Data Analytics 1: Quantitative Analysis

Harris School of Public Policy
University of Chicago
Winter 2018
Time: Thursdays, 6:00-8:50 p.m.
Location: 1871

Instructor
Anthony Fowler
anthony.fowler@uchicago.edu

Teaching Assistant
Miguel Morales-Mosquera
mmoralmo@uchicago.edu

Course Description

This class will provide an introduction to quantitative analysis in public policy. Much of the class is devoted to learning about the effects of policies and answering empirical, policy-relevant questions from observational data. In doing so, the course provides an introduction to critical, quantitative thinking in general. Students will be introduced to the basic toolkit of policy analysis, which includes sampling, hypothesis testing, Bayesian inference, regression, experiments, instrumental variables, differences in differences, and regression discontinuity. Students will also learn how to use a statistical software program to organize and analyze data. More importantly, students will learn the principles of critical thinking essential for careful and credible policy analysis.

Materials

The following book is required:

Bueno de Mesquita, Ethan and Anthony Fowler. 2017. Critical Thinking in a Data-Driven World. Book manuscript (available on the course website).

The following books are optional:

Angrist, Joshua D. and Jörn-Steffen Pishke. 2014. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press.

Ellenberg, Jordon. 2014. *How Not to Be Wrong: The Power of Mathematical Thinking*. Penguin Books.

Mumford, Stephen and Rani Lill Anjum. 2014. *Causation: A Very Short Introduction*. Oxford University Press.

For several of the course assignments, students will utilize a statistical software program. Stata is strongly recommended unless a student is already highly proficient in an alternative (e.g., R, SPSS, GAUSS). Students can access Stata through vLab using their CNetID.

The following online resources may be useful for students learning to use Stata for the first time. We will also provide helpful information on the course website and in class.
http://stats.idre.ucla.edu/stata/seminars/notes/
http://stats.idre.ucla.edu/stata/webbooks/reg/

Requirements

Students are expected to read all assigned readings carefully and come to class ready to discuss them. Although class attendance and class participation are not explicitly incorporated into the final grade, success on the assignments and exam will depend upon engagement with the readings and the class sessions.

All assignments should be submitted through the course website in pdf format by 11:59 p.m. on the stated due date. There is a word limit for each assignment, and the grader will stop reading after you have exceeded that limit. All assignments should be formatted in a professional way, as if you were presenting them to a boss, client, or colleague. Out of fairness to all students, late assignments will not be accepted under any circumstances. Please plan ahead accordingly. If unforeseeable circumstances prevent you from successfully completing an assignment, please turn in whatever you have by the deadline, and you will at least receive partial credit.

Grading

Students will be graded based on two writing assignments, two data analysis projects, and a final exam. All items will be graded out of 20 possible points. At the end of the course, the grades will be re-weighted, with each student's worst grade down-weighted by half and each student's best grade up-weighted by 50 percent. In other words, each assignment will have an average weight of 20 percent, but a student's best assignment will count for 30 percent of their grade, and their worst assignment will count for 10 percent of their grade. The rationale is to reward students for exceptional performance on an assignment and to mitigate the effects of one bad assignment, which could have resulted from bad luck or extraneous circumstances. Uri Simonsohn calls this TWARKing. To read more about it, along with arguments for why this is better for learning, measurement, motivation, and student happiness, see datacolada.org/56.

After re-weighting, each student will have a grade out of 100 points. A score of 90 points will ensure an A, 85 points will ensure an A−, 80 points will ensure a B+, 70 points will ensure a B, 60 points will ensure a B−. Remember that these thresholds apply to your re-weighted grade, meaning that a good score helps you more than bad performance hurts you.

After each assignment has been graded, you will be able to see your grade on the course website, along with any comments specific to your assignment. I will also e-mail the class with overall feedback on the assignment.

If you would like to contest a grade, you must do so in writing within one week of receiving your grade for that assignment or exam. Send me a brief e-mail and explain why you feel you were mis-graded. If I agree that an error was made, I will regrade your entire assignment, at which point your grade may go up or down.

Questions

If you e-mail me or a TA about the course, please include "Data Analytics I:" in the subject heading, followed by the specific subject of your e-mail.

If you have an administrative question about the course (e.g., grading, deadlines, requirements, expectations, procedures, etc.) that applies to all students but is not answered on the syllabus, please ask it in class. If the question is personal in nature, e-mail me.

If you have an intellectual/academic question, ask it in class. Other students will likely benefit from it.

Please do not ask an instructor or teaching assistant to read your assignment before you turn it in. This wouldn't be fair to the rest of the class as we cannot possibly do this for every student. However, if a specific question arises as you're working on your assignment, feel free to ask it in class or via e-mail.

Because this is a part-time program and most students are unable to come to the Hyde Park campus during the day, there will be no regular office hours. We will try to make ourselves available to answer questions before, during, and after class as well as via e-mail.

Assignments and Exam

Writing Assignment #1: Failure to Compare

Although correlation requires variation, many analysts fail to make adequate comparisons before making claims about correlations. Find a recent case where an analyst wants to make claims about a correlation but doesn't conduct the relevant comparison or collect the relevant data to assess that correlation. Your example could come from a news article, academic article, or related source and it should not be too similar to the examples discussed in class or in the readings. Summarize the argument made and discuss the inferential problem. What alternative analysis could the author have conducted instead in order to make a more informative comparison? Your write-up should be less than 1,000 words.

Data Analysis #1: Predicting Voter Turnout

Students will be provided with data from the 2011 North Carolina voter file on the course website. Specifically, students will have access to the year of birth, registration year, race, gender, county, city, zip code, precinct, and turnout history of 2,000 randomly selected registered voters. For a randomly selected 1,000 of those voters (the target population), data on their 2010 turnout is missing. Using data from the other 1,000 registered voters (the sample population), students should predict the probability that each individual in the target population voted. This is similar to the kind of exercise a quantitative analyst might conduct before an election. Perhaps you want to know how likely a person is to vote so that you can decide whether to contact them with a get-out-the-vote intervention.

Students should turn in a brief written report (1,500 words or less) explaining how they generated their predictions and how they settled upon that particular strategy. Students should also turn in a data set with their predicted probabilities for the target population. See instructions below. Students must follow these instructions precisely in order to receive credit for this part of the assignment.

Half the points on this assignment will be determined by the written report and the extent to which students have carefully justified their predictions. The other half will be determined by the actual accuracy of the predictions. I know which individuals in the target population actually voted, so using this data and student predictions, I will compute the mean squared error for each student, and the number of points out of 5 will increase monotonically as the mean squared error decreases. As indicated above, in order to receive a score on this part of the exam, students must provide a predicted probability for all 1,000 individuals in the target population and they must follow the instructions for formatting their data set.

Instructions for submitting predictions: In addition to their written report, students will also upload their predicted probabilities for each individual in the target population. In the provided data set on registered voters in North Carolina, each registrant has a unique id number (the id's for the target population range from 1 to 1,000). Students should submit their predictions in a .csv file with 1,000 observations (one for each registered voter) and 2 columns (one for each registrant's id number and one for their predicted probability of voting). The two variable names should be "id" and "pred" and both must be lowercase. The "id" variable should have integers ranging from 1 to 1000, and the "pred" variable should have continuous numbers ranging from 0 to 1. At least one week before the assignment is due, each student will be given their own id number, and the name of their csv file, when they turn it in should be "predictions_X.csv" where X is replaced by their unique number. An example of the file that students should submit with the hypothetical id number of 99 (and with

predictions generated using only turnout in 2008) will be available on the course website. Furthermore, Stata code for reading in the assigned data set and producing the sample output is also available on the course website.

Writing Assignment #2: Causal Inference

Comparisons and correlations may not be informative about causal relationships. Find an example of a researcher, journalist, policymaker, etc. who makes an error by wrongly interpreting a correlation as causal. As before, your example should not be closely related to any example discussed in class or in the readings. Explain the evidence presented, and explain why you think this correlation is not informative about the effect of interest. Furthermore, propose an alternative research design that would provide more credible causal evidence on this particular question. Your proposed design should be feasible, meaning that a real researcher or organization with finite resources could actually implement it. Again, your write-up should be no more than 1,000 words.

Data Analysis #2: Candidate Divergence

Using data from the U.S. House of Representatives, estimate the extent to which Democratic and Republican candidates would differentially represent the same district. Implement whatever research design you feel is appropriate, and discuss its pros and cons relative to other designs that you considered. The written report should be brief (1,500 words or less). The report should include enough detail such that another student could easily understand and reproduce your work, but extraneous details should be excluded. Please paste your Stata code at the end of your report.

Final Exam

There will be a final exam during the last class session. Students are free to use books, notes, laptops, and online resources while taking the exam. However, students should not discuss the exam or communicate with any other person while taking the exam.

Deadlines and Dates

January 22 – Writing Assignment #1
February 5 – Data Analysis #1
February 19 – Writing Assignment #2
March 5 – Data Analysis #2
March 8 – Final Exam

Class Sessions and Readings (*optional reading)

1. January 4

Thinking and Data: Substitutes or Complements?

Bueno de Mesquita and Fowler, Chapter 1
*Angrist and Pischke, Introduction
*Ellenberg, Introduction (When Am I Going to Use This?)

Correlation: What is it and what is it good for?

Bueno de Mesquita and Fowler, Chapter 2
*Ellenberg, Chapters 1 and 3

2. January 11

Causation: What is it and what is it good for?
Bueno de Mesquita and Fowler, Chapter 3
* Mumford and Anjum, Chapters 1, 5, and 10

Correlation Requires Variation
Bueno de Mesquita and Fowler, Chapter 4

3. January 18

Regression for Description and Prediction
Bueno de Mesquita and Fowler, Chapter 5

Inferences about Relationships
Bueno de Mesquita and Fowler, Chapter 6

4. January 25

Overfitting, Multiple Testing, Reporting Bias
Bueno de Mesquita and Fowler, Chapter 7
*Ellenberg, Chapters 6, 7, and 9

Regression to the Mean
Bueno de Mesquita and Fowler, Chapter 8
*Ellenberg, Chapters 14-15

5. February 1

Compare Apples to Apples
Bueno de Mesquita and Fowler, Chapter 9

Regression, Matching, and Selection on Observables
Bueno de Mesquita and Fowler, Chapter 10
*Angrist and Pischke, Chapter 2

6. February 8

Randomized Experiments
Bueno de Mesquita and Fowler, Chapter 11
*Angrist and Pischke, Chapter 1

Noncompliance and Instrumental Variables
Bueno de Mesquita and Fowler, Chapter 12 (IV section)
*Angrist and Pischke, Chapter 3

7. February 15

Regression Discontinuity Designs
Bueno de Mesquita and Fowler, Chapter 12 (RD section)
*Angrist and Pischke, Chapters 4

Differences-in-Differences Designs
Bueno de Mesquita and Fowler, Chapter 12 (DD section)
*Angrist and Pischke, Chapters 5

8. February 22

Inferential Problems, redux
Bueno de Mesquita and Fowler, Chapter 13

Translating Evidence into Beliefs, Bayes' Rule
Bueno de Mesquita and Fowler, Chapter 14
*Ellenberg, Chapter 10

9. March 1

Consider Adaptation
Bueno de Mesquita and Fowler, Chapter 15

Turn Statistics into Substance
Bueno de Mesquita and Fowler, Chapter 16
*Ellenberg, Chapters 4-5

10. March 8

Combining Theory, Design, Data, and Critical Thinking
Bueno de Mesquita and Fowler, Chapter 17

Final Exam