Economics 5/Political Science 5D
Introduction to Social Data Analytics

UC San Diego
Winter 2024

Professor: David Arnold (daarnold@ucsd.edu)

Teaching Assistants

- Anjali Pai (a1pai@ucsd.edu)
- Ruby Mittal (rumittal@ucsd.edu)
- Yi Zhou (yiz192@ucsd.edu)
- Muhammad Karim (mukarim@ucsd.edu)
- Maddison Erbabian (merbabian@ucsd.edu)

Overview

This course has three main goals. The first goal is to introduce you to interesting and important social science questions. Each chapter will introduce 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. You should feel free to ask questions about quizzes and labs at office hours.

Class will not be held on Monday January 15th in observance of Martin Luther King Jr. Day or Monday February 19th in observance of President's 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. The last question of every lab quiz will be to submit the completed lab script. **If you do not submit a completed lab script, you will automatically receive a zero on the lab, regardless of your other answers.**

- **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. These are open-note/open-textbook, but you should work independently on the quizzes.

- **Midterm** (25%): The Midterm will be held in class on **Wednesday February 7th (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 generally graded on a relative curve. 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. This is only a rough guide. These numbers are only meant help you understand how well you are doing in the class. They are **not** concrete rules. In particular, there are objective standards as well. For example, if you get above a 90 percent as your final course grade, you will be guaranteed an A- at the very least.

Important Due Dates

- Weekly labs are due Sundays at 11:59 PM.
- Weekly quizzes are due Friday at 11:59 PM.
- **Midterm:** Wednesday February 7th in class.
- **Final:** Saturday March 23rd 3:00-6:00 PM (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?** We will use an online textbook that was written for the course and is available through Canvas for free. The software we will be using (Excel, Stata, and R) is also available through Canvas. Check the first module which contains general course materials. One of the pages goes through all the installation instructions.

2. **Are there any technology requirements:** We will be using Excel, 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, there is also a possibility for a long-term loaner laptop from the University. Please go here: https://library.ucsd.edu/computing-and-technology/computers-and-laptops/index.html to receive more information.

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 your TA or my 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.

6. **I think a question on midterm/final was graded incorrectly** For midterms and finals, regrade requests will be done through Gradescope. For a regrade request, please state clearly why you think your solution was graded incorrectly.

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