Political Science 6: Introduction to Data Analysis

Instructor: Marika Csapo, mcsapo@ucla.edu

Class Meetings: Tu/Th, 10:45a - 12:50p, Bunche 2209A
Office Hours: Th, 1:30 - 2:30 p (or by appt), Bunche 3rd Floor TA Lounge

Summer Session A 2016

Course Objectives and Philosophy

The goal of this class is to facilitate introductory-level knowledge and practice of statistical analysis and graphics, and more importantly, to encourage careful and critical consumption, interpretation and usage of quantitative research. You can think of this as a "driver's education" class with an emphasis on defensive driving—only the vehicle you will be driving will be a powerful analytical tool, called R. Most anyone can be trained to come to a complete stop before the line at a signal; however, keen awareness of potential problems down the road, a wary eye, and conscientiousness are skills that are more subtle, but in the long run determine the difference between a competent driver and a public safety hazard. That is correct; using statistics carelessly is akin to wreckless driving. The most savvy of practitioners internalizes the objectives of safe driving, not just the rules, and so is equipped to deal with uncharted territory and unique situations in a safe, if improvised, manner.

Students in this class will learn by the end of the quarter to succinctly summarize and distill useful information about the world from data, to discriminate between more and less appropriate summaries for your purposes, and to differentiate between what you can and cannot say about the world with this information. To do well in the class and get the most out of the material, you need to attend lecture regularly, where I will put a large emphasis on conceptual understanding and interpretation. I repeat, DO attend class. If you attend class and still find the material confusing, I will happily meet with you outside of class to work on it. Office hours should NOT, however, be treated as a substitute for attending lecture. Please exchange emails with other students so that they can share their notes if you do happen to miss a class. I post lecture slides on the website for your convenience. These are meant to complement, not replace, lecture.

Finally, do not get bogged down in thinking about this as a math class. For our purposes, math is a means to an end, not the objective itself. In an era where just about any commonly-used statistical formula is easily accessible on Wikipedia, knowing what questions to ask and where to look for answers are bigger assets than memorizing formulas. Furthermore, once you really understand this stuff conceptually, you can probably intuit the formula that you would need to answer your question, and hopefully even think of several alternatives to it and why one might be better than the other for your question. The math does, at times, contribute to the conceptual understanding and so we will use it to that end. You do not need to be a math whiz to do well in this class.

Homework and Grades

The assignment and evaluation structure of this class is based on the philosophy that the best way to get good at data analysis is to jump right in and do it. The marginal returns to reading about how to do statistics or watching someone else compute statistics for an hour are incredibly small compared to the returns to spending that hour doing and interpreting data analysis yourself. Therefore, homework will be the most important study tool for this class. In putting more weight on homework and projects less weight on exams, I hope to discourage binge consumption of the material, which will not serve you well in the long-run. What you get out of this class will be proportional to how much effort you put into the homework. Your focus for the homework should not just be on documenting your results, but also on carefully articulating and interpreting those results. There is no required text.

Depending on your learning style, diving right into doing data analysis may be a little bit uncomfortable. Lets try to set aside our self-judgment and comparison to others and just be curious about the material. If something is hard or confusing, that's okay; it will get easier. Do not let that discourage you from deconstructing the task into pieces and tackling them slowly, one-by-one; aim for a slow accumulation of knowledge. Using this approach, you will understand considerably more at the end of the quarter and will remember it for longer afterward. It is not my goal to make the homework painful. As a pain management strategy, the homework holds your hand through most tasks, and then asks you to stretch yourself a little bit on a couple of tasks. For the most part, homework will be due every Monday (note, however, there is nothing due on Monday, 7/11, and that the last assignment, Homework 5, is due on Friday, 7/29). You should submit your homework to `turnitin.com` via `my.ucla.edu` by 11:59 pm on the day it is due. Late homework loses one point per day.

Homework and Quiz Dates

Assignment	Due Date	Contribution to Grade
In-Class Assignments	Varies	15%
Homework 1	6/27	15%
Homework 2	7/04	15%
Midterm (In-Class)	7/12	20%
Homework 3	7/18	15%
Final Project (Take-Home)	7/29	20%

One point will be added to your lowest homework grade when you submit your end-of-the quarter electronic evaluation of instruction on `my.ucla.edu`. I will not know what you wrote, only whether or not you submitted it.

Course Policies

Collaboration: Students are welcome and even encouraged to discuss homework problems and R techniques outside of class. However, each student MUST do their own write up, using their own examples, and their own interpretation of the results. Turning in the same write up as another student (all or in part) will be considered academic dishonesty. Collaboration on in-class quizzes is not allowed and will also be considered academic dishonesty.

Academic Dishonesty: Students that do not turn in their own work will receive an "F" on the assignment or quiz, and depending on the severity of the infraction, possibly in the class. Cheating on exams must also be reported to the university.

Required Skills: If you have a functional knowledge of arithmetic and algebra, you have enough math training to do fine in this course. Expressing yourself clearly and precisely in writing is a skill that is both less common and, for this course, possibly more important. You will need to be able to articulate the motivation for your analysis (for example, the research question), interpret the results of your analysis, and acknowledge the limitations of the analysis. Writing is key. You also need to have very basic computer skills (which I expect everyone of your generation to have). You do not need to know how to program. You will learn by doing and lecture will give you what you need.

Hardware and Software: You need access to a computer in order to do the homework (one can be borrowed from the Young Research Library, if necessary). Depending on what type of learner you are, it may behoove you to bring a laptop to class. You will need to download R and RStudio to the computer on which you plan to do your homework. Both are free. Download R first at `http://cran.r-project.org/`, then RStudio at `http://www.rstudio.com/products/rstudio/download/`. Any of the last couple of generations of the software should be fine.

Reading Materials: There are no assigned texts for this class. Your best resource as an adjunct tool to lecture is the internet (do not underestimate the value of instructional youtube videos!). If you feel compelled to purchase a text, check out Moore and McCabe's *Introduction to the Practice of Statistics*, any edition.

Lecture Schedule

Lecture 1, Tuesday, June 21
What *are* statistics? What *is* Statistics? What is a random variable? Introduction to R. Numeric summaries, box plots and histograms of a single random variable.

Lecture 2, Thursday, June 23
Summarizing and visualizing a relationship between two random variables. Scatterplots, trend lines, covariance and correlation coefficient.

Lecture 3, Tuesday, June 28
Bivariate regression and Ordinary Least Squares.

Lecture 4, Thursday, June 30
Introduction to the Central Limit Theorem. Diagnostic techniques for evaluating Ordinary Least Squares, including non-linearity.

Lecture 5, Tuesday, July 5
Linearizing non-linear relationships and variable transformations. Aggregation issues and Simpson's Paradox.

Lecture 6, Thursday, July 7
More diagnostics—identifying confounding factors. Gauss-Markov assumptions. Problems in step-wise regression.

Lecture 7, Tuesday, July 12
Alternatives to step-wise regression. Partial Regression. Multivariate regression.

Lecture 8, Thursday, July 14
Variants of multivariate regression (dummy variables, "fixed effects" or factors and interaction terms). Interpreting R^2. Predicted values and counterfactual prediction.

Lecture 9, Tuesday, July 19
Causal claims, spuriousness, confounding and statistical control. Causal design and randomization. Internal and external validity.

Lecture 10, Thursday, July 21
Sample simulation. Introduction to Statistical Inference.

Lecture 11, Tuesday, July 26
Hypothesis testing, p-statistics and Type I and Type II error tradeoff.

Lecture 12, Thursday, July 28
Summation. Preview of relevant future topics. Philosophical concerns.