you can still get credit for the lab by completing the lab and then completing the quiz related to that lab.

Class will not be held on Monday May 29 in observance of Memorial Day.

Course Materials

All of the course materials will be made available through Canvas. This includes the software we will be using in the class, as well as a link to the online textbook for the course.

Assessment

Your grade will be based on a combination of:

- **Lab** (10%): Each week you will need to complete the online quiz associated with the lab. If you attend lab the answers for the lab will be covered during the lab itself. If you do not attend, you can still get credit by completing the lab on your own time and completing the associated Canvas quiz.

- **Quizzes** (15%): These are separate from the lab quizzes you will turn in. There will be weekly quizzes. You will be able to drop the grade of the lowest quiz. Quizzes will be posted on Wednesday night and must be completed by Friday at 11:59 PM. In general, they will be a mix of multiple choice and occasional short answer questions.

- **Midterm** (25%): The Midterm will be held in class on **Wednesday May 3rd (week 5 of the course).**

- **Final** (50%): The final will be a cumulative exam, but will be weighed much more heavily towards the R portion of the class (weeks 6-10).

The course is graded on a relative curve. Letter grades will depend on your ranking in the class. In past classes, generally you can be confident you will receive an A of some type (A+,A, or A-) if you are in the top 30 percent. You can be reasonably confident you will receive at least a B of some type (B+,B, or B-) as long as you are in the top 70-75 percent of the class. These numbers are only meant as guide to help you understand how well you are doing in the class. They are **not** concrete rules.

Important Due Dates

- Weekly labs are due Sundays at 11:59 PM.
- Weekly quizzes are due Friday at 11:59 PM.
- **Midterm:** Wednesday May 3rd in class.
- **Final:** Wednesday June 14th (06/14/2023) at 8:00-10:59 AM. Location: TBA.

Academic Honesty and Plagiarism

All graded work must be done by you. If you are unfamiliar with the University's policy on academic integrity, please see http://senate.ucsd.edu/Operating-Procedures/Senate-Manual/Appendices/2. There is a zero-tolerance policy for academic integrity violations, and if you are found to have violated the University's academic integrity policy you will receive a failing grade in the course.

Course FAQs

1. **What do I need to buy for the course?** Nothing. We will use an online textbook that was written for the course and is available through Canvas. The software we will be using (Excel, Stata, and R) is also available through Canvas with installation instructions on the homepage.

2. **Are there any technology requirements:** We will be using Stata and R throughout the course. You can download these on either PCs or Macs, but **chromebooks are very difficult to download the programs on.** There are computers in the library that have these programs installed. The Data and GIS lab: https://library.ucsd.edu/computing-and-technology/data-and-gis-lab/index.html has computers that have these programs. However, if you attend lab in person, most of the lab will be using your laptop to complete a coding assignment. If you do not have a laptop that is capable of downloading R and Stata, we are looking into getting loaner laptops that can be used during certain lab times. Please reach out to daarnold@ucsd.edu if you have concerns about these technology requirements or would like to look into using a loaner laptop for labs.

3. **How can I stay organized in this class?** In the Modules tab in Canvas, there will be a page named Week X: Plan for the Week, for every week in the course. This page includes a calendar of everything that is going on in the week, including readings, lectures slides, lecture code, and links to quizzes and labs. The final step for each week is a list of deliverables you are expected to finish by the end of the week.

4. **I have a question about the course, where should I go to ask this question?**

    a. First, check the course website and the syllabus. For example, if the question is: is there a quiz this week, you should navigate the Modules and find the given week for the course. Many questions can be answered by looking through the Plan for the Week on Canvas.

    b. If you still can't find the answer to your question, you can ask it on Piazza (see the Piazza tab from within Canvas). **This should be where you post most of your questions. Please, however, do not post questions that include code or partial answers to quizzes/labs.**

    c. If you have a question related to the course that you don't think is relevant for other students, you can send an email to the class email: econ5poli5ducsd@gmail.com. This email is monitored by me and the teaching assistants. We will try to reply to emails within 24 hours, but this may take longer on weekends.

    d. If you have a particularly sensitive question that you don't want to be read by others, you can send me an email at daarnold@ucsd.edu

5. **I'm on the waitlist, what are my chances of getting off?** Waitlists at UCSD are automated, and I can't manually allow anyone off the waitlist. While we try to expand the class when there is excess demand, we are sometimes constrained by classroom size and TA availability. For Spring 2023, unfortunately we are not able to expand the course beyond the current size.

6. **I think a question on my quiz or test was graded incorrectly.** For quizzes or labs you can submit a regrade request to the course email address with the title ECON5/POLI5D REGRADE REQUEST. For midterms and finals, regrade requests will be done through Gradescope.

Course Schedule

The schedule below lays out what is covered each week both in terms of the empirical application as well as the coding and software.

Week 1: Introduction to Excel

- Empirical Application: Instructor Incentives and Student Performance, by Andy Brownback and Sally Sadoff (2020)
- Data tables
- Functions
- Pivot tables

Week 2: Introduction to Stata

- Empirical Application: Intergenerational Mobility Rates by College. Data comes from Opportunity Insights
- The Stata Graphical User Interface (GUI)
- Do-files
- Basic data analysis commands
- Interpret and constructing histograms

Week 3: Data Wrangling in Stata

- Empirical Application: Racial Discrimination in Traffic Stops. Data comes from the Stanford Open Policing Project
- Introduce concept of data wrangling
- Learn the append, merge, and collapse commands
- Bar charts in Stata
- Ways to improve data visualization in Stata

Week 4: Regression in Stata

- Empirical Application: Disrupting Education using Technology, by Muralidharan, Sing, and Ganimian (2019)
- Estimate and interpret linear regressions in Stata
- Introduce concept of fitted values and residuals
- Visualize and plot the results of regressions in Stata

Week 5: Binned Scatter Plots (MIDTERM WEEK)

- Empirical Application: The Legacy of Colonial Medicine by Lowes and Montero (2021)
- Binned scatterplots
- Missing values and value labels

Week 6: Introduction to R

- Empirical Application: Resume Experiments, by Bertrand and Mullainathan (2004)
- Objects and variables in R
- Introduction to data frames
- Subsetting data frames in R

Week 7: Data Wrangling in R

- Empirical Application: The Rug Rat Race, by Garey Ramey and Valerie Ramey
- If statements
- For loops
- Introduction to the tidyverse package

Week 8: Data Visualization in R

- Empirical Application: China's War on Air Pollution by Greenstone, He, Jia and Liu)
- Histograms, scatter plots, and box plots in R

- Using dates in R
- Ggplot2

Week 9: Linear Regression in R

- Empirical Application: The Butterfly Ballot
- Linear regression in R
- Plotting the regression line
- Predicted values and residuals

Week 10: Functions in R

- Empirical Application: The Impact of Unconditional Cash Transfers by Haushofer and Shapiro
- Building your own functions in R