PSCI 1800 910

Introduction to Data Science

[Preliminary syllabus]

Instructor: Pedro Vicente de Castro
PhD student
Political Science Department
University of Pennsylvania
vcpedro@sas.upenn.edu

Lecture time: Mondays and Wednesdays, 12:00 pm – 1:30 pm

Office hours: Fridays, 12:00 pm – 1:30 pm, by appointment

Course description

Understanding and interpreting large, quantitative data sets is increasingly central in political and social science. Whether one seeks to understand political communication, international trade, inter-group conflict, or other issues, the availability of large quantities of digital data has revolutionized the study of politics. Nonetheless, most data-related courses focus on statistical estimation, rather than on the related but distinctive problems of data acquisition, management and visualization – in a term, data science. This course addresses that imbalance by focusing squarely on data science. Leaving this course, students will be able to acquire, format, analyze, and visualize various types of political data using the statistical programming language R. This course is not a statistics class, but it will increase the capacity of students to thrive in future statistics classes. While no background in statistics or political science is required, students are expected to be generally familiar with contemporary computing environments (e.g. know how to use a computer) and have a willingness to learn a variety of data science tools.

Learning goals

The goal of PSCI 1800 910 is to produce basic competency in working with data using the R environment. There are four broad learning goals of the course:

1. The ability to load, clean, and analyze data sources using R.
2. The ability to produce graphs, figures, and other basic analyses to gain insight about those data.
3. A non-mathematical working knowledge of the key concepts of uncertainty and causality.
4. The ability to talk and write about statistical findings in a concise and clear fashion.

Working with data

Individuals leaving PSCI 1800 910 will have the ability to download and process “raw” data from the internet into a workable dataset. Topics in this module include: understanding objects & functions, under- standing different data formats, sub-setting data & creating new variables, cleaning and reshaping data, using ‘for’ & ‘if’ statements, and merging datasets.

Producing output

Individuals leaving PSCI 1800 910 will have the ability to produce – from a "cleaned" dataset – figures, tables, and other insights. This includes: generating a table of summary statistics, creating cross- tab tables, determining the correlation between variables, running simple bivariate and multivariate regressions, choosing and implementing the right plot to better understand their data (histograms, scatterplots, violin plots etc.).

Understanding uncertainty and causality

While PSCI 1800 910 will have very little math, students leaving the class will have a basic understanding of uncertainty and causality.

For uncertainty: students will understand how any particular dataset is one possible example of an infinite number of datasets that could be generated. They will understand that this means any statistic that we calculate is but one of an infinite possible number of statistics. They will understand that the distribution of these possible statistics is knowable, and allows us to estimate the likelihood of our result being "noise".

For causality: students will understand the fundamental problem of causal inference. They will know that no research design can ever determine true causality. They will understand why randomized experiments are the "gold standard" of determining causality, and how observational research can be approached in an experimental framework to improve our estimates.

Talking and writing about statistical findings

Students will leave PSCI 1800 910 with the ability to talk through charts and figures that might be found in political science journals. They will be able to talk about their own statistical findings in a way that communicates importance to a general audience. Similarly, they will be able to write about statistics in a way that is clear, free of jargon, and quickly communicates to readers the main point of the analyses.

Teaching strategies

Lectures

The best – no, the only – way to learn how to code is by doing. Nevertheless, seeing other people code and having them walk you through each step of what they are doing can be helpful. Most of our lecture time will be used in this way: live coding. But I don't want you to just sit and watch. This is a hands-on class, so you will be coding with me. On our first class I will walk you through downloading and setting up R and RStudio. But on every class after that I expect you to have R running and to be coding with me all the time.

Tutorials and graded challenges

No one has ever finished learning how to code. The number of tasks coding can help us perform done is virtually endless. Also, new tools and techniques are developed every day. As a someone who works with data, whether you are a data scientist at a company or a political scientist in academia, part of your work will involve learning by yourself how to perform a given task you have to do or how to apply a given tool or technique to your purposes. A crucial resource are tutorials: code someone else has written and commented in a detailed way with the intention of showing you how to do something. The internet is filled with tutorials, which makes learning how to do new things with code easy. In PSCI 1800 910, we are going to rely heavily on tutorials. On every week you will have a tutorial to complete and submit. Due dates are below.

At the end of every tutorial, you will be presented with one or more “challenges”. These are tasks I won’t provide the code for. Your grades depend on your performance on those. Each challenge is worth 20% of your final grade.

Just like, as someone who works with data, you will often have to learn something new, you will also often find it helpful to ask others for help. This is encouraged in this class. You are encouraged to work with your peers on the weekly challenges. But I don’t want you to just copy and paste someone else’s code. You won’t learn like that. So, you have to submit your own, original code.

Tutorial due dates:

1. Getting started, 06/05
2. Data wrangling, 06/12
3. Asking questions with data, 06/19
4. Iterations 06/26
5. Describing relationships 07/03

Communication

I will create a Slack channel in which your can ask me questions and I will try to respond in a reasonable time. You can also use it to talk to your peers and exchange tips and tricks – which, again, is incentivized in this class. I just ask you that you refrain from posting your entire challenge answers. You can post a couple of lines of code and, ideally, explain what they are doing, in reply to someone else’s question. But you can’t just post all of your code.

Preliminary class schedule:

Class 1, 05/29: Introduction to R: base R, dplyr, and R Markdown

Class 2, 06/03: Reading and tidying data

Class 3, 06/05: Combining and reshaping data

Class 4, 06/10: Describing data: summary measures and crosstabulation

Class 5, 06/12: Data visualization: introduction to ggplot

Class 6, 06/16: Iteration: loops and functions

Class 7, 06/19: Text: strings and regular expressions

Class 8, 06/23: Describing relationships: correlation and regression

Class 9, 06/26: Uncertainty and simulation

Class 10, 07/01: An introduction to causal inference

Class 11, 07/03: Review