Economics 5/Political Science 5D
Introduction to Social Data Analytics
UC San Diego
Winter 2022

Instructor
- David Arnold (daarnold@ucsd.edu)

Teaching Assistants
- Nathaniel Bechhofer (nbechhof@ucsd.edu)
- Emily Fallick (efallick@ucsd.edu)
- Anjali Pai (a1pai@ucsd.edu)

Overview

As data about individuals, organizations, and governments become increasingly available, data analysis is transforming the way we think about the economy, politics, and society. In this class, each week we will introduce an empirical application, often taken directly or inspired from research done by faculty at UCSD. We will study the empirical application of the week by learning to code in software popular in the social sciences: Excel (1 week), Stata (4 weeks) and R (5 weeks).

Normally, I consider the syllabus my contract with the class. However, this quarter some flexibility might be desirable due to all the uncertainty related to the ongoing pandemic. The syllabus that follows is my current prediction of how the class will be structured and assessed. While I will do what I can to keep to the predicted assessments for this course, the evolving situation may make it necessary for me to make changes. If that happens, I will make sure to inform you as early as possible, and to explain as best as I can the rationale behind the change.

Lectures and Labs

There will be two classes per week. Following campus guidance, all lectures between January 3rd and January 17th will be held remotely (See Zoom LTI Pro tab in canvas for links to the class). All remote lectures will be held at the regularly schedule course time and then posted afterward. Tentatively, all lectures from January 18th onward are planned to be held in person. All in-person lectures will be recorded and posted after the lecture via video podcast. In addition, all the lecture material, as well as bonus material, is covered in a series of videos on Canvas. These videos are good references if a certain aspect of lecture was unclear to you, or you would like to review a certain portion of 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. Again, the first two labs will be held virtually. From January 18th onward, the plan is to hold the lab both in-person and virtually. If you attend in person, you will work on sections of the lab together with your classmates that attend in person. If you attend virtually, you will be placed in virtual breakout rooms to work on the lab.

Each week you will need to turn in the completed lab (which is graded based on completion) through Canvas. If you attend lab, the answers will be covered during the lab. We recommend attending if you can, but if you cannot attend you can still get credit for that weeks' lab by turning it in through Canvas. However, we will not post recordings of the lab, so you will need to go complete the lab without the aid of the teaching assistants or your fellow classmates.

Assessment

Your grade will be based on a combination of:

- **Homework** (35%): 4 problem sets will be given throughout the quarter. Problem sets will contain analytical, computational, and data analysis questions. Each problem set will be counted equally toward the calculation of the final grade. The following instructions will apply to all problem sets unless otherwise noted:
    - Homework will be due slightly over a week after they are posted (see due dates section for exact dates). Each day a homework is late will reduce the grade by 10 percentage points.
    - Copies of the homework write-up should be turned in via the Gradescope tab on Canvas by the due date.
    - Although it is permissible to discuss conceptual questions with other students enrolled in class, each student must submit their own writeup of the solutions that shows their independent work on the assignment. **In particular, one should not copy someone else's answers or code or share their answers or code with anyone**. **Asking someone to send you their code (or sending your own code) is cheating.** We also ask you to write down the names of the other students with whom you solved the problems together at the top of your solutions submission. Solutions/code that appear overly similar between students will be reported to the Academic Integrity Office.

- **Take-home Final Project** (40%): Students will complete an independent project that demonstrates mastery of the material taught during the quarter. The project will be due on **Friday, March 18th at 5:00 PM**, but updates will be due throughout the quarter as part of the homework submissions. See the final project prompt for specific details on the final project. Late submissions will lose a letter grade for every day (or part thereof) late. No submission more than three days late will be accepted.

- **Lab and Quizzes** (25%):
    - **Lab** (10%): Each week you will need to submit a completed lab. If you attend lab synchronously (either in-person or via Zoom), the answers for the lab will be covered during the lab itself.

    - **Quizzes** (15%): 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 around 10 questions.

Academic Honesty and Plagiarism

All graded work must be done by you. To be explicit, ***sharing solutions or code are violations of academic integrity*** and will be reported. If you are unfamiliar with the University's policy on academic integrity, please see http://senate.ucsd.edu/Operating-Procedures/Senate-Manual/Appendices/2.

Course Website

Syllabus and course materials
The syllabus, assignments, solutions, and other course materials will be posted on Canvas. All problem sets will be turned in via the Gradescope tab on Canvas. If you are not familiar with Gradescope, take some time to familiarize yourself before the first assignment is due:

https://www.youtube.com/watch?v=u-pK4GzpId0&feature=youtu.be

Quizzes and labs will be turned in via Canvas.

Online Q&A
Given the online format of this class, it will be immensely helpful to share questions and answers on a platform that all students may access. We will use Discussions in Canvas to answer emails. If you have a question, please navigate to the Discussion tab in Canvas and find the relevant category to post your question.

If you have a question that you do not want to make public, you can email your TA or instructor depending on the subject matter.

Course Materials
Since we will be learning Excel, Stata, and R, we will draw on several different resources. Many of these resources will be videos from YouTube, blogs, and online sources. All are freely available online or have been provided by the authors. A few of the primary sources are listed below:

Excel Easy Tutorial

Princeton Stata Tutorial

UCLA Stata Resources

Software

This course will consist of three different statistical software programs commonly used by social scientists: Excel, Stata, and R. Excel and Stata both require licenses that are available for free to UCSD students. R is open-source and is free to everyone. Instructions on how to install the three software packages are located on the Canvas homepage.

Important Due Dates

- Weekly labs are due Sundays at 11:59PM
- Homework 1 (Stata) assigned Monday January 17th, due Wednesday January 26th at 11:59PM
- Homework 2 (Stata) assigned Wednesday January 26th, due Friday February 4th at 11:59PM
- Homework 3 (R) assigned Monday February 14th, due Wednesday February 23rd at 11:59PM
- Homework 4 (R) assigned Wednesday February 23rd, due Friday March 4th at 11:59PM
- **Final Project due Friday, March 18th at 5:00 PM.**

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: More Regression in Stata

- Empirical Application: Avoiding the Ask, by Andreoni, Rao, and Trachtman
- Uncertainty
- Standard errors, p-values, and confidence intervals
- Produce nicely formatted regression tables in Stata

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 and Markdown in R

- Empirical Application: The Impact of Unconditional Cash Transfers by Haushofer and Shapiro
- Functions in R
- Introduction to R markdown